# Behavioral Economics – Enhanced: Machine Learning and Decision-Making

Inauguraldissertation
zur
Erlangung des Doktorgrades
der
Wirtschafts- und Sozialwissenschaftlichen Fakultät
der
Universität zu Köln

2020

vorgelegt von
Carina Ines Hausladen
aus
Pemfling (Deutschland)

Referent: Prof. Dr. Martin Fochmann
Koreferent: Prof. Dr. Dr. h.c. Christoph Engel
Tag der Promotion: 05.10.2020

Advisors: Martin Fochmann & Christoph Engel Carina I. Hausladen

# Behavioral Economics – Enhanced: Machine Learning and Decision-Making

## Abstract

In this thesis, I investigate decision-making in the fields of behavioral economics, experimental economics, and law and economics. The research questions I ask are: Can we nudge people towards being more honest? Can we use language to find out who lies? Which factors influence a judge's decision, and how do people cooperate? Specifically, I investigate contributions in a public goods game, (dis-)honest decision-making in a die-in-the-cup and tax compliance game. Furthermore, I investigate the bounds of rational decision-making in the context of the law. To answer the posed questions, I apply – alongside traditional econometrics – machine learning methods: I use natural language classification to predict decisions based on text data. Furthermore, I use time-series clustering to reduce complexity and thereby enable theory building and interpretation.

Keywords: Behavioral Economics, Experimental Economics, Law and Economics, Machine Learning, Natural Language Processing, Mouse Tracking, (Dis-)honest Decision-Making

# Contents

# List of Figures

# List of Tables

# 0
# Introduction

Honesty, cooperation, and law-abiding behavior are values to which all members of our society should adhere. However, instances where these values are violated, are easily found: In Germany, in the year 2018, the value of detected tax fraud alone was around 2.6 billion euros (BMF 2020), and the lack of the willingness to cooperate prevents climate protection measures to be implemented. Generally speaking, unethical, uncooperative, and law-neglecting behavior cause problems from which all members of our society suffer. To face these problems, many behavioral economists dedicate their time and effort to the topics mentioned above. With this dissertation project, I join these scholars in their endeavor. More specifically, I ask: Can we nudge people towards being more honest? Can we investigate language to find out who lies? Which factors influence a judge's decision, and how do people cooperate? In order to answer these questions, I work both with experimental and field data. The methods I deploy are highly interdisciplinary and reach from classical econometrics to machine learning.

The central goal of behavioral economics is to investigate the bounds of rationality of agents' decisions. Each of the four chapters of my dissertation deal with this central theme. In chapter 1, I study a game where participants can lie without facing the risk of being punished. In this context, rationality would predict individuals would opt for the highest possible monetary payoff, but in reality, many subjects report honestly even if this particular choice is associated with a suboptimal payoff. Nevertheless, is honesty the intuitive behavior, or is the selfish behavior a person's first reaction? So far, studies have found support for both options, leaving this question an open puzzle. Chapter 1 is an attempt to solve the dispute.

Instead of investigating the default behavior, in chapter 2, I predict participants' behavior. More specifically, I analyze written language obtained from group chats to determine who of the participants of a tax evasion game plan to lie.

Not all economic games are structured such that the rational player has to lie to earn the highest payoff. In a public goods game, the Nash Equilibrium is to freeride. Despite this theoretical prediction, experimental data displays various strategy profiles with positive contributions. In chapter 3, I investigate those strategy profiles by a data-driven approach.

I outlined three games in which experimental studies have found behavior that could not be predicted by standard economic theory. In chapter 4, I investigate such "irrational" decisions outside of the laboratory. A rational decision in the legal context should exclusively be based on law and statues. However, research shows that external variables, such as socio-demographic characteristics, additionally influence a judge's decision. To capture such a relationship, the dependent variable, namely the legal decision, needs to be modeled in a measurable form. In chapter 4, I map the written opinion text of judges on a two-dimensional scale representing political ideology.

Methods beyond traditional econometrics. Each of the four chapters of this work advances the frontiers of our understanding of bounded rationality. Beyond that, three chapters draw upon machine learning methods allowing to not only predict but also substantiate choices. The usage of these methods makes this work unique, because so far, most research in the field of behavioral economics draws upon traditional econometrics for data evaluation. However, even if data from laboratory experiments are structured and only a fraction of the size of what would be considered Big Data, complexity still exists, which is hard or even impossible to reduce by traditional econometrics. These problems arise, for example, for text data and multi-round decisions.

Much behavioral experimental research allows for communication between participants, for example, in the form of chats. However, in many cases, chat data is hardly analyzed because language is too complex to be captured by a simple analysis tool. A possible solution is natural language processing, which allows to analyze the text in a resource-efficient way.

In chapter 2, I use natural language processing to analyze chat texts from a tax evasion experiment. More specifically, a classifier is trained to label chat texts as either "honest" or "dishonest", depending on the income stated.

In chapter 4, I again deploy supervised machine learning. Much research in the field of empirical legal studies quantifies the political ideology of judicial opinions. So far, human coders assess whether the text is considered as conservative or liberal. Due to the resource intensity of the labeling process, a frequently used database in this filed holds political ideology labels for less than 5% of all judicial opinions available. I train a classifier based on the available labels to predict political ideology labels for the remaining 94% of the opinions.

Apart from text, another instance of data that requires machine learning techniques is multi-round decisions, with temporal and inter-group dependencies. More concretely, in chapter 3, I use unsupervised machine learning in the form of clustering algorithms to find strategies played by participants in a public goods game.

Contribution to the literature. This work contributes to the literature by demonstrating how methods commonly used in computer science can be leveraged for the field of behavioral economics. In the following, I demonstrate how each of the four chapters advances the knowledge in a particular field of research.

Chapter 1 contributes to the experimental literature by investigating (dis-)honest decision making. The paper proposes a solution to a long-standing dispute about whether a human's nature is selfish or not. A few papers have already argued that the diverging findings in this respect are due to a methodological flaw. In Chapter 1, this flaw is satisfyingly corrected for the first time. We achieve this goal by introducing various improvements to a widely-used game in the literature on lying, namely Fischbacher and Föllmi-Heusi (2013)'s dice-in-the-cup game. Most experimental studies increase the probability that a participant will stick with her default by limiting the time available to make a decision. However, we believe that the dice-in-the-cup game triggers a particular default behavior: Most participants have fantasized repeatedly about rolling and reporting the number six in the past. That is the winning strategy in most board games. Therefore, we introduce a novel type of die, namely a *color* die, which excludes such a lying default. We furthermore introduce a novel randomization device, a dice tower, to exclude the possibility that incomplete randomization distorts reports. Finally, the study is among the very few ones in this field of research to track mouse movements on the decision screen. Instead of analyzing a one-dimensional decision, a participant's mouse trajectories provide detailed insight into the decision process.

Based on the methods just mentioned, we conduct an experimental study that includes four treatments. Participants either roll a regular (i) or color (ii) die. After doing so, they are asked to report the upward-looking face under time pressure (iii) or not under time pressure (iv). Their payoff depends on this report, which is stated by clicking on one of six choice options displayed in a circular manner on a computer screen. While doing so, mouse movements are tracked. We analyze the reports with an ordered logit regression within a Bayesian framework. The resulting treatment coefficients are significant and point in the direction hypothesized: Under time pressure, participants rolling a colored die report lower results than participants rolling a numbered die. It is the mouse movements that add the final puzzle piece: Depending on the treatment, some trajectories start at a less and end at a more lucrative option, but we also find trajectories that start at a more and end at a less lucrative option. Based on the regression coefficients and the insights gained by the mouse trajectories, we confidently state: Honesty and dishonesty are *both* default behaviors. However, the default crucially depends on the situation: In a novel situation, such as rolling a colored die, the easiest and quickest answer is the truth.

In chapter 2, I again investigate (dis-)honesty. The chapter contributes to the usage of text in behavioral experimental research. This chapter is the first to systematically test various supervised classification methods on datasets obtained by experimental research.

Many experimental studies investigate channels that influence lying behavior, such as group interactions which are found to increase dishonesty. Those interactions often take place in the form of written text messages, mostly analyzed in a very resource-intensive way: Researchers read the text and categorize it to provide a qualitative description of the content. However, the text can be used for yet another purpose: It holds the potential to directly *predict* lying. More precisely, chapter 2 consists of two parts: In the first part, a supervised classifier is trained based on chat data to predict participants' decisions, both provided by an existing experimental study. For the second part, a new experimental intervention was conducted. Based on the resulting data, the generalizability of the classifier is tested.

The first part aims to train a predictive classifier by testing various configurations: The main parameters varied are the shapes of the dependent and independent variable, embedding techniques, and different classification algorithms. The results show that pretrained Word2Vec embeddings combined with stacking multiple classification algorithms reach the best performance. However, the predictive power of the classifier is weak, it reaches an F1 score of .411. Nevertheless, this result was expected, given the adverse conditions the model faces: Less than 700 samples are available for training and the distribution of the labels to be predicted is heavily skewed. Overall, chapter 2 shows successfully that choosing the classification configuration carefully is extremely important: The best configuration reaches a F1 score double as high as the worst configuration.

The second part of chapter 2 is concerned with generalizability. For that purpose, we design a new experimental study that alters the initial (dis-)honesty framework concerning three significant dimensions: (i) Instead of reporting (taxable) income, participants work for a fictitious company and report surplus hours. Additionally (ii), the direction of the lie is changed: The more hours are reported, the higher is the payoff; and (iii) the group size shrinks from three to two. Given that the pretrained model is already weak, it might not be possible to express an unequivocal statement, concerning the classifier's generalizability. Nevertheless, we set the two AUCs – .597 for the first, and .529 for the second dataset – in context and conclude that the model does not generalize to another experimental behavioral setting.

In the final part of Chapter 2, we investigate various concepts potentially related to lying. We find significant correlations for risk attitudes, beliefs about others' behavior, and the experience of joy during the experiment. To proxy joy, we use a dictionary-based approach to estimate the degree of positivity associated with a chat text. This investigation allows us to find out in which direction causality points: As participants chat before reporting surplus hours, we conclude: A joyful group chat is more likely to result in a false report.

Chapter 3, as do chapter 1 and chapter 2, contributes to the behavioral experimental literature. More precisely, chapter 3 investigates public good games and thereby makes a methodological and a substantive contribution. On the methodology side, it shows in which ways clustering can be used to infer the composition of the type space. On the substantive side, it shows that existing theories about behavioral types can only explain a very narrow fraction of the data.

Based on a pilot, we select an algorithmic clustering configuration known to have excellent performance. We use multivariate clustering and feed the algorithm with pairs of experiences and choices.

One might naïvely think that the algorithm will find as many clusters as there are distinct behavioral programs. With simulation, we show why this approach must fail. We simulate all combinations of five behavioral programs that have been theorized in the literature: altruists, conditional cooperators, far-sighted freeriders, hump-shaped contributors, and short-sighted freeriders. For investigating these five behavioral programs, we need many more clusters. Yet we also show that we do not need the theoretical maximum of 350 clusters; this would approach next to unusable for real data, as one would need a considerable amount of data for that many clusters to be credible. We use internal cluster validation indices to determine that the appropriate trade-off between underusing and overusing the evidence is reached with approximately 40 clusters. Some of these clusters are indeed pure, in the sense that all combinations of experiences and choices stem from participants of the same simulated type. Yet many are not, as different behavioral problems may generate indistinguishable behavior. We do not see this as a limitation of the approach. Instead, it demonstrates what can be achieved with clustering: one is informed about distinct patterns and must discuss whether they could be generated with alternative behavioral programs.

We apply this methodology to a large dataset consisting of 16,474 observations. Results clearly show that the true type space is much more abundant than thus far assumed by the literature. Only very few of the clusters that we find in the experimental data can be rationalized with any of the five theoretical behavioral programs used to simulate data. We find practically no altruists and only very few outright selfish participants. Very few participants are near-perfectly reactive, as assumed by the canonical model of conditional cooperation. No cluster requires the assumption that players are hump-shaped, and perversely react to good experiences.

Chapter 4 is again concerned with supervised learning and text. As opposed to chapter 1 and chapter 2, it does not contribute to behavioral experimental literature but to the field of empirical legal studies. It does so in two ways: It replicates a famous paper by Landes and Posner (2011) that relates judge characteristics to the political ideology of judicial decision-making. Furthermore, it contributes to an intensely used database in the field: The Songer database provides political ideology labels only for roughly 5% of judicial opinions. By deploying supervised classification, we generate robust predictions for the remaining 95% of judicial opinions.

More precisely, in the first part, we replicate the linear regression framework proposed by Landes and Posner (2011). We add multiple robustness checks, such as aggregating the dependent variable on another level, multiway-error clustering, and an extreme bounds analysis. By doing so, we replicate the original study's most robust findings: The party of the appointing president crucially influences whether a circuit court judge's opinion is considered conservative or liberal.

In the second part, we explore multiple classification configurations to train a highly predictive classifier: We achieve the highest performance with a tf-idf weighted bag-of-words combined with a Ridge classifier. The latter is adjusted by isotonic calibration and achieves a F1 score of .67 on a binary label. To assess the robustness of both our predictions and the regression framework, we repeated the empirical analysis on the full sample, including hand-labeled and predicted labels. Results show that, again, the most robust findings of the initial regression framework still hold.

Author contributions. In the following, I provide a detailed list of how different authors contributed to the following chapters.

Chapter 1 is coauthored by Olexandr Nikolaychuk (ON). CIH acquired the funding, preregistered the experiment, conducted half of the experimental sessions, visualized the results, wrote the initial draft, and presented the project at a lab meeting at the Max Planck Institute for Research on Collective Goods (May 2018), the Young Scholar's Research Workshop in Neuchâtel (September 2018), the ESA conference in Utrecht (May 2019), and at the C-SEB gender symposium (October 2019). ON developed and tested the code for the experimental app, conducted half of the experimental sessions, and extensively reviewed the draft. CIH and ON were both equally involved in developing the hypotheses, the statistical analysis, and the experimental setup. The project was supported financially by the Max Planck Institute for Research on Collective Goods (1000€) and the Center for Social and Economic Behavior, University of Cologne (4000€).

Chapter 2 is coauthored by Martin Fochmann (MF) and Peter Mohr (PM). CIH acquired the funding, preregistered the experiment, developed the oTree code, conducted the experiment, trained the classifier, analyzed the experimental data, and wrote the initial draft. Furthermore, CIH presented the paper at the MPI thesis workshop in Wittenberg (March 2019) and the MPI lab meeting (July 2020); MF provided the training data, supervised the experimental implementation, the statistical analysis and provided critical comments to the paper. MF and PM both developed the vignette of the experimental study. The project was supported financially by the Center for Social and Economic Behavior, University of Cologne (3000€).

Chapter 3 is coauthored by Christoph Engel (CE) and Marcel H. Schubert. CIH implemented the local regression approach as proposed by CE, and visualized the results. In addition, CIH presented the project at a lab meeting of the Amsterdam Cooperation Lab at VU Amsterdam (February 2020). CIH and MHS were both equally involved in data curation and tested different configurations and data setups. CE and MHS wrote the code for simulating the data. CE identified the literature gap, formulated the research goals, and supervised the implementation of the clustering specifications. CIH and MHS provided the initial draft of the methods section. CE critically reviewed the latter and wrote the remaining parts of the paper.

Chapter 4 is coauthored by Marcel H. Schubert (MHS), and Elliott Ash (EA). CIH prepared the data used for replication, and conducted the regression analysis as well as robustness checks. Furthermore, CIH developed and implemented the original version of the code that tested the different initial classification setups. MHS was responsible for scraping and preprocessing the text data. Furthermore, he implemented the final grid search on the cluster. CIH and MHS were both equally involved in writing the original draft and presenting the paper at the PELS replication conference in Claremont (April 2019). Furthermore, CIH presented the project at the MPI lab meeting (November 2019). EA supervised the research activity, proposed analysis methods and visualizations of the results, and provided part of the data to be analyzed. He furthermore extensively reviewed the draft.

*Only time (whatever that may be) will tell.*

Stephen Hawking

# 1 Color Me Honest! Time Pressure and (Dis-)Honest Behavior

We introduce a modified version of the die-in-the-cup paradigm to study (dis-)honest behavior under time pressure. Replacing the regular die with one that has a distinct color on its either side enables us to manipulate the amount of familiarity with the randomization device. This both removes the limitations of the original paradigm and allows for a test of theories that suggest that (dis-)honest behavior is affected by the relative difficulty of generating false reports. We also replace the cup with a simple mechanical device for better control over the very process of rolling the die and collect mouse movement data from the participants to investigate the present behavioral archetypes. Our main finding is that time pressure leads to more dishonest behavior, but only if the regular die is used. We also find that when given the time to deliberate, the participants generally report lower values if the regular rather than the color die is used.

## 1.1 Introduction

Everyday life offers ample opportunities for gains through misreporting. Whether individuals take the chance or not depends on the circumstances. One popular example is that of a mother visiting an amusement park with her child. She arrives at the register and is told that children under the age of six enter for free. Her child has just turned six. What do you think, will the mother tell the truth and therefore have to pay the entrance fee, or will she claim the child to be younger than six to dodge it? Would you expect her answer to be different, had she known the rule beforehand?

Just like in the above example, this research project investigates lying behavior in situations where one has an opportunity to gain some material benefit (Tang 2012). More specifically, we are interested in the effect of time pressure on truth-telling.

Existing literature appears somewhat conflicted as studies found both positive and negative effects of time pressure on truth-telling. Most studies use the die-in-a-cup paradigm (Fischbacher and Föllmi-Heusi 2013) to measure such behavior, which is not surprising as the paradigm has certain advantages. Nevertheless, we cannot but also notice some limitations that it has.

First, there is no intrinsic value to the state of the world that the participants report in a regular die-in-a-cup scenario. When telling the experimenter what they rolled, they are effectively saying how much money they would like to receive as opposed to what happened.

We also believe that the use of the all too familiar six-sided die with pips (where progressively higher values are almost exclusively associated with higher payoffs) makes it overly easy for the participants to misreport both voluntarily and involuntarily. One needs not look at the die or even roll it. As long as one wants to maximize the own payoff, it is clear what their claim should be. It is quite common to attempt to "break the spell" by making the six a suboptimal report. Furthermore, while we agree that it may help alleviate the issue, it would be overly optimistic to assume that to be the end of the story.

With this in mind, we suggest a variation of the die-in-a-cup paradigm where the regular die is replaced with one that has a distinct color on each of its sides. This way, (i) there is a meaningful state of the world for the participants to report; and (ii) without knowing in advance the association between the colors and payoffs, there is no readily available report to fall back onto.

## 1.2 Literature

As a starting point of our analysis, we look into experimental studies from the fields of behavioral economics and experimental psychology that consider truth-telling as a function of time pressure and experience. While some find lying to be intuitive, others reach precisely the opposite conclusion. In order to make sense of such contradictory results, we discriminate among the studies on the bases of their manipulation and reporting framework.

Literature on lying tends to use the following three types of manipulation to provoke quick and intuitive reactions: cognitive load, ego depletion, and time pressure. These are followed by an individual decision situation or a game that allows the participants to lie in order to gain a monetary benefit.

Capraro (2017) induces time pressure and deploys a variant of the deception game (Gneezy 2005). In a similar context, Gunia et al. (2012) sort participants' decisions according to their response times. In the end, the two studies come to opposing conclusions: The former finds the participants to be more honest under time pressure, whereas the latter finds honesty to be positively associated with deliberation time.

Lohse, Simon, and Konrad (2018) put their participants under time pressure to report the outcome of a lottery that determines their payoff. The authors conclude that time to reflect increases one's awareness of the misreporting opportunity.

Tabatabaeian, Dale, and Duran (2015) ask the participants to predict the outcome of a virtual coin flip and find that dishonest reports are associated with shorter decision times. Houser, Vetter, and Winter (2012) ask the participants to report the outcome of an actual coin flip and find them to be more likely to lie if they receive nothing in an earlier dictator game.

Perhaps the most commonly used decision scenario in the literature on lying is the aforementioned die-in-a-cup paradigm (Fischbacher and Föllmi-Heusi 2013). Bereby-Meyer et al. (2018) and Van't Veer, Stel, and Beest (2014) use it in conjunction with cognitive load while Foerster et al. (2013) and Shalvi, Eldar, and Bereby-Meyer (2012) pair it with time pressure. The first two studies agree that cognitive load leads to truth-telling while the other two disagree as far as the observed effect of time pressure is concerned.

Shalvi, Eldar, and Bereby-Meyer (2012) conclude that time pressure results in more lying, whereas Foerster et al. (2013) find the opposite. Foerster et al. (2013) claim that the participants of Shalvi, Eldar, and Bereby-Meyer (2012) were able to come up with a lie before the start of their time pressure condition. In an attempt to tackle the issue, they augment their setup to contrast reports of individual rolls with reports of a series of rolls (with short breaks in-between).

While we agree with the general statement of Foerster et al. (2013), we believe that the issue requires a more substantial treatment. In our opinion, the regular die as a means of randomization is all too familiar to student participants. Even those who do not have first-hand experience with it are to be expected to be familiar with the concept through popular culture. Therefore, it is likely that participants' reports collected within the standard die-in-a-cup paradigm are informed by their earlier experiences outside the laboratory[1].

Numerous brain studies have shown that practice helps improve the efficiency of knowledge retrieval and response inhibition across various task domains (Brehmer, Westerberg, and Bäckman 2012; Hu, Rosenfeld, and Bodenhausen 2012; MacLeod and Dunbar 1988;

[1] We are not saying that it is reasonable to expect most to cheat when playing, e.g., a board game. It is sufficient for our argument to expect most to know what is feasible and desirable when rolling a die.

Milham et al. 2003; Olesen, Westerberg, and Klingberg 2004; Pirolli and Anderson 1985; Walczyk et al. 2009). In particular, rehearsed lies are associated with less conflict than spontaneous lies as evidenced by lower relative activity in the anterior cingulate cortex (Ganis et al. 2003).

As regards to documented behavioral effects are concerned, rehearsed lies are associated with slower reaction times than spontaneous lies (DePaulo et al. 2003; Johnson, Henkell, et al. 2008; Walczyk et al. 2009). Ganis et al. (2003) even report that they could not detect statistical difference in response times between rehearsed lies and truthful answers. It is also noteworthy that very little practice is required to alter the cognitive cost of lying (Van Bockstaele et al. 2012), and the effect carries over across various decision tasks (Hu, Rosenfeld, and Bodenhausen 2012; Van Bockstaele et al. 2012).

## 1.3 Hypotheses

Following the die-in-a-cup paradigm, we take the regular six-sided die with pips as the control treatment. As the first manipulation, we replace the regular die with a distinct color on each of its sides (hereafter, *color* die). This way, (i) there is a meaningful state of the world for the participants to report; and (ii) without knowing in advance the association between the colors and payoffs, there is no readily available report to fall back onto. As the second manipulation, we put the participants under time pressure to elicit their spontaneous reactions.

Altogether, with a 2 × 2 factorial design as summarized in Table 1.1, we aim to test the following hypotheses.

**Table 1.1:** Hypotheses and Experimental Design

| | regular die | | color die |
|---|---|---|---|
| time pressure | $R^P$ | $>$ | $C^P$ |
| | $\vee$ | | $\wedge$ |
| no pressure | $R$ | $=$ | $C$ |

Hypothesis 1 $R^P > R$.

When dealing with the *regular* die, the participants are expected to report higher values under time pressure (indicated with $^P$). As long as there is no time for deliberation, they will fall back onto the readily available reports.

Hypothesis 2 $C^P < C$.

When dealing with the *color* die, the participants are expected to report lower values under time pressure. Since there is no readily available report to fall back onto, generating a dishonest one requires more time than telling the truth.

Hypothesis 3 $R \approx C$.

Without time pressure, the participants are expected to report similar values when dealing with the *regular* and *color* die. When provided with sufficient time for deliberation, they will be (dis-)honest to the same extent regardless of the type of the die.

Hypothesis 4 $R^P > C^P$.

Under time pressure, the participants are expected to report higher values when dealing with the *regular* die than when dealing with the *color* die. Since there is no readily available report to fall back onto in the latter case, they will have to resort to telling the truth.

## 1.4 Experimental Setup

To achieve our research goals, we modify the standard die-in-a-cup paradigm (Fischbacher and Föllmi-Heusi 2013) in the following two ways. First, as we explained earlier, we replace the regular die with a color die. Second, we replace the cup with a dice tower (section A.4 provides a picture of the tower).

The dice tower is a significant improvement over the cup as it enables control over the quality and duration of the randomization phase. With the cup, both are at the discretion of the participant who can, e.g., cheat by producing ineffective lateral movements or violate the time pressure manipulation by taking their time to shake the cup. With the dice tower, one needs only to tip the die over the ledge, which ensures proper randomization and stable timings across the participants.

The experiment is comprised of four stages. Stages 1, 3 and 4 are executed with the help of computer terminals using oTree (Chen, Schonger, and Wickens 2016) (section A.4 provides screenshots). Stage 2 (rolling the die) is executed in the physical space and synchronized across the participants. All participants are divided into four groups, according to Table 1.1.

Before each session, the requisite die[2] is placed on top of the dice tower and covered with an opaque paper lid in order to rule out potential priming.

In stage 1, the participants receive general instructions on paper, and treatment specific instructions on the computer screen. When everyone is ready, stage 2 begins where the participants remove the lid and tip the die over the ledge following a five-second countdown.

In stage 3, the participants are presented with six buttons arranged in a circular pattern and occupying most of the screen real estate. Depending on the treatment condition, each

[2] Color: green on top, gray facing the participant; regular: one on top, two facing the participant

button corresponds to one of the six colors or one of the six faces of a regular die[3], and the associated payoff is revealed as long as the mouse cursor is hovering over it. The participants submit their report by clicking any of the six buttons. In the background, their decision process is being tracked through the cursor movements (hereafter, mouse movements).

In the time pressure conditions, the participants have a limited amount of time to report their die roll. Following Dana, Weber, and Kuang (2007), they are not given a precise cutoff point but rather an interval. Based on a pilot session with 12 participants, we determined the optimal interval to be between 6 and 12 seconds.

In stage 4, the participants are asked to provide answers to three blocks of questions. First, we do manipulation checks (where applicable) by asking if the participants can clearly distinguish between all six colors used in the experiment, if they have a favorite color, if they can recall the payoff associated with the color they reported (incentivized), if they felt time pressure, and if they felt some general pressure during the experiment. The second block contains three problems of the Cognitive Reflection Test (Frederick 2005) as a crude measure of cognitive ability. In the final block, the participants are asked to fill out a basic demographic questionnaire as well as to report on their prior experience with laboratory experiments and games involving dice.

## 1.5 Results

The experiment was conducted in the economics laboratory of the Friedrich Schiller University of Jena in November 2018 and in the MPI decision lab in Bonn in June 2019. Overall, we collected 234 observations. We excluded color-blind participants and those who failed to report on time (in time pressure conditions). The final sample contains 229 observations, of which 61% are females; 87% are business administration and economics students. The average age is 24.4 years (SD 7.68).

**Table 1.2:** Means by Treatment

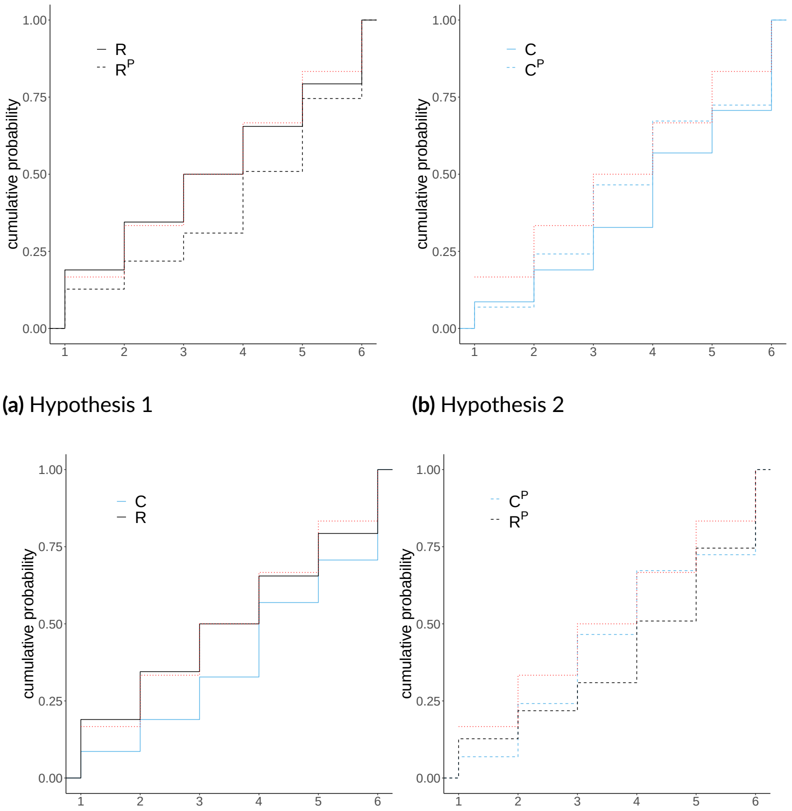
| | $C^P$ | $C$ | $R^P$ | $R_S^P$ | $R$ | $R_S$ |
|---|---|---|---|---|---|---|
| means | 3.09 | 3.33 | 4.09 | 4.38 | 3.52 | 3.42 |
| SD | 1.84 | 1.57 | 1.70 | 1.50 | 1.81 | 1.84 |
| p-value | | | | 0.21 | | 0.68 |

Two-sided p-values compare our means ($R$ and $R^P$) to those ones by Shalvi, Eldar, and Bereby-Meyer (2012) ($R_S$ and $R_S^P$).

Table 1.2 provides a basic overview of the data by summarizing mean reports (with SD) across the treatment conditions. It also contains the results of Shalvi, Eldar, and Bereby-Meyer (2012) that can be directly compared to ours when the regular die is being used. We

[3] Particular arrangement randomized across the participants.

observe the mean reports of 4.09 (1.7) and 3.52 (1.81) relative to their 4.38 (1.5) and 3.42 (1.84) with and without time pressure, respectively. These differences are not statistically significant at the 10% level (t-test, two-sided p-values of .21 and .68, respectively), and we therefore conclude that our sample is not qualitatively different from those usually found in the literature.



**(a)** Hypothesis 1

**(b)** Hypothesis 2

**(c)** Hypothesis 3

**(d)** Hypothesis 4

**Figure 1.1:** Empirical Cumulative Distribution Functions by Treatment
The dashed red line denotes the full honesty benchmark. The color blue denotes rolling the color die. The dashed lines in blue and black denote the time pressure conditions.

Even though mean comparison can be useful as a quick measure of differences across the treatment conditions, we do not regard it as a sufficient statistic of interest for this type of data. Instead, we consider Figure 1.1 providing the empirical distribution functions for each condition grouped according to Hypotheses 1–4.

As one can see, the empirical distribution function in condition $R$ stochastically dominates that one in condition $R^P$. This means that when dealing with the regular die, the participants tend to provide higher reports under time pressure. They also tend to report higher dice rolls relative to the reference uniform density of telling the truth (the dashed red line). Both of these findings are consistent with Hypothesis 1.

If the color die is used instead, the opposite appears to be the case as the empirical distribution function in condition $C$ stochastically dominates that one in condition $C^P$[4]. The participants tend to report lower values under time pressure, and even though they still over-report relative to the truth-telling benchmark, they do so to a smaller extent. This behavior is consistent with Hypothesis 2.

When we compare the reports without time pressure, the empirical distribution function in condition $C$ stochastically dominates one in condition $R$. In reference to the truth-telling benchmark, it appears that the participants hardly lie on the aggregate level when dealing with the regular die but tend to over-report when dealing with the color die instead. This finding is not consistent with Hypothesis 3 and we explore it further using the collected mouse tracking data in the next section.

With regards to the empirical distribution functions in conditions $C^P$ and $R^P$, no obvious pattern emerges there. If anything, it appears that under time pressure, both participant groups tend to over-report relative to the truth-telling benchmark but in somewhat different ways. The regular die report distribution is simply skewed to the right, whereas in the color die report distribution, a sizable portion of the probability mass is associated with the reports of three and four while five is considerably underreported. As far as Hypothesis 4 is concerned, we deem these findings inconclusive.

Following the visual analysis, we investigate the hypotheses within a Bayesian framework. Since the die roll report has an inherent ordering as a dependent variable, we opt for the ordered logit as the general framework for the statistical analysis of the results. We implement the estimation with normally distributed, vague priors ($\sim \mathcal{N}(0, .0001)$), sampling from two separate chains, where $10,000$ samples of each chain were used for adaption and $100,000$ samples were used for burnin. After the burnin, we collect $100,000$ samples from each chain. Furthermore, we use a step wise algorithm based on the Akaike information criterion (AIC) to determine the necessary set of control variables. Equation (1.1) denotes the final specification:

$$report_i^* = \beta_1 \cdot R_i^P + \beta_2 \cdot C_i + \beta_3 \cdot P_i \cdot C_i + \beta_4 \cdot EHI_i + \beta_5 \cdot ECON_i + \beta_6 \cdot PED_i + \varepsilon_i \quad (1.1)$$

where $i$ indexes the participant, and $report_i^*$ corresponds to the latent variable. $R_i^P$ and $C_i$ equals one if the participant is assigned to the treatment condition with time pressure and a color die, respectively; $EHI_i$, $ECON_i$ and $PED_i$ each equal one if the participant reports experience with laboratory experiments, studies economics, or studies psychology or education, respectively.

[4] Except for the observed frequencies of reporting the lowest roll, which are very close regardless.

**Table 1.3:** Ordered Logit Regression within a Bayesian Framework

| | Odds Ratio | Estimate | [HPD 95%] | ESS | BF (>0) | psrf |
|---|---|---|---|---|---|---|
| $R^P$ | 1.89 | 0.64 | [-0.04,1.3] | 188 | 29.78 | 1.01 |
| $C$ | 1.94 | 0.66 | [-0.01,1.32] | 163 | 35.60 | 1.01 |
| $P \cdot C$ | 0.41 | -0.90 | [-1.82,0.02] | 594 | 32.04 | 1.00 |
| H2 | 0.77 | -0.26 | [-0.92,0.39] | | 3.65 | |
| H4 | 0.79 | -0.24 | [-0.91,0.42] | | 3.16 | |
| EHI | 1.86 | 0.62 | [0.12,1.1] | 215 | 140.44 | 1.00 |
| ECON | 2.14 | 0.76 | [0,1.53] | 1732 | 40.60 | 1.00 |
| PED | 1.72 | 0.54 | [-0.16,1.25] | 607 | 14.15 | 1.00 |

$R^P$: treatment no color, pressure; $C$: treatment color; $P \cdot C$: treatment color, pressure; H2: tests Hypothesis 2; H4: tests Hypothesis 4; EHI: ample experience with laboratory experiments; ECON: field of study is economics; PED: field of study is psychology or education; HPD: highest posterior density; ESS: effective sample size; BF: Bayes factor; psrf: potential scale reduction factor.

Table 1.3 provides the coefficient estimates, as well as estimates for Hypothesis 2 and Hypothesis 4. H2 denotes testing the hypothesis $\beta_1 + \beta_3 < 0$; H4 denotes testing the hypothesis $\beta_2 + \beta_3 < 0$. As H2 and H4 are not estimated by Equation 1.1 directly, neither the effective sample size (ESS), nor the potential scale reduction factor (psrf) are provided.

$\beta_1$ ($R_i^P$) tests Hypothesis 1. Its Bayes Factor is 22.89 which, according to Jeffreys (1998)'s scale of interpretation, provides *strong* support for the hypothesis that rolling the regular die under time pressure increases reports as opposed to not being under time pressure.

$\beta_2$ ($C_i$) tests Hypothesis 3. Its Bayes Factor is 27.74, providing *strong* support for the claim that when not under time pressure, rolling the regular die decreases reports as opposed to rolling the color die. We, however, hypothesized not to find any significant difference between these two treatment groups. Therefore, section 1.6 further investigates this unexpected finding.

Hypothesis 2 cannot be investigated directly by Equation 1.1. Instead, $\beta_1 + \beta_3 < 0$ needs to be tested. The resulting Bayes Factor is 3.65, providing *substantial* support for the claim that rolling the color die, participants report lower results under time pressure than when not under time pressure.

Similarly, Hypothesis 4 cannot be investigated directly by Equation 1.1. Instead, $\beta_2 + \beta_3 < 0$ needs to be tested. The resulting Bayes Factor is 3.16, providing *substantial* support for the claim that when under time pressure, rolling the regular die results in higher reports than when rolling the color die.

As far as the control variables are concerned, the following results were achieved: The Bayes Factor for $\beta_4$ ($EHI_i$) is 115.69, providing *decisive* support that participants having taken part in multiple experimental studies report higher results. The Bayes Factor for $\beta_5$ ($ECON_i$) is 34.77, providing *very strong* support that participants studying economics report higher results. Finally, the Bayes Factor for $\beta_6$ ($PED_i$) is 12.18, providing *strong* support that participants studying either psychology or education report higher results.

To summarize, an ordered logit regression within a Bayesian framework confirms three out of four hypothesis. More precisely, we find *strong* support for Hypothesis 1, *substantial* support for Hypothesis 2, and *substantial* support for Hypothesis 4. Hypothesis 3 was not confirmed, instead we find *strong* support against it. We initially hypothesized no difference in means for conditions $R$ and $C$; however, it turned out that rolling the regular die leads to lower reports than rolling the color die when not under time pressure. To investigate this unexpected result, we analyze participants' mouse movements in section 1.6.

## 1.6 Mouse Movements

To report the number or color rolled, participants needed to click on one of six choice options on the computer screen. On this screen, mouse movements were tracked to gain a more nuanced explanation of the decision making process.[5]

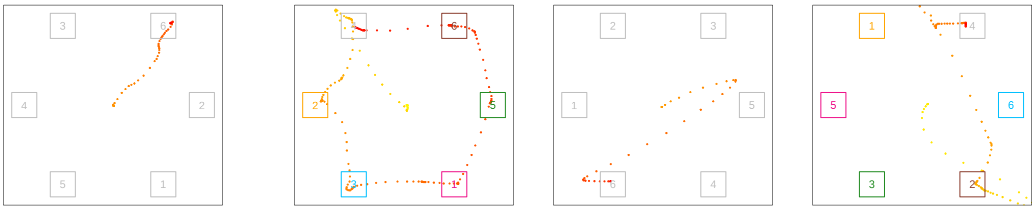
Recent models show that even before choosing a particular target, participants can hold several movement plans in their mind and display a movement according to their average (Alonso-Diaz, Cantlon, and Piantadosi 2018; Dotan, Meyniel, and Dehaene 2018; Erb et al. 2016; Friedman, Brown, and Finkbeiner 2013; Pinheiro-Chagas et al. 2017). This idea that movement can start even before the final decision is why we track participants' mouse movements on the decision screen: The trajectories can provide a valuable insight into a participant's decision making process.

We analyzed the mouse movements by plotting trajectories by participants and categorizing them concerning the movement's start- and endpoint. By that procedure, we identified four types, of which Figure 1.2 shows representative trajectories.

The A type explores only one option and clicks on it. The majority of participants, 75.43%, belong to this category. The $A = \Omega$ type (9.48%) explores several choice options. However, this type ultimately reports the option which it explores first. The $A < \Omega$ type (11.64%) explores several choice options. The option reported is greater than the option explored first. The $A > \Omega$ type (3.45%) again explores several choice options. The option reported is smaller than the option explored first.

Before interpreting the four types by treatment – Table A.3 shows a summary – the reader is asked to remember that participants in the color treatment need to hover over a choice option to reveal the associated payoff. By contrast, the payoff in conditions $R$ and $R^P$ was identical to the number of pips on the choice option displayed. Consequently, if participants in this treatment want to maximize their payoff instead of reporting the number rolled, they can directly go to the desired option without taking a detour.

[5] Mouse movements are unlikely to be driven by other factors than the payoff related to the color. Section A.2 provides more details concerning this statement.



**(a)** A **(b)** $A = \Omega$ **(c)** $A < \Omega$ **(d)** $A > \Omega$

**Figure 1.2:** Behavioral Archetypes Found

The majority of A types rolled the regular die. These participants could be either payoff maximizers or reporting the number they rolled. By contrast, A types rolling the color die are unlikely to be payoff maximizers: 5/6 of those would need to take a detour to find the highest payoff and therefore, would not be an A type. The most likely interpretation is that these participants indeed report the number they rolled.

The majority of $A = \Omega$ types rolled the color die. Two possible interpretations could explain their detour: curiosity and temptation. In the open answer field, many participants stated that they were curious or wanted to check whether all six payoff options were available. Another interpretation is that participants went to the face they rolled, subsequently were tempted to maximize payoff but ultimately go back to their initial choice. For participants rolling the regular die, possible interpretations are temptation or guilt: In the same manner, as participants rolling the color die, they could go to the number they rolled initially, get tempted to maximize their payoff, but ultimately decide to report the number rolled. The thought process could also go the other way round: $A = \Omega$ types rolling a regular die could have chosen the payoff maximizing face first, felt guilty, and thought about reporting their actual role, but in the end, they decided for the payoff maximizing strategy.

The interpretation for $A < \Omega$ types is similar, regardless of whether the participant rolled the regular or the color die. The most reasonable explanation for this behavior is payoff maximization: Participants move the mouse cursor to the option they rolled first but finally reported a more lucrative option.

$A > \Omega$ types, rolling a regular die, are likely to be motivated by a feeling of guilt: Those participants navigate to the payoff maximizing option first, feel guilty, and, finally, the feeling of guilt wins. Therefore, they chose the face they had rolled. The most reasonable explanation in the case of the color die is that they randomly picked one color, felt guilty, and chose the number rolled in the end.

**Table 1.4:** Participants in Treatments $R$ and $C$ by Behavioral Archetype

| | $R$ | $C$ |
|---|---|---|
| A | 51 (87.93%) | 34 (58.62%) |
| $A = \Omega$ | 1 (1.72%) | 14 (24.14%) |
| $A < \Omega$ | 3 (5.17%) | 9 (15.5%) |
| $A > \Omega$ | 3 (5.17%) | 1 (1.72%) |

Absolute counts by type, percentages in parentheses.

### 1.6.1 Interpreting the Findings Regarding Hypothesis 3

The categorizations and interpretations outlined above help explain the puzzling findings regarding Hypothesis 3: When not under time pressure, participants rolling the color die ($C$) significantly report higher numbers than participants rolling the regular die ($R$). Table 1.4 counts participants by type. It clearly shows that most participants rolling a regular die belong to category A. By comparison, more participants rolling the color die belong either to the $A = \Omega$ or $A < \Omega$ type. A Fisher's exact test confirms these intuitions: The behavioral archetype is not independent of treatments (p-value = $8.07 \cdot 10^{-5}$).

How does the analysis of the behavioral archetypes connect to the puzzling findings concerning Hypothesis 3? If we assume that average reports are similar for types A and $A = \Omega$ then the difference in means concerning treatments $R$ and $C$ is mostly driven by $A < \Omega$ types. By definition, those types report higher numbers. As treatment $C$ includes three times as many $A < \Omega$ types as treatment $R$, we expect that it is mainly this type driving the difference in mean reports.

The question left is why are there more $A < \Omega$ types in treatment $C$ than $R$? The explanation is found in participants' statements: When asked to explain their behavior, many participants in treatment $C$ reported that they had been just curious which numbers were behind the different colors. $A = \Omega$ and the $A < \Omega$ types reflect this intention. While $A = \Omega$ types report honestly, $A < \Omega$ types most likely do not. The $A < \Omega$ type, however, less likely reflects a dishonest intuition and rather the giving in to temptation: Given participants' explanations as well as the high share of $A = \Omega$ types in the same treatment, we interpret that $A < \Omega$ started with innocent curiosity but discovered a lucrative choice option on the way. It was too hard to move on, and finally, they gave in to the temptation. In other words, treatment $C$ might have seduced participants to higher reports even if they started with honest intentions.

### 1.6.2 Types Proposed by Fischbacher and Föllmi-Heusi (2013)

We are not the only ones who introduced typing to analyze treatment effects. Fischbacher and Föllmi-Heusi (2013) introduced the types "honest", "partial liars" and "income maximizers". They estimate each type's shares based on the rules of probability, assuming uniformly distributed reports. Fischbacher and Föllmi-Heusi (2013) interpret these estimates as upper bounds.

Our mouse-tracking data allows us to precisely determine the share of the types proposed by Fischbacher and Föllmi-Heusi (2013). Honest subjects are either A or A $=$ $\Omega$ types. Partial liars are A $<$ $\Omega$ types, reporting a value smaller than 6. Income maximizers are A $<$ $\Omega$ type, reporting a value $=$ 6. The reader is asked to keep in mind that the above mapping is only possible for the participants in the color treatment. Only in this treatment, we can be sure that mouse movements directly map to the thinking process. Table 1.5 reports shares, either estimated by the rules of probability (first two columns) or by visual inspection (last two columns).

Table 1.5 shows that the share of income maximizers is higher when estimated with probabilities than with mouse tracking. Upper bound estimations according to Fischbacher and Föllmi-Heusi (2013) for honest participants and partial liars are therefore not correct. The comparison implies that participants should not be assigned to Fischbacher and Föllmi-Heusi (2013) types based on the rules of probability. Instead, mouse tracking allows for an adequate allocation of the type space.

**Table 1.5:** Types Proposed by Fischbacher and Föllmi-Heusi (2013)

| | $C^P$ | $C$ | $C^P_m$ | $C_m$ |
|---|---|---|---|---|
| honest reporters | 41.38 | 51.72 | 88.33 | 86.21 |
| partial liars | 0.00 | 0.00 | 5.00 | 3.45 |
| income maximizers | 13.10 | 15.17 | 6.67 | 10.34 |

$C^P$ denotes the treatment where participants roll dice with colors under time pressure. $_m$ denotes counts based on mouse movements. The absence of $_m$ denotes counts based on probability.

## 1.7 Conclusion

In this paper, we investigate lying behavior in situations where opportunities to gain material benefits are available. To this end, we implement a 2 × 2 experimental design allowing for manipulations of the amount of time available to the participants and the familiarity with the randomization device.

As the core of the framework, we use a modified version of the die-in-the-cup paradigm, where the regular die with pips is replaced with one that has a distinct color on its either side. Besides, we replace the cup with a dice tower, which allows for better control over the very process of rolling the die.

Our contribution is two-fold: First, we add to the discussion of the effect of time pressure on (dis-)honest behavior. Second, we improve the existing methodology in the field by allowing for more abundant and realistic scenarios to be implemented in the lab.

Our immediate findings suggest that time pressure leads to more dishonest reports, but only if the regular die is used. This finding is in line with theories that suggest that such action is affected by the relative difficulty of generating false reports. On a more intuitive note, our findings suggest that when faced with a new situation, one's immediate reaction is to tell the truth. However, when one has lied repeatedly in a specific situation, and when a cue triggers this behavior, e.g., a regular die, it is most natural to draw upon this very default when under time pressure.

Furthermore, when given the time to deliberate, the participants generally report lower values when dealing with regular dice than with color dice. This finding was not hypothesized but could ex-post be rationalized by consulting the mouse trajectories collected on the decision screen: It seems as if the colored choice options sparked the innocent curiosity to explore multiple options. During this search, also lucrative choice options were discovered. Once the temptation was found, it was hard to resist. However, at this stage, this interpretation is merely speculation; further research needs to confirm this hypothesis.

On a more general note, the mouse-trajectories suggest various behavioral archetypes concerning (dis-)honest behavior: The A type makes no detour and directly reports once choice. The $A = \Omega$ type explores all options but reports the options on which the search started. The $A > \Omega$ trajectory started on a more lucrative choice than that one chosen, and the $A < \Omega$ trajectory started on a less lucrative choice option than that one chosen. Especially these last two trajectory-types suggest that lying and truth can both be default behaviors.

The trajectories provide dimensions of a decision that are not present in a one-pointed report. We believe that mouse-tracking data can enrich the insights gained from decisions made in various contexts and provides fruitful research opportunities. For example, covering the payoff associated with one's choice by colors, as we did it, paired with tracking mouse movements makes an investigation of search patterns about lying under time pressure possible.

Outside of the laboratory, based on our findings, we would advise policymakers to construct a novel situation, if honest decisions are crucial for the public good. Another intuitive conclusion is that if one wants to be more self-disciplined about personal values such as honest decision-making, one should pay close attention to how to construct the environment: Tempting and immoral choices should be made hard to reach.

*Human: What do we want?!*
*Computer: Natural Language Processing!*
*Human: When do we want it!?*
*Computer: When do we want what?*

Posted by sachintripathi007 on Reddit

# 2

# Predicting (Dis-)Honesty: Leveraging Text Classification for Behavioral Experimental Research

Many laboratory experiments in experimental behavioral research offer participants the opportunity to chat with each other. The chat is often directly related to a numeric variable, e.g., the amount of money sent by a dictator. This kind of data can be leveraged for supervised classification, where text serves as input and the numeric report as a label. However, experimental data have different properties – such as being small or heavily skewed – as compared to datasets typically used for text classification. This paper systematically investigates various classification setups to test whether supervised learning can at all classify chat texts obtained by experimental behavioral research. More specifically, we train a classifier to predict whether a group reported (taxable) income honestly, based on chat texts obtained from a tax evasion experiment. The classifier's generalizability is tested on data obtained from a new experiment where participants play in a different setup. Results show that the predictive performance of the best model is rather weak (F1 score = .411). This result was expected as the dataset is tiny and heavily skewed. However, we successfully show that a careful selection of the classification configuration doubled the weakest setup's performance. Furthermore, we found that the pretrained classifier does not generalize to another context. This result was most likely driven by an already weak model, as well as structurally different communication in the new experimental setting.

## 2.1 Introduction

In many behavioral experiments, (numeric) decision data is collected alongside process data, such as texts generated through group chats. Such data can be interpreted as labeled data, where the chat text serves as input and the numeric decision as the label. This property makes experimental data an exciting candidate for supervised learning, as real-world text data, by contrast, rarely possess this characteristic.

Even though having gold-standard labeled data available, behavioral research does not exploit this property. In this discipline, text data is – as by now – exclusively used in the line with process data. Capra (2019), Elten and Penczynski (2020), Fochmann et al. (2019), and Kocher, Schudy, and Spantig (2018) assign labels derived from theoretical reasoning, to the obtained texts by hand. Andres, Bruttel, and Friedrichsen (2019) and Capra (2019) construct word-clouds with the chat texts. Both hand-assigned labels and word-clouds are used to distinguish treatment groups.

Arad and Penczynski (2018), Burchardi and Penczynski (2014), Georgalos and Hey (2019), and Penczynski (2019) use (semi-)supervised learning to assign labels to text. These authors are among the rare ones in the discipline to leverage machine learning. Their approach is different from ours because they assign labels by hand and train an algorithm on this association. They do not connect decision with process data, as this paper does.

Furthermore, this paper systematically tests multiple configurations of classification setups. Burchardi and Penczynski (2014), Georgalos and Hey (2019), and Penczynski (2019), and Arad and Penczynski (2018) each just use one configuration: They combine a bag of words approach with a linear classifier. By contrast, our paper aims to give a well-researched recommendation for using a classification setup specifically tailored to experimentally collected chat data.

Such a systematic review of techniques is necessary because supervised classification might fail because this kind of data often shows problematic characteristics such as small size and an imbalanced share of labels. More specifically, this paper uses supervised machine learning to classify group chats, which resulted in honest or dishonest reports. Subsequently, the classifier is tested on a new, unseen dataset to evaluate its predictive potential in a different experimental setting.

## 2.2 Model Architecture

Text classification is a prevalent task in Natural Language Processing. Therefore, literature provides an exhaustive palette of feature engineering techniques and classifiers. However, not all of these methods are equally well suited for the type of data we plan to investigate. Additionally, experimental behavioral data has characteristics that can be exploited to increase the classification's performance. The following section describes the multiple configurations tested to train a classifier that learns the association between process and decision data.

### 2.2.1 X and y variations

The experimental behavioral data has characteristics that can be exploited in order to increase the classification's performance. In experimental games, such as a public good game, a group chat precedes the actual (numeric) decision. The subsequent decision is taken individually, e.g., each group member makes an individual contribution to a shared public good. Consequently, the numerical decision could be predicted by the group chat or an individual's chat messages only. As stated in the literature, many scientists in this field assign labels or categories manually to the text data. These labels can be exploited for predictions, too. Often, there are chat snippets to which no label is assigned because these are filler sentences, which are commonly arising in chat communications. Excluding them from the input data could reduce noise and therefore increase accuracy. Furthermore, spelling mistakes are common in user-generated content. Thus, the variations proposed above can be implemented on text with and without spellchecking.

A decision task most often follows the communication phase. The numeric decision is context-dependent; however, in most cases, this decision can be mapped to a broader, binary concept, such as cooperativeness in public good games. Setting this threshold to form categories of a previously continuous variable can be theorized or derived by data inspection. Through the lens of a classifier, is communication in the case of full cooperation structurally different than in the case of partial cooperation? A systematic comparison of classification results based on various thresholds can answer this question. In other words, the threshold chosen by this procedure defines the labels such that the classifier can easily distinguish them. We want to highlight that the chosen threshold is not "better" than others, it is just an "easier" threshold for the classifier to work with.

### 2.2.2 Embeddings

Classification algorithms only work with numerical input. Therefore, words need to be transformed into numbers. A simple and efficient baseline for text classification is to represent sentences as bag of words or bag of n-grams (Harris 1954). We implement a bag of words approach either based on absolute counts or tf-idf weighted counts.

However, treating words as atomic units has many disadvantages, such as that there is no notion of similarity between words. A possible remedy is to use vector representations that can preserve the meaning of words. In the following, we focus on static embeddings because they require fewer data than dynamic embeddings. The three most widely known algorithms to train word embeddings are Word2Vec (Mikolov et al. 2013), GloVe (Pennington, Socher, and Manning 2014), and fastText (Joulin et al. 2017).

Word2Vec trains a neural network based on texts. The resulting embedding captures whether words appear in similar contexts. GloVe focuses on word co-occurrences over the whole corpus. Its embeddings relate to the probabilities that two words appear together. The method builds on the idea that semantic relationships between words can be derived from the co-occurrence matrix (Pennington, Socher, and Manning 2014). fastText improves on Word2Vec by taking word parts into account, too. This trick enables the training of embeddings on smaller datasets, generalization to unknown words, and to capture partial information about the local word order (Joulin et al. 2017).

Besides training word embeddings on the actual corpus at hand, we deploy pretrained embeddings: Those vectors were trained on an external, much bigger corpus. This method seems especially useful because text data obtained by laboratory experiments are limited in size.

Additionally, we combine word embeddings to form sentence embeddings: Directly averaging all word embeddings that occurred in the text has proven to be a stable baseline across multiple tasks (Banea et al. 2015; Hu, Lu, et al. 2014; Socher et al. 2013). However, means could be a relatively weak way to describe the distribution of word embeddings across a text. To weight features concerning saliency, we apply tf-idf weighting of the individual feature vectors, as proposed by Kenter and De Rijke (2015).

Instead of averaging word embeddings over the sequence, we additionally train paragraph vectors directly. More specifically, we use document embeddings and distributed memory (PV-DM) (Le and Mikolov 2014) instead of distributed bag of words. The central idea is to randomly sample adjacent words from a paragraph and predict a center word from the randomly sampled set of words by taking the context words and a paragraph id as input.

### 2.2.3 Classifiers

Logistic regression, support vector machines (Joachims 1998), or Naïve Bayes (Zhang 2004) are considered as efficient baselines for text classification. Furthermore, we test one non parametric model, namely k-nearest neighbors (Sun and Chen 2011), and non-linear models like Random Forests (Breiman 1998, 2001) and XGBoost (Chen, Schonger, and Wickens 2016). We also test ensemble techniques such as bootstrap aggregating (Breiman 1996), short boosting, and model stacking (Wolpert 1992). Another model tested is a 2-layer perceptron.

The following paragraph provides a brief explanation of the classifiers deployed: A support vector machine finds the decision boundary to separate different classes by maximizing the margin. A Naïve Bayes classifier takes advantage of probability theory and the Bayes' theorem: It calculates the probability of each class for a given text and then outputs the class with the highest score. K-nearest neighbor classification is based on a majority vote: A text is assigned to the class with the most representatives within the nearest neighbors of the point representing the text in space. A random forest classifier consists of multiple individual decision trees. Each tree predicts the class of a given text, and the class with the most votes becomes the model's prediction for a given text. XGBoost denotes a specific implementation of gradient boosted decision trees designed for speed and performance. Bagging is an ensemble technique that combines the results of multiple classifiers trained on different subsamples of the same data set. The technique reduces the variance of predictions. Model stacking is another ensemble learning technique which combines the predictions of several base models: A meta-algorithm makes the final prediction based on the predictions of multiple base models. Each model's votes are weighted by a certain weight to derive the final prediction. Stacked ensembles tend to outperform the individual base models. The two-layer perceptron is a simple feedforward neural network, composed of an input layer that receives the text and an output layer that predicts the class; in between is one hidden layer.

### 2.2.4 Evaluation

To evaluate the performance of the X-, and y- variations, the embeddings, and the classifiers, standard evaluation metrics, such as accuracy, precision, recall, F1 score, and AUC are reported.

The accuracy denotes the fraction of correct predictions. However, it is not suitable to evaluate a classifier's performance on imbalanced data: If the classifier exclusively predicts the majority label, the accuracy is high, but the classifier performs poorly.

To evaluate an imbalanced dataset like ours, the F1 score is the preferred metric. The F1 score is the harmonic mean of precision and recall. For binary classification, precision quantifies how many of the texts that are predicted positive are positive[1]. Recall calculates how many of the actual positives are labeled as positive.

The F1 score is the preferred metric to evaluate a model's performance *within* one dataset. When comparing models *across* datasets, however, an independent prevalence metric needs to be consulted. The F1 score is solely interested in the performance of the positive class. Therefore, it is sensitive to different class distributions when comparing across datasets. By contrast, AUC is prevalence independent because this measure is built from a separate evaluation of the two classes (for an excellent explanation see Straube and Krell 2014). AUC is short for Area Under the ROC Curve (AUC). The ROC curve (receiver operating characteristic curve) plots the false positive rate on the x-axis and the true positive rate on the y-axis. By doing so, it shows a model's performance at all thresholds. Consequently, AUC $\in \{0, 1\}$ measures the two-dimensional area underneath the ROC curve. The measure equals 1 if the model predicts correctly. It equals 0 if the model predicts the inverted ground truth. If the AUC equals .5, the model is not able to effectively separate classes. AUC is our metric of choice when comparing a model's performance across different thresholds to binarize y, and across two experimental datasets.

To evaluate the embedding techniques, we additionally consult a visual representation. The trained embeddings should capture the meaning of a given word. For this purpose, the embeddings are plotted in a two-dimensional space. Before doing so, the dimensionality of the embeddings needs to be reduced, from 300, which is the standard dimension size in

[1] For the classification of imbalanced datasets, it is common to associate the minority label – which are the truth-tellers in our case – as the positive class (1).

literature, to two. For that, we use principal component analysis (PCA) (Halko, Martinsson, and Tropp 2011). PCA performs linear mapping in such a way that the variance of the data in the low-dimensional representation is maximized.

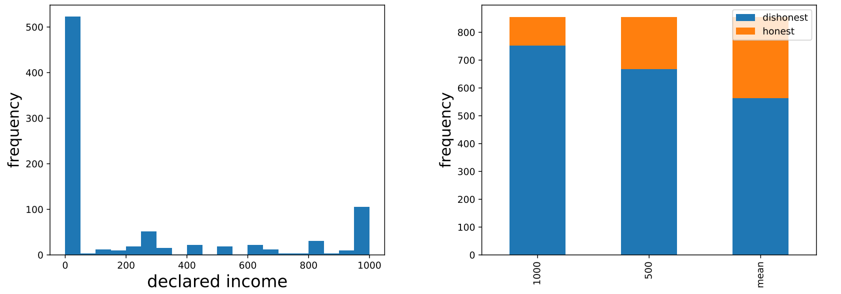
## 2.3 Experiments

We evaluate the proposed configurations on an experimental behavioral dataset. The configuration that reaches the highest F1 score is used to train the final model. The generalizability of the resulting model is subsequently assessed on a second, unseen behavioral experimental dataset.

### 2.3.1 Finding the Best Configuration

Dataset. Fochmann et al. (2019) provides the dataset on which the configurations are trained and tested. In the context of a behavioral experiment, participants play a tax evasion game where the vignette states the (taxable) income they earned. Participants discuss whether to report their income truthfully or not in a group chat. Subsequently, they individually state their income. The dataset includes 141 subjects within 47 groups, taking overall 855 decisions.

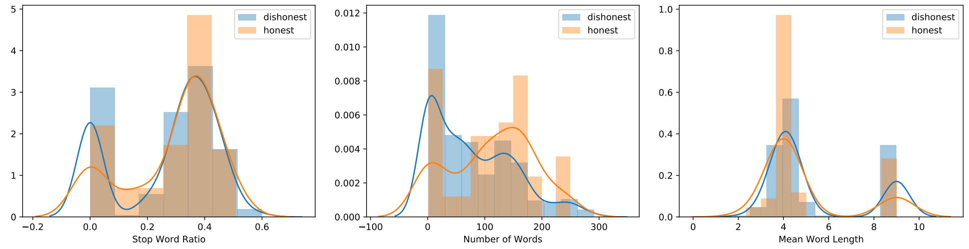
The label, or y, to be predicted is the income stated by participants. This paper aims to predict the broader concept of truth-telling and not specific levels of the income stated. Whether the stated income is classified as truth or not depends on the threshold applied. Different settings might require a different threshold; however, for the purpose of this paper, the threshold is exclusively chosen based on performance criteria. The full honesty benchmark was to report 1000 tokens. Consequently, all reports from 0 to 999 can be called a lie. Other thresholds tested are the mean (317), and half of the endowment (500).



**(a)** Distribution of the Declared Income **(b)** Categories based on Specific Thresholds

**Figure 2.1:** Taxable Income Stated

Figure 2.2 plots three different indicators that describe a group's communication. Each subfigure plots two distributions: The orange distribution reflects honest, the blue one



**(a)** Stop-Word Ratio

**(b)** Number of Words

**(c)** Mean Word Length

**Figure 2.2:** Distribution of Three Indicators by Label on Experimental Chat Data

dishonest communication. To achieve this classification, the dependent variable – the reported income – was mapped from a continuous to a binary scale based on the full (dis-)honesty threshold. Figure 2.2a plots the stop word ratio's distribution. Both distributions are bimodal, with one peak around .1 and another peak at .4. The peak at .1 is less pronounced for honest group chats indicating that these texts include fewer stop words than texts classified as dishonest. Figure 2.2b plots the distribution of the number of words per group chat. Both distributions show a noticeable peak at the bin, indicating 0 to 20 words. The suspicious reader might ask: Why is a considerable share of the group chats so short? The answer lies in the structure of the experimental setup. Depending on the treatment, participants stay in the same group for up to nine rounds. Very often, there is extensive discussion at the beginning of the experiment; but in later rounds, participants coordinate on reporting the same number as in previous rounds. Apart from this peaking first bin, the two distributions allocate their masses differently: Group chats classified as dishonest include fewer words than as honest classified group chats. This shift in distributions could be interpreted as coordinating on an honest requires more discussion and seems less intuitive than coordinating on a dishonest answer. Figure 2.2c plots the distribution of the mean word length per group chat. The distributions are again bimodal, both peaking at 4 and 8.5 words. Mean word length is very similar for both honest and dishonest communication, suggesting that the complexity and structure of the language used are quite similar across both labels. Overall, Figure 2.2 provides a qualitative interpretation of the text data. It is particularly interesting to compare this figure to chat data collected within a new experimental setup in another part of this paper. The comparison shows how communication structurally differs across contexts.

Apart from providing chat text, Fochmann et al. (2019) furthermore provide 34 labels which the authors assigned to the chat after reading its content. These labels define the chat, for example, as a concrete money proposition or lying strategy. Section B.1 provides a complete list of these labels. To each chat, no label, one label or multiple labels could be assigned. To exploit this information, we train variations of the classifiers only on these labels or based on chat texts to which labels were assigned. The later variation excludes chats.

The chat-based configurations are built either on standard preprocessed or spellchecked and then preprocessed text. The spell checker implemented uses the Levenshtein Distance (Levenshtein 1966) to find permutations within an edit distance of two from the original word. It then compares all permutations to words in a word frequency list. Those words that appear more often in the list are more likely correct (Norvig 2007). Overall, the spell checker corrected 599 out of 3474 (17.24%) unique words in the text.

All text-based variations are preprocessed using the package spaCy (Honnibal and Montani 2017). First, the text is tokenized, and punctuation and stopwords are removed. Then, depending on the feature representation, bi-grams[2] or uni-grams are built. Finally, all tokens are lemmatized[3]. We remove the words occurring less than five times, which results in 2, 678 unique words.

After preprocessing, the data is split: The train set contains 684 examples (80%) and the test set 171 examples. The split is a stratified group split, meaning that the distribution of the labels in the whole dataset is preserved while ensuring that the same groups do not appear in both train and test set. Furthermore, we randomly oversample the minority class in the training set because the labels' shares are imbalanced.

Results. In section 2.2, we propose various parameters to vary, which results in a vast set of combinations. As it is computationally not efficient to test all possible combinations, we step-wise choose the best configuration to proceed with.

The performance metrics reported in the following section have to be interpreted in context: The training set is tiny (less than 700 samples), and the data is heavily skewed. It can not be expected to achieve a highly predictive classifier; it is also possible that it is beyond the capacity of any machine learning model to distinguish.

Most of the following tables report a high recall, which can be explained by the low predictive quality of the models combined with the imbalanced class distribution of the dataset: When optimizing the F1 score, it is easier for a model with low predictive quality to achieve a high recall than to achieve a high precision on the minority class.

Furthermore, the following tables mostly report an accuracy below 50%. As will be explained in the following section, the prediction threshold is chosen, such that the F1 score is maximized. This procedure results in a decline in the accuracy. Additionally, accuracy is an inconclusive metric when working with an imbalanced dataset.

In the following descriptions of the tables, we consult the F1 score for comparing models *within* a dataset, and the AUC for comparing models *across* datasets.

[2] Bi-grams are built with Gensim's (Rehurek and Sojka 2010) "phrase detector", implying that only bi-grams of common phrases are built.

[3] Stemming, as opposed to lemmatization is typically faster as it merely chops off the end of the word. Lemmatization, however, is more accurate because it derives the root form of a word.

Best Combinations of X and y. First, we test all possible $X$- and $y$-combinations. For this step, we chose a baseline setup where tf-idf vectorized text is input to a support vector machine (SVM). The regularization parameter $C$, the kernel $k$, and the degree of the polynomial kernel function $d$ were subject to a five-fold cross-validated grid search, where we choose $C \in [-1..4]$, $k \in \{linear, polynomial, rbf\}$[4], and $d \in \{2, 3\}$. Furthermore, based on the best parameters proposed by the grid search, the model is refitted ten times to reduce prediction variance. Within each fit, the classification threshold is chosen such that it maximizes the F1 score (as proposed by Lipton, Elkan, and Naryanaswamy 2014). The reported model metrics are averaged over the ten fits.

Each subtable of Table 2.1 shows the best performing combinations for each variation of $X$ and $y$. Table 2.1a shows no variation in performance concerning the spellchecked and the original text. The spellchecker corrected only 17% of the unique words present in the text. Possibly, the correct adjustions were balanced by those where the spellchecker mistakenly changed a word's meaning.

Table 2.1b represents a comparison of different models *across* datatsets. The labels' imbalances vary by the threshold chosen. Therefore, we consult the AUC to select the best performing threshold. The highest AUC (.558) is reached when the dependent variable is binarized based on its mean (317).

Table 2.1c shows that labels as input reach the highest F1 score (.500). This result is not surprising: The 34 labels were assigned by a researcher after reading the chat texts and are therefore far less noisy than chat data. However, assigning labels manually is labor-intensive and prone to mistakes. After all, this paper aims to provide an *automated* approach to text classification. The second best F1 score (.362) is achieved by using chat texts grouped on the subject level ("chat, subject"). Chat texts do not require labor-intensive preprocessing, and as the loss in performance is tiny, we stick with this $X$ variation for the remaining analysis.

Overall, Table 2.1 implies to binarize $y$ based on the mean of the reported income, and deploy chat texts grouped by individuals as input $X$.

Features. We construct features as described in subsection 2.2.2. We either train embeddings on our corpus or deploy pretrained embeddings. We downloaded the latter vectors from the Website "deepset.ai" (Rusic, Pietsch, and Möller 2020) which were trained on the German Wikipedia corpus scraped in 2015. The performance of the different text representation techniques is assessed in two ways. For both methods, a linear logistic regression was used.

Table 2.2 sorts the feature representation based on the F1 score. Overall, the range of the observed F1 scores is narrow (F1 score $\in [.35, .49]$). The performance of the four best variations is almost indistinguishable, showing a standard deviation of only .004. Among those four are the bag of words approaches, and the pretrained Word2Vec embeddings. The unweighted bag of words performs best (F1 score $=$ .490), closely followed by Word2Vec em-

[4] rbf stands for radial basis function.

**Table 2.1:** Best Combinations by Input Variation

**(a)** data

|  | F1 | prec | rec | AUC | acc |
|---|---|---|---|---|---|
| original | 0.362 | 0.227 | 0.963 | 0.558 | 0.286 |
| spell-checked | 0.362 | 0.227 | 0.963 | 0.558 | 0.286 |

**(b)** y

|  |  |  |  |  |  |
|---|---|---|---|---|---|
| <mean | 0.362 | 0.227 | 0.963 | 0.558 | 0.286 |
| <500 | 0.295 | 0.185 | 0.848 | 0.529 | 0.427 |
| <1000 | 0.209 | 0.126 | 0.857 | 0.544 | 0.449 |

**(c)** X

|  |  |  |  |  |  |
|---|---|---|---|---|---|
| label | 0.500 | 0.389 | 1.000 | 0.928 | 0.846 |
| chat, subject | 0.362 | 0.227 | 0.963 | 0.558 | 0.286 |
| chat, group | 0.329 | 0.214 | 1.000 | 0.501 | 0.337 |

F1 score, precision, and recall are reported for the minority label (= 1) "honest".

beddings, pretrained on the German Wikipedia corpus, and tf-idf averaged over the texts (F1 score = .489). fastText embeddings rank second lowest with an F1 score of .354, indicating that a bag of n-grams was not more informative than a bag of words.

Secondly, feature representations should be able to preserve word meanings, at least to such degree that similar words are close in space. For Figure 2.3, the embeddings' dimensionalities were reduced via PCA. Furthermore, it plots only the 30 most common features to keep the visualization readable.

As expected, Figure 2.3b and Figure 2.3c successfully group words of similar contexts: These embeddings leverage pretrained vectors and therefore had the chance to learn a word's meaning from a vast corpus. Figure 2.3b shows that "kontrollieren" (to control), "verdienen" (to earn), "verlieren" (to loose), and "verstoßen" (to comply) are close in space, and thereby precisely capture the setting of the experiment: If the group does not *comply* with the rules and they get *controlled*, they *loose* part of their *earnings*.

Figure 2.3d shows successful grouping, as well. This result is interesting, as fastText embeddings were solely trained on our corpus but nevertheless manage to grasp the word's usage in our context. Figure 2.3a, by contrast, does not succeed in doing so.

Overall, the performance metrics suggest that pretrained Word2Vec embeddings perform second best and seem to learn a word's meaning reasonably well. Therefore, for the

**Table 2.2:** Comparison of the Embeddings' Performance

| | F1 | pr | re | AUC | acc |
|---|---|---|---|---|---|
| bag of words | 0.490 | 0.335 | 0.953 | 0.563 | 0.373 |
| Word2Vec (pre, tf-idf) | 0.489 | 0.336 | 0.982 | 0.509 | 0.372 |
| Word2Vec (pre, smpl) | 0.486 | 0.329 | 0.971 | 0.482 | 0.353 |
| bag of words (tf-idf) | 0.482 | 0.332 | 0.922 | 0.565 | 0.376 |
| Word2Vec (tf-idf) | 0.481 | 0.321 | 1.000 | 0.554 | 0.320 |
| Doc2Vec | 0.361 | 0.232 | 0.909 | 0.500 | 0.317 |
| GloVe (pre, tf-idf) | 0.356 | 0.223 | 0.959 | 0.494 | 0.272 |
| GloVe (pre) | 0.355 | 0.222 | 0.967 | 0.469 | 0.262 |
| fastText | 0.354 | 0.239 | 0.740 | 0.569 | 0.436 |
| Word2Vec | 0.350 | 0.275 | 0.659 | 0.547 | 0.489 |

The classifier used was a logistic regression implemented with the squared euclidean norm as penalty. "pre"-trained vectors were deployed; "tf-idf" weighting was applied. F1 score, precision, and recall are reported for the minority label (= 1) "honest".

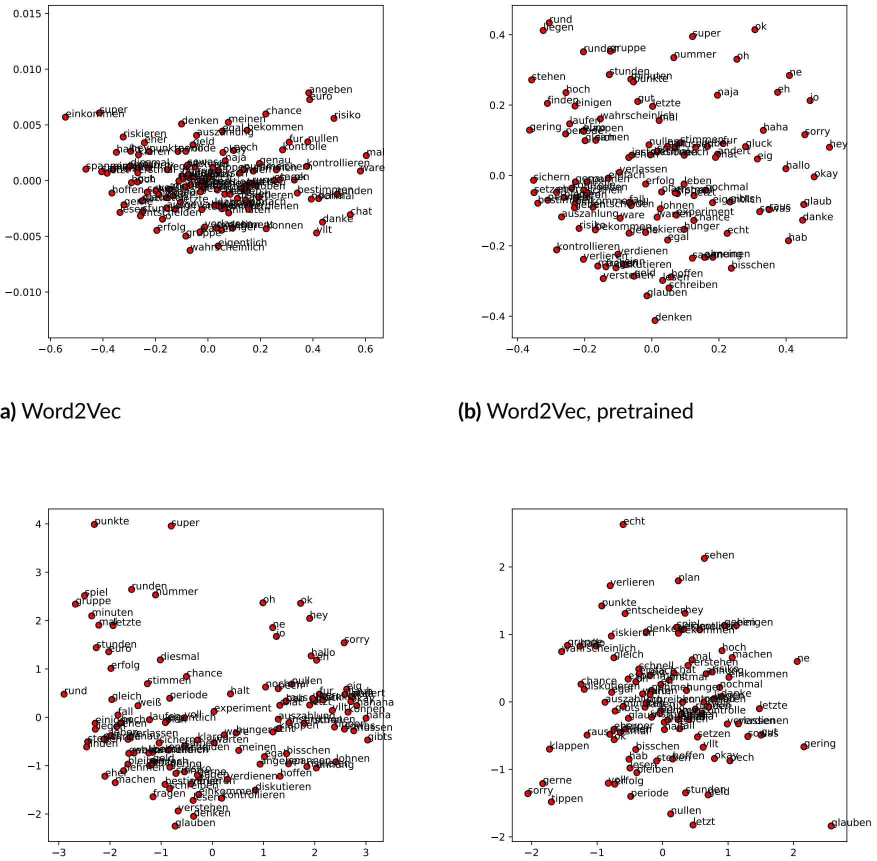
remainder of the analysis, we deploy pretrained Word2Vec embeddings, which we average and tf-idf weight for each subject's chat text.

Classifiers. All classification models' hyperparameters were tuned based on a 5-fold-cross-validated grid search. Which parameters were tuned, depended on the particular classifier chosen. Based on the best parameters proposed, each model is refitted ten times to reduce prediction variance. Within each fit, the classification threshold is chosen such that it maximizes the F1 score (as proposed by Lipton, Elkan, and Naryanaswamy 2014). The reported model metrics are averaged over the ten fits.

Table 2.3a summarizes the performance metrics for all classifiers tested. As it was the case for Table 2.2, in Table 2.3a, the range of the F1 scores is small: $F1 \in [.35, .41]$. The best performing classifier (F1 score = .411) is a stacking classifier, the weighted average of the predictions of different models, as presented by Table 2.3b. Its – in comparison – outstanding performance is not surprising as a combination of estimators reduces their individual biases (Wolpert 1992). Weights as presented by Table 2.3b have to be understood as weight per vote and are chosen such that the F1 score is maximized. In our setting, the random forest classifier earns the highest weight, followed by the 2-layer perceptron and the linear logistic regression.

To summarize, Table 2.3a suggests that there is not much variation in performance across classifiers, and stacking multiple classifiers reaches the highest performance. Therefore, we stick with this ensemble model for the rest of the analysis.

Overall, we show that binarizing the dependent variable at the mean, combined with



**(a)** Word2Vec

**(b)** Word2Vec, pretrained

**(c)** GloVe, pretrained

**(d)** fastText

**Figure 2.3:** Word Embeddings, Reduced in Dimensionality by PCA

grouping the chat texts by subjects, reaches the highest performance. Pretrained and tf-idf averaged Word2Vec embeddings are the most reliable representation of the text. Finally, stacking multiple classifiers outperforms the single ones. This combination yields an F1 score of .411. Nevertheless, this score is rather weak, but this number has to be interpreted in context: A setup with $<$ 700 training samples and the labels' highly imbalanced share could be beyond any machine learning model's capacity. Nevertheless, we show that carefully selecting the configuration can achieve considerable gains in performance. The lowest F1 score reached was .209, which is approximately half of the best model's F1 score.

Apart from the F1 score, we consult the AUC, too. Amongst all models tested, the stacking classifier shows the highest AUC (.597). By contrast, XGBoost yields the lowest AUC of .474 and therefore performs worse than a random guess. We do not expect that XG-

Boost learns to make "negative" predictions; instead, we consider the difference of roughly .03 as noise. Following that logic, we would interpret an AUC of .53 as a random guess, too. However, the stacking classifier yields an AUC of .597, which we, therefore, do not consider a random guess. The predictive power is still weak, but the model did learn to predict (dis-)honesty to some extend.

To summarize, even though the trained classifier's predictive power is weak, it is better than a random guess, which can be considered a success, given the unfavorable circumstances of tiny and heavily skewed data. Furthermore, this paper successfully shows that testing multiple configurations is necessary to find a suitable classification setup for behavioral experimental data.

**Table 2.3:** Results for All Classifiers Tested

**(a)** Performance Metrics for All Classifiers tested

| | F1 | prec | rec | AUC | acc |
|---|---|---|---|---|---|
| Stacking | 0.411 | 0.292 | 0.741 | 0.597 | 0.556 |
| KNN | 0.390 | 0.268 | 0.778 | 0.581 | 0.492 |
| SVM | 0.364 | 0.227 | 1.000 | 0.526 | 0.266 |
| RF | 0.364 | 0.236 | 0.894 | 0.550 | 0.343 |
| NN | 0.356 | 0.225 | 0.926 | 0.559 | 0.298 |
| Bagging | 0.356 | 0.220 | 1.000 | 0.543 | 0.238 |
| LLR | 0.355 | 0.229 | 0.893 | 0.522 | 0.319 |
| XGBoost | 0.349 | 0.215 | 1.000 | 0.474 | 0.214 |

**(b)** Model Weights for Stacking

| | Model Weights |
|---|---|
| KNN | 0.931 |
| NN | 0.805 |
| LLR | 0.618 |
| RF | 0.060 |
| SVM | 0.000 |
| XGB | 0.000 |

F1 score, precision, and recall are reported for the minority label (= 1) "honest".

### 2.3.2 Testing Generalizability

Ideally, the classifier trained generalizes to another setting. Therefore, we collected chat data via a new behavioral experiment. The experimental setup was preregistered[5], programmed with oTree (Chen, Schonger, and Wickens 2016), and conducted online via the server of the MPI Decision Lab[6] in May 2020. Furthermore, the project was supported by the Center for Social and Economic Behavior, University of Cologne, with 3000€.

The instructions state that the participant works as an employee for a fictitious company. Together with one additional team member, she completed a project for which both team members worked the same surplus hours. Both team members could coordinate in a chat about the number of surplus hours they wished to state: The higher the stated amount,

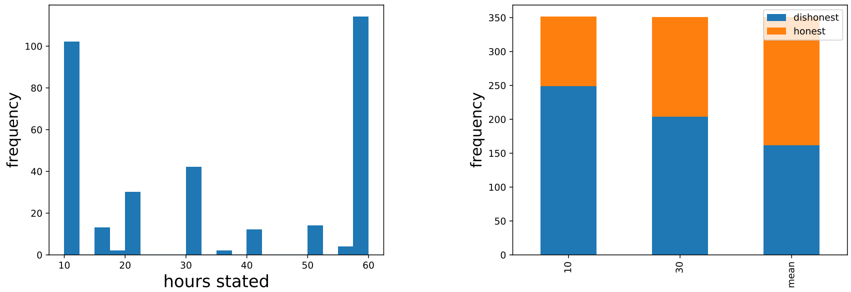
[5] `https://doi.org/10.1257/rct.5049-1.2000000000000002`

[6] `https://www.coll.mpg.de/124252/decision-lab`

the higher was the salary that the fictitious company would pay. If the team members' reports diverged, the couple was controlled. Couples that reported the same amount of hours were randomly controlled with a probability of 30%. If the participant got controlled and reported more than ten surplus hours, she payed a fine. Section B.3 provides screenshots of the experiment.

On a more general note, the new experimental setup alters the setting by Fochmann et al. (2019) concerning three significant dimensions: Firstly, the context is no longer a tax evasion setting, but participants report surplus hours. Secondly, the direction of the lie is switched: It is optimal to overreport in the surplus hour setting, whereas it was optimal to underreport in the tax evasion setting. Thirdly, the group size is reduced from three to two.

Data. Overall, 351 observations were collected. Participants were, on average, 24.8 years old, and 60% of them were female. Figure 2.4a shows the distribution of surplus hours stated. It is binarized based on the full honesty benchmark (10), the mean report (34.1), and half of the full dishonesty benchmark (30). The distribution of the stated income is bimodal, where the greatest masses lie at the full honesty benchmark (10) and the full dishonesty benchmark (60). Comparing Figure 2.4a to Figure 2.1a, it is evident that the participants in the new experiment report considerably more honestly than those in Fochmann et al. (2019). For the classification, we binarized the income stated based on the mean of the reported surplus hours.



**(a)** Distribution of Surplus Hours Stated

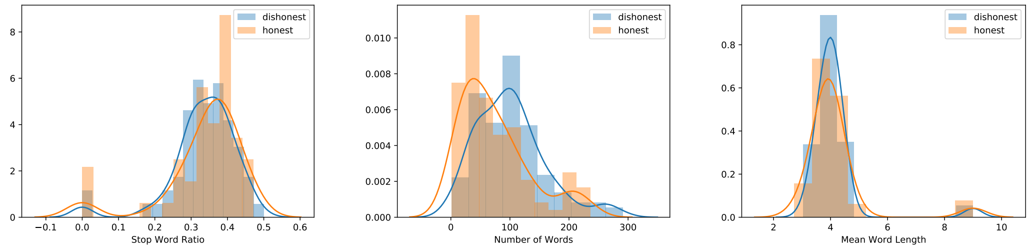
**(b)** Categories Based on Specific Thresholds

**Figure 2.4:** Surplus Hours Stated

Is communication structurally different in the two experiments? Simple word counts suggest that it indeed is: The text data in Fochmann et al. (2019) counts 3, 003 unique words; the text data in the new experiment counts 1, 398 unique words, where only 847 words intersect.

Figure 2.5 provides a more detailed answer to the question posed. According to Figure 2.5a, the distributions of the stop word ratios across the two labels peak roughly at the same values. Figure 2.2a shows similar behavior, and peaks occur in both datasets around 0 and 4. Furthermore, the distribution of the mean word length is bimodal for honest chats and unimodal for dishonest chats. The highest peaks for both distributions occur around 4, which is true for Figure 2.2c, too. Finally, Figure 2.5b displays that chats classified as honest contain fewer words than chats classified as dishonest. This relationship is the other way round for Figure 2.2b.

Overall, the word counts and the distributions displayed in Figure 2.5 suggest that communication is structurally different across the two datasets.



**(a)** Stopword Ratio **(b)** Number of Stopwords **(c)** Mean Word Length

**Figure 2.5:** Distribution of Three Indicators by Label on the New Experimental Chat Data

Comparing AUC. To assess the in section 2.3 trained model's generalizability, we assess its out-of-context predictions on the complete dataset obtained by the new experimental setup. As described in the previous section, the configuration for the model chosen are pretrained Word2Vec embeddings, combined with a stacking classifier. Table 2.4 presents the out-of-context predictions across three thresholds. As the thresholds affect the imbalance of the labels, we consult the AUC instead of the F1 score. The best AUC (.529) is achieved, by binarizing the observed distribution at the mean of the reported surplus hours.

Next, we compare the out-of-context results (on the second experiment) to the results on the test set of the first experiment. As this is a comparison *across* datasets, we again consult the AUC instead of the F1 score. On the test set of the first experiment, the model reaches an AUC of .597. On the out-of-context predictions, the model yields an AUC of .529. Both AUCs are very low, considering that .500 is equivalent to a random guess. The reader is asked to remember that we consider a variation of .03 as noise. Consequently, while the classifier's predictive performance in the first experiment can be interpreted as slightly better than a random guess, the AUC does not indicate any predictive power in the second experiment. Therefore, we conclude that the classifier can not generalize sufficiently well to another context.

This interpretation was already suggested by introspecting the data: The data obtained by the new experimental setup shows that subjects report far more honestly than in the study by Fochmann et al. (2019). Furthermore, distributions show that in the experiment by Fochmann et al. (2019), groups chatted *less*, if they reported honestly. In the new experimental setup, groups chatted *more* if they reported honestly. Furthermore, the number of intersecting words in both datasets is small. These structural differences likely prevented the already weak model from generalizing to a new context.

**Table 2.4:** Out-of-Context Performance of the Pretrained Classifier

| y | F1 | prec | rec | AUC | acc |
|---|---|---|---|---|---|
| > mean | 0.704 | 0.548 | 0.995 | 0.529 | 0.553 |
| > 30 | 0.591 | 0.425 | 0.986 | 0.510 | 0.433 |
| > 10 | 0.455 | 0.305 | 0.922 | 0.506 | 0.365 |

Classification was based on pre-trained Word2Vec embeddings, averaged over texts by tf-idf weighting and a stacking classifier. F1 score, precision, and recall are reported for the minority label (= 1) "honest".

## 2.4 Assessing Robustness

The previous subsection addressed the generalizability of the classifier. This subsection addresses its robustness concerning two major components: Is text an independent predictor in laboratory experiments? Moreover, does the text reflect concepts that previous literature found to predict lying?

Controlling for Experimenter Demand Effects. To rule out omitted variable bias, in the new experimental setup, we checked for experimenter demand effects, which possibly could influence both: chat behavior and the actual report. Therefore, we ask whether participants changed their chat behavior because they anticipate the experimenter would read their communication. On a binary scale, only 4% of participants (overall 14) stated to have changed their chat behavior. Participants could additionally explain their choice. Five people state they thought about the chat being read but did not change their behavior. Four people state that they put more effort into proper grammar and spelling. Furthermore, four people state they wrote less text, and two people state that they reported more honestly.

Out of these four categories, the most critical is the fourth one. However, only two participants stated to having answered more or less honestly, which is negligibly low. The third category, having written less text, is more likely to be motivated by privacy concerns than appearing (dis-)honest in the face of the experimenter. The first two categories are uncorrelated with the dependent variable. Overall, the low number of participants who changed behavior and the categorization of answers show that there is no reason for concern that participants changed their behavior due to an experimenter demand effect. Therefore, we conclude that chat text in laboratory experiments is a suitable independent variable based on which valid predictions can be made.

Various Concepts to Influence Lying. After participants had stated the number of surplus hours, they answered control questions which targeted concepts that were shown in previous experimental research to influence lying behavior. For example, we asked about various feelings and emotions felt during the experiment, risk attitudes, or the field of study. Section B.2 provides more details about the questions and their scales. We visually inspect the distributions of these answers, as displayed by Figure B.1 and Figure B.2. Concepts that show little variation in responses are unlikely to be a strong predictor for lying behavior in our setting. Therefore, we only select concepts that show substantial variation to be included in a linear regression.

Table 2.5 deploys the regression results, which highlight significant correlations of the concepts just mentioned. At this point, it is not our goal to investigate causality. The following four coefficients were found to significantly influence the dependent variable: When the believed number of dishonest people increases, the participant's report increases (belief = .450). Furthermore, the more joy a participant experienced, the more surplus hours she stated (joy = 1.378). A positive and significant relationship holds for risk attitudes, as well: The more risk affinity a participant stated, the higher the reported surplus hours (risk = .905). Likewise, political orientation has a significant influence on lying behavior: The more left-wing a participant claimed to be, the lower was the number of surplus hours stated. In other words, right-wing oriented participants were more dishonest (politics = $-1.081$).

The following three coefficients did not significantly influence the dependent variable: Having participated in more than one economics course, the experience with laboratory experiments and attitudes towards lying did not affect lying behavior. Especially the latter finding, together with the strong influence of experiencing joy, might hint at the limited external validity of laboratory experiments: Lying in a game might be perceived as substantially different than lying in real life. Playing a game and getting the most out of it (e.g., through lying) seems to be a playful and fun experience for participants.

In the following, we investigate the chat messages in terms of those concepts that significantly influence the number of surplus hours stated.

Searching for keywords targeting beliefs and political orientation did not yield meaningful results. Only one couple speculated about other groups' behaviors. No couple spoke about political issues, which was expected, given the experiment's context.

To investigate risk attitudes, we searched for two keywords' patterns. Results show that 31% of all texts include a discussion about either "risk" or "safety". This percentage is rather high when keeping in mind that these are only two keywords. Nevertheless, it is consistent with the analysis by Fochmann et al. (2019), who found that risk-related arguments mainly drive the influence of group interaction on behavior.

**Table 2.5:** Linear Regression Analysis

| | *Dependent variable:* |
|---|---|
| | hours stated |
| joy experienced | $1.378^{***}$ |
| | (0.360) |
| beliefs | $0.450^{***}$ |
| | (0.036) |
| risk attitude | $0.905^{**}$ |
| | (0.405) |
| political orientation | $-1.081^{*}$ |
| | (0.641) |
| econ classes | $-1.836$ |
| | (2.022) |
| lying attitude | $-0.059$ |
| | (0.475) |
| lab experience | 0.308 |
| | (0.762) |
| Constant | $-0.596$ |
| | (6.679) |
| Observations | 318 |
| $R^2$ | 0.426 |
| Adjusted $R^2$ | 0.413 |
| Residual Std. Error | 16.408 (df = 310) |
| F Statistic | $32.849^{***}$ (df = 7; 310) |

*Note:* $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$; hours stated $\in [10..60]$, 10≡full honesty; joy experienced $\in [1..10]$, 1≡experienced no joy; beliefs $\in [0..100]$: 0≡0 people state more than the true amount; risk attitude $\in [1..11]$, 1≡not risk-prone; political orientation $\in [1..9]$, 1≡left-wing; econ classes $\in [1..2]$, 1≡participated in less than one econ class; lying attitude $\in [1..10]$, 1≡one should never lie; lab experience $\in [1..5]$, 1≡never participated in one

The concept of risk could be investigated directly. A direct investigation is, however, not straightforward concerning the concept of joy. Therefore, as a proxy, we use Rauh (2018)'s German Sentiment Dictionary, which includes 17, 330 terms indicating positive sentiment. Figure 2.6 shows that the dictionary based approach approximated joy sufficiently well, as texts with a higher positivity score are associated with more hours stated. The regression analysis can not make any claims about causality, but Figure 2.6 suggests the direction of this correlation: Participants chatted before they stated their surplus hours. This ordering could be interpreted as experiencing positive feelings and emotions such as joy, made participants more likely to lie in the subsequent task.

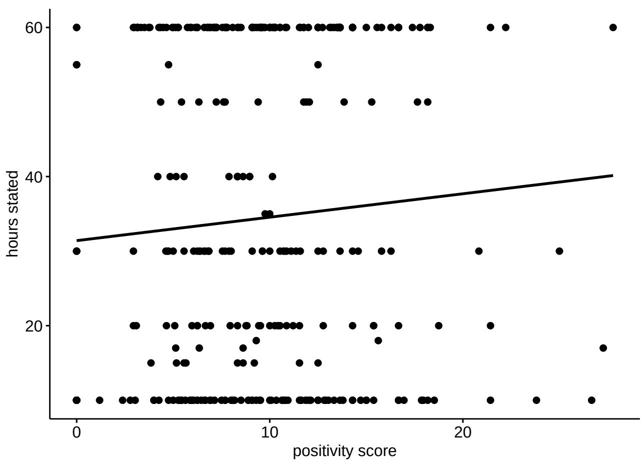


**Figure 2.6:** Relation of the Positivity Score and Surplus Hours Stated

To conclude, this subsection shows that the concepts of joy and risk attitudes are highly predictive and are indeed topics that participants have in mind when developing their decisions.

## 2.5 Discussion and Conclusion

In this paper, we test multiple text classification setups to find the best configuration for behavioral experimental text data. Furthermore, we test whether a classifier trained in the context of one behavioral experiment generalizes to another. For that purpose, we conduct a new behavioral experiment.

Results exhibit that even if an individual participated in a group chat, only the individual's text messages are most predictive of her decision. When mapping the decisions to a broader concept, the mean report as binarization threshold performs best. Concerning the feature representation, pretrained Word2Vec embeddings perform best and can preserve the meanings of the words considerably well. Furthermore, a stacking classifier outperforms all individual models tested. The best configuration yields an F1 score of .411, indicating that the classifier's predictive quality is low; However, this result was expected as the dataset is tiny and heavily skewed. Nevertheless, this paper succeeded in showing that carefully selecting the classification configuration yields a considerable gain in performance.

Finally, we assessed the generalizability of the classifier. It might not be possible to express an unequivocal statement, as the classifier's performance is already weak. Nevertheless, we set the two AUCs – .597 for the first, and .529 for the second dataset – in context and conclude that the model does not generalize to another experimental behavioral setting.

Furthermore, the behavioral experiment reveals interesting concepts connected to participant's lying behavior. Our findings confirm the established results in the literature that beliefs about other's lying behavior and risk attitudes strongly influence one's (dis-)honesty. Additionally, we found that experiencing joy is positively correlated with lying. Analyzing text data sheds light on the direction of this correlation: It seems as if positive sentiments experienced during the group chat encouraged lying in the subsequent decision.

As we could reach a considerable gain in performance just by varying configurations, we expect that training a predictive model would be possible, if a considerably bigger dataset was available. A classifier that was pretrained by these methods could be used in multiple ways.

In the context of experimental behavioral research, it could be leveraged to maximize intervention effects. Assignment to treatment in experimental research is usually random. However, evidence shows that the heterogeneity in reaction to treatment is not random (Engel 2019). Assigning one group to – let us say an honesty treatment while neglecting its type might – in the worst case – crowed out intrinsic motivation and thereby diminish the intended intervention effect (Fehr and Rockenbach 2003; Frey and Oberholzer-Gee 1997; Gneezy and Rustichini 2000). A possible way out is to assign the intervention not randomly but only to groups showing specific characteristics, e.g., an intention to report dishonestly. In this work, we show that group chats are well suited to inform the researcher about the groups' attitudes and planned actions. A predictive model could assess in real time the group's intentions. If the classifier labels a group's intention as dishonest, for example, the group at hand would automatically be assigned to an honesty treatment to alter the behavior towards a more honest response. We successfully tested real-time classification in oTree and can provide the research community with this functionality upon request.

Though we implemented quite an array of configurations, the current paper leaves space for future research. The classifier's predictive performance is low and very likely increases with the size of the training data. It would be interesting to train the best model we found on data of at least double the current training set size.

Furthermore, in many behavioral experimental settings, participants decide repeatedly over multiple rounds. Screening the chat data by Fochmann et al. (2019) reveals that the early chat rounds are the most informative ones. Most groups stick with their initial decision and, consequently, chat about unrelated topics in later rounds. A classification model can exploit such additional information: Decisions in $t-n$ could serve as input features for predictions in $t$. Another way to reflect the time component would be to assign a higher weight to chat texts in $t - n$ as opposed to chat texts in $t$.

*However beautiful the strategy, you should occasionally look at the results.*

Winston Churchill

# 3
# Charting the Type Space: The Case of Linear Public Good Games

Behavior in economic games is not only noisy. One has reason to believe that heterogeneity is patterned. A prominent application is the linear public good. It is widely accepted that choices result from participants holding discernible types. Proposed types, like freeriders or conditional cooperators, are intuitive. However, the composition of the type space is neither theoretically nor empirically settled. In this paper, we leverage machine learning methods to chart the type space. We use simulation to understand what can be achieved with machine learning. We rely on these insights to find clusters in a large (N = 16,474) set of experimental data. We discuss ways in which these clusters could be rationalized.

## 3.1 Introduction

Standard theory predicts the tragedy of the commons. Everybody maximizes individual profit and exploits socially minded choices of others. If members of the community interact repeatedly, but it is known when an interaction will stop, the gloomy prediction still holds. A robust experimental literature shows that, in the aggregate, results look different. In a standard symmetric linear public good, average contributions typically start considerably above zero but tend to decline over time (Chaudhuri 2011; Ledyard 1995; Zelmer 2003). A substantial theoretical literature rationalizes these results, usually by introducing some form of social preference into the utility function (for an excellent overview see Fehr and Schmidt 2002). While such extensions of motives can generate a starting point above zero, it is more difficult for them also to explain the downward trend. For this, one needs a reactive element. It has been prominently introduced into the literature with the concept of conditional cooperation (Fischbacher, Gachter, and Fehr 2001). A conditional cooperator is willing to act unselfishly provided she expects or knows that others will do so as well. In principle, the downward trend could result from the fact that conditional cooperation is imperfect. While participants would not be outright selfish, they would still try to outperform their peers, albeit only slightly (Fischbacher and Gachter 2010). Yet one of us has shown that the data do not support this explanation. Instead, the downward trend results from bad experiences. If participants, in the previous period, have been overly optimistic about the contributions of their peers, they adjust their beliefs and, in turn, their contributions, in the subsequent period. Critically they overreact to negative experiences (Engel and Rockenbach 2020).

This is where the present project starts. If the population were homogeneous, and entirely consisted of conditional cooperators, there could not be a downward trend. The source of the trend, and hence the need for at least some form of institutional intervention to sustain cooperation, must be heterogeneity. Even if many individuals are in principle good-natured and happy to cooperate in good times, their willingness to do so is fragile. If they experience exploitation, they react. While the claim is intuitive that populations are heterogeneous, understanding the character of this heterogeneity is inherently difficult. One needs estimates about the utility functions of group members: is an individual outright selfish? Is she so strongly motivated by the common good that she does not care about others' choices? Or does she react? If so, what does she react to? And how strongly? There could also be mixed types: individuals freeride or cooperate for that matter, unconditionally as long as a certain threshold is not crossed. Reaction functions might have an exploratory component: while an individual is in principle of a certain type, she occasionally tests the waters by contributing more or less than suggested by her ordinary reaction function. Reaction functions could be non-linear. Conditional cooperators might, for instance, be happy to tolerate an occasional bad experience (maybe attributing it to others having made a mistake), but they might lose faith, and react very strongly if bad experiences repeat. There might be individuals who try to educate their groups by showing them what could happen if others do not stop misbehaving. For that purpose, they might once contribute nothing, and go back to high contributions in the following period. Reaction functions may also depend on the effects of occasional exploitation. In the standard setting (group size 4, marginal per capita rate .4) 3 loyal members still make a small profit if they continue to cooperate (and accept that the free rider gains a windfall profit).

All these behavioral programs resonate with data from public good experiments. However, these are only ex post rationalizations. Moreover, not every dataset could be reasonably explained with all of these behavioral programs. Before the field can move forward and better targeted interventions can be designed, one needs a much deeper understanding of behavioral heterogeneity. Ultimately it would be highly desirable to formally define, and experimentally test, these reaction functions. However, a necessary first step is exploratory: which reaction functions exist, and how prevalent are they? Charting the type space is the aim of the present project. We start by assuming that the theoretical possibilities for the composition of the type space are partly understood. We further note that reactions may differ in kind, but also in degree, which is why parameters must be estimated. This is why we revert to machine learning. We use a reasonably large dataset of earlier linear public goods to find types and discuss reaction functions that would rationalize the reaction patterns.

In principle, choice data is well suited for our endeavor. The choices of others in previous periods are the only information to which participants can react in an anonymous linear public good. For each individual, we can check whether, and if so, how they have reacted to past choices of the remaining members of their group. We can represent the development of their choices over time as a timeseries. We can use the rich set of methods developed in the machine learning community for clustering the timeseries of choices, giving the algorithm the possibility to use the average choices of the remaining group members in the previous period as an input. From these clusters we can extract what machine learners call a prototype.

This approach, however, presupposes that reaction functions can indeed be inferred from choices. Arguably this will depend on at least two features of the data: the precision with which an individual participant has reacted to experiences, and the character of these experiences. The former depends on the noise rate. Potentially individuals have a particular reaction function, but they do not act upon it at all times. The latter depends on group composition and initial choices. To illustrate: in a group of three straightforward free riders, a conditional cooperator can be expected to quickly make choices that are indistinguishable from the choices of native free riders. Discriminating between the choices of conditional cooperators and of free riders will be the more difficult, the lower the initial contribution of a conditional cooperator. It should be equally challenging to discriminate between conditional cooperators and genuinely cooperative participants if a group of native cooperators surrounds a single conditional cooperator.

Before using machine learning for clustering participants in real data, we, therefore, investigate with simulated data the framework conditions under which potentially powerful algorithms can find types. In simulations, we can systematically vary the composition of the type space, the definition of individual types, and the noise rate. This first step yields one crucial insight: machine learning methods find patterns. If the choice program of an individual is reactive, the same choice pattern may result from different reaction functions, depending on the choices the remaining group members have made in the previous period. Consequently, there is no one to one mapping between patterns and types. This must be reflected in the design of the clustering algorithm. We show that interpretation becomes much easier if one estimates a number of patterns that is considerably bigger than the expected number of types and hence reaction functions.

Simulation also helps us with two further tasks. We can estimate the richness of the data that is required for making the exercise meaningful. Furthermore, we can check in which ways fine-tuning the algorithm improves estimation.

As explained above, we do not take it for granted that the type space has already been understood completely. A major motive for our project is the possibility that there are further types that have not been theorized. Yet for our simulations, we need to build in types that have already been conceptualized. In the simulations, we work with groups consisting of different fractions of the following five types: altruists, whom we define as participants who do not react to experiences, and who start with relatively high contributions. Such participants may exhibit variance, the more so, the higher the noise rate. However, they show no trend. The corresponding type at the lower end is total free riders. They in principle do not make contributions to the public project, but may occasionally deviate from this program. Pure conditional cooperators start with relatively high contributions but adjust them to experiences. Following Fischbacher, Gachter, and Fehr (2001), we allow for hump shaped contributions: up to a value near half the endowment, they increase contributions in reaction to good experiences, but they exhibit a perverse reaction to even better experiences. Following Engel and Rockenbach (2020), we finally implement farsighted free riders. For some initial periods "they feed the cow" by making substantial contributions but then start "milking" it by reducing their contributions below average contributions in the previous period.

The remainder of this paper is organized as follows: in Section 3.2, we situate our endeavor in the literature. In Section 3.3, we discuss the choice and fine-tuning of the algorithm. In Section 3.4, we explain the data generating process induced by public good experiments. In Section 3.5, we use simulation to demonstrate how clustering algorithms can help the researcher find types, defined by how they react to experiences with the choices of other participants. In Section 3.6, we apply the method to a sizeable set of experimental data. We offer interpretations of the empirical type space. Section 3.7 concludes.

## 3.2 Literature

It has often been noted that choices in public good experiments are not homogeneous (see only Fischbacher and Gachter 2010; Fischbacher, Gachter, and Fehr 2001). But the literature has only relatively recently begun to define the type space more precisely. Amin, Abouelela, and Soliman (2018) use theory derived from Fischbacher, Gachter, and Fehr (2001) to classify 72 participants from a new experiment into seven types, and then use simulation to find out which fraction of which type is required to sustain cooperation in a linear public good. Lucas, Oliveira, and Banuri (2012) show with simulation that cooperation is hard to sustain in a linear public good if the group consists of heterogeneous types (which they take from Fischbacher and Gachter (2010)). Arifovic and Ledyard (2012) develop a model that combines social preferences with learning. In the framework of this model, conditional cooperation is not a type but develops endogenously. They use data from, among others, Isaac and Walker (1988) and Andreoni (1995) to calibrate their model, and argue that it has a good fit. We have a different goal. On the one hand, we do not expect individual choices to be merely noisy. We consider the possibility that heterogeneity is patterned. On the other hand, we do not assume that the behavioral forces that drive this heterogeneity are already fully understood. On the contrary, we want to find patterns that are hard to reconcile with extant theoretical concepts. The purpose of our exercise is hypothesis generation. Testing these hypotheses would require a series of new experiments. That is beyond the scope of the present paper.

Engel (2020) also uses machine learning to organize the type space for experimental data, and demonstrates the approach with data from Fischbacher and Gachter (2010). However, he has a different research question. He wants to compare the performance of a finite mixture model (that estimates the type space and choices conditional on type simultaneously) with a two-step approach (that first estimates the type space from the data, and then choices conditional on type in a mixed effects model that interacts the types estimated in the first step with the effect of experimental manipulations). He also uses a different approach for estimating types, using the coefficients of local (per participant) regressions as inputs for a classification and regression tree.

A third group of contributions is more remote. Game theory usually starts with a complete definition of the game. Yet when they are exposed to one of the games of life, individuals often do not know that much. They must learn what game they are playing. This task is even more laborious if they cannot exclude the heterogeneous population with whom they play. Vorobeychik, Wellman, and Singh (2007) use machine learning methods for the task. Games can be too complex for solving them analytically. Then solutions must be found computationally. Ficici, Parkes, and Pfeffer (2012) make the game tractable by first compressing a large number of agents into a manageable number of clusters, and then solve the simplified game analytically.

Closest in spirit are Bapna et al. (2004) and Lu et al. (2016). Both papers aim at classifying bidding strategies in online auctions (Bapna et al. 2004) and in flower auctions (Lu et al. 2016), using machine learning methods. We have a different game (a dilemma), experimental data, and exploit the power of algorithms for the classification of time series data.

## 3.3 Method

Clustering time series data Repeated experiments produce time series data. It is meaningful to relate the choices of an individual at a given point in time to the choices this individual has made at an earlier point in time, and that she will make at a later point in time. From the development of choices over time, one can infer the program this individual has followed. One can capture the dependence of choices over time with the help of parameters of an appropriate transformation, and then cluster individuals with classic algorithms for static data (Liao 2005); this is how Engel (2020) proceeds, using the coefficients of linear local (per participant) regressions as input for the classifier. This straightforward approach may well be sufficient for many practical applications. Yet the approach requires that the local regressions adequately capture the characteristics of an individual's choice program. As in this project, we want to find the best way to characterize these programs, we prefer a classifier that remains open to unexpected features of the individual timeseries. This is why we exploit the raw time series and use classifiers that have been developed explicitly for time series data (for overviews see Liao 2005; Sardá-Espinosa 2017).

Multivariate clustering Actually, many standard experiments are not only repeated. They are also interactive and produce panel data. In an interactive experiment, the program of an individual participant may react to the experiences she has made with others' choices. This may hold for a cognitive reason: the individual learns from others; or for a motivational reason: the individual wants to react to the choices of others. In principle, the reactive component of the individual choice program could be captured by regressing individual choices on the experiences resulting from the choices made by other group members. Yet this approach assumes that the reaction to experiences is systematic. We are open to this possibility but do not want to impose it on the design of our estimation. This is why, instead, we provide the algorithm with the exact information that participants receive in the experiment. It consists of the average choice of the remaining group members in the previous period. The algorithm thus simultaneously receives two times series: the development of the choices over time that each participant has made; and the corresponding development of the average choices made by the remaining group members in the respective previous period.

Choice of the clustering algorithm Multiple methods have been developed for clustering (raw) multivariate time series (for overviews again see Liao 2005; Sardá-Espinosa 2017). For our data generating process, partitional algorithms outperform hierarchical algorithms. A partitional algorithm assigns each time series to exactly one cluster. The number of clusters is predefined. The procedure starts with randomly chosen centroids, and iteratively improves cluster assignment until the distance between clusters is maximized (Hastie, Tibshirani, and Friedman 2009, chapter 13). We use k-means, as implemented in the `dtwclust` package in `R`.

Our choice of input requires that we use an algorithm that can handle multivariate data. This excludes the otherwise powerful TADPole algorithm. We do, however, not consider this as a significant limitation. The main appeal of TADPole is efficiency, not accuracy (Begum et al. 2015). And it turns out that less efficient alternative algorithms still work reasonably well both with our simulated and experimental data.

In principle, our data generating process would also be amenable to the use of a "fuzzy" algorithm. An attractive candidate would be the fuzzy equivalent of k-means, the c-means algorithm (Bezdek 2013). In the spirit of a finite mixture model (McLachlan and Peel 2004), fuzzy algorithms assign individual time series only probabilistically to any cluster. Yet, even when using simulated data (arguably more clean), the algorithm often does not converge. This is why we do not report estimations with this algorithm. We again think that the resulting limitation is not severe. For comparison with alternative algorithms, we anyhow would have had to use a "crisp" version, that assigns the individual time series to the most likely candidate.

Distance measures All time series clustering relies on a measure for the dissimilarity of two series. Specifically, partitional algorithms need to define the proximity of a data point (time series) to the centroid. It is the distance measures that most profoundly distinguish algorithms for the clustering of time series from algorithms used for cross-sectional data. Two series may have a different length. Seeing the similarity may require that one series is shifted in time. If two series are very close at some point but take on a pronouncedly different shape, one would not want to cluster them as similar. Dynamic time warping (DTW) tackles these challenges. It only requires that the beginning and the end of the time series are firmly matched, and allows all intermediate points to be shifted forward or backward, to construct a better match. The method uses dynamic programming for the purpose (Berndt and Clifford 1994).

In its basic form, this method is computationally very costly. The procedure may occasionally even lead to pathological matches. Both motivate the imposition of constraints. They limit the area that can be reached by the algorithm. We use the constraints proposed by Cuturi (2011) and implemented in the `dtwclust` package as option `sdtw` (for "soft DTW"), and allow for windows of size 2.[1] The algorithm thus seeks for similarity in the present period, but also two periods before and after. Given much of our experimental

[1] As a rule of thumb, a window of about 10% of the length of the time series is sufficiently wide (Ratanamahatana and Keogh 2004, sec.2.1 with refs.).

data has only ten periods, and experiences begin only in the second period, this window-size seems appropriate for our data generating process.

Prototype extraction The purpose of clustering is grouping the data by some measure of proximity. This would not be necessary if the data within each cluster were perfectly homogeneous. But ultimately, one wants to interpret the cluster. This requires defining a prototype, an observed or constructed time series that best characterizes the cluster. If one uses a partitional algorithm, the prototype is also used during the clustering process, as a cluster centroid. We use the centroid based on soft DTW (`sdtw_cent`). The method uses DTW matches, rather than the raw data, to find the most characteristic time series, relying on soft DTW as the starting point.

## 3.4 Data Generating Process

Linear Public Goods While we believe our method to be applicable more generally for finding patterned heterogeneity in repeated, interactive experiments, our specific object of investigation is a linear public good. The following profit function defines the game:

$$\pi_{it} = e - c_{it} + \mu \sum_{k=1}^{K} c_{kt} \tag{3.1}$$

where $\pi$ is profit of individual $i$ in period $t$. Every period, the individual receives an endowment $e$. She can keep the endowment or contribute $c$ to the group's public project. Marginal per capita rate $\mu < 1 < K\mu$ creates the dilemma. As $\mu < 1$, each individual is best off keeping the entire endowment for herself. Yet as $K\mu > 1$, the group is best off if all members contribute the entire endowment. Most frequently, $e = 20, \mu = .4, K = 4$ have been chosen (Chaudhuri 2011; Ledyard 1995; Zelmer 2003). Then three loyal group members still make a small profit. This serves as a buffer against the rapid decline of contributions.

Simulated type space In their seminal paper, Fischbacher, Gachter, and Fehr (2001) argue that (in their one-shot version of this game) there are three types: freeriders, conditional cooperators, and "hump-shaped" players. In his reanalysis of Fischbacher and Gachter (2010), Engel (2020) further finds a small, but discernible fraction of altruists. In their reanalysis of Fischbacher and Gachter (2010), Engel and Rockenbach (2020) use a combination of belief and choice data to distinguish a fifth group, which they call far-sighted free riders. In our simulations, we allow for these five types. We focus on a partner design. Groups stay together for the full duration of the game. We always allow for an individual random effect $\eta_i$ and residual error $\varepsilon_{it} \perp \eta_i$, which we both define to be normally distributed with mean 0 and standard deviation .3 ($\sim \mathcal{N}(0, .3)$). Thus, we implement the type space as defined in Table 3.1, where $c_{-i,t-1}$ is the average contribution of the remaining group members in the previous period.

We have two types that exhibit variance (between participants due to $\eta_i$, and within participants due to $\varepsilon_{it}$), but that do not react to experiences: short-sighted freeriders and altruists. The contributions of these types do also not have a trend. They are random walks, albeit with opposed starting points. By contrast, the remaining three types are reactive, which may depending on the choices of the remaining group members $c_{-i,t-1}$, lead to a trend. We have (true) conditional cooperators start in the middle of the action space. In early periods ($t < \tau$) far-sighted freeriders mimic conditional cooperators, but from period $\tau$ on, they freeride. Such participants "feed the cow" for a while, to then "start milking" it.[2] Finally, we simulate hump-shaped participants such that they start rather low, at 5, and have them behave like conditional cooperators as long as the remaining group members, in the previous period, have on average not contributed more than half of the endowment. If $c_{-i,t-1} > 10$, they exhibit a perverse reaction. The more others have contributed, the less they contribute themselves.

**Table 3.1:** Simulated Type Space

| type | $t = 1$ | $t > 1$ |
|---|---|---|
| short-sighted freerider | 0 | 0 |
| far-sighted freerider | 10 | $c_{-i,t-1}$ if $t < \tau$<br>0 if $t \geq \tau$ |
| conditional cooperator | 10 | $c_{-i,t-1}$ |
| hump shaped | 5 | $c_{-i,t-1}$ if $c_{-i,t-1} \leq 10$<br>$-c_{-i,t-1}$ if $c_{-i,t-1} > 10$ |
| altruist | 20 | 20 |

We have groups of size $K = 4$, and we allow for $n = 5$ types. Participants choose their contributions to the public good simultaneously, which is why their order does not matter. We consider the possibility that types are present more than once in a group. Hence we have a problem of unordered sampling with replacement. This gives us a total type space of

$$N = \binom{n+k-1}{k} = \frac{(5+4-1)!}{(5-1)!4!} = 70 \tag{3.2}$$

different group combinations. In our simulations, we include each of these 70 combinations of types four times.

[2] In our simulations, we set $\tau = 5$.

Confusion matrix Simulation is routinely employed to test the performance of an estimator. One generates a data set where one knows ground truth and checks whether a proposed estimator reasonably reconstructs the simulated parameters. If an alternative estimator outperforms a competing estimator, one adopts the better performing method. Simulation gives the researcher confidence in using an estimator with data where she does not know ground truth.

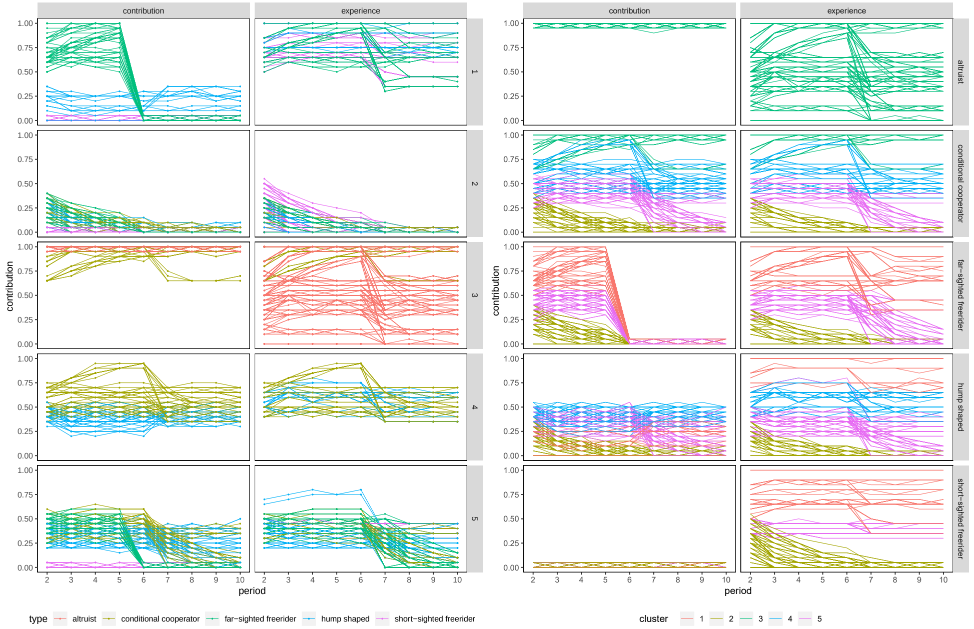
When applied to our estimation problem, the seemingly straightforward criterion for choosing an estimator would be the frequency of identifying the simulated types. Assessed with this criterion, the results reported in Table 3.2 are sobering. Each of the five types is precisely 224 times present in the dataset. Yet the size of the clusters ranges from 135 to 320. All clusters except the third are fairly impure: participants from different simulated types are put into the same cluster. In clusters 2 and 5, there is a prominent type, but it is not in the majority. In cluster 4, the most prominent type (conditional cooperators) is in the majority, but the cluster is tiny.

Yet it is most worrisome that, even knowing ground truth, it is hard to match clusters with types. Numbers are printed in italic if the most frequent type per cluster, and the most frequent cluster per type, coincide. In the example dataset, this holds for all types but the hump-shaped players. But even if one were to use these possible unique matches, only 48.3% of all participants would be matched.

**Table 3.2:** Types vs. Clusters

| cluster | 1 | 2 | 3 | 4 | 5 | Total |
|---|---|---|---|---|---|---|
| short-sighted freerider | 48 | *140* | 0 | 0 | 36 | 224 |
| hump shaped | 24 | 60 | 0 | 58 | 82 | 224 |
| conditional cooperator | 0 | 60 | 28 | *77* | 59 | 224 |
| altruist | 0 | 0 | *224* | 0 | 0 | 224 |
| far-sighted freerider | 64 | 60 | 0 | 0 | *100* | 224 |
| Total | 136 | 320 | 252 | 135 | 277 | 1120 |

Clusters are patterns, not types Figure 3.1a shows why the attempt fails to validate 5 clusters by comparing them with 5 simulated types. The algorithm visibly does a good job at clustering the data. But it clusters patterns of observed contributions, and of observed experiences. There is no one-to-one mapping of 5 patterns to 5 types. Cluster 1 collects choices when experiences are excellent, but choices are more selfish. This holds for far-sighted freeriders, for hump-shaped players, and even for a few short-sighted freeriders if they are matched with particularly cooperative types. Cluster 2 assembles individuals and groups that exhibit the classic downward trend, which can occur for all types but altruists. Cluster 3 stresses unswerving cooperativeness, irrespective of experiences. This, of course, holds for altruists and those conditional cooperators in a particularly cooperative environment. In cluster 4, reactive types manage to cooperate reasonably well, whether they are conditional cooperators or hump shaped players. By contrast, cluster 5 assembles reactive types that do not manage to sustain cooperation, at least not in the long run.



**(a)** 5 Clusters for 5 Types

**(b)** 5 Types in 5 Clusters

**Figure 3.1:** 5 Clusters for 5 Types on Simulated Data

It is even more instructive to consider which types are put into which clusters (Figure 3.1b). Altruists are in a cluster of their own (but some conditional cooperators are also put into this cluster). All other types are distributed across multiple clusters. For conditional cooperators, there is a close mapping between experiences and choices. Depending on the quality of experiences, they are distributed over 4 clusters. For far-sighted freeriders, the gap between their contributions once they start cashing in and experiences are critical. This, in principle, also holds for hump shaped players. Yet if their contributions are relatively high and reflected in their experiences, they are put into one cluster with a fraction of conditional cooperators. Finally, short-sighted freeriders are split into three clusters, depending on the cooperativeness of experiences.

Hence upon closer scrutiny, there is not a problem with the performance of the algorithm. It just does not do what one might have naïvely expected. The object of classification is not types, but time series. Three of the types that we have simulated are reactive themselves. Unless the environment exclusively consists of short-sighted freeriders or altruists (which only holds for 2 of 70 simulated group compositions), individuals with a consistent reaction function respond to various environments. If we impose 5 clusters, the algorithm must distribute pairs of experiences and choices across these clusters as best it can.

If one allows for types to be reactive, one cannot directly infer reaction functions from the data. Precisely because types are allowed to be reactive, the same reaction function may lead to distinctly different choice patterns. Just considering choice patterns would be misleading as well. One would miss the possibility that, in specific environments, multiple types exhibit very similar behavior. In Figure 3.1a, the point is most forcefully illustrated by the biggest cluster, cluster 2. Since overall cooperativeness is low in these groups, the choices of conditional cooperators, hump-shaped players, far-sighted freeriders, and short-sighted freeriders look very similar.

One needs an indirect strategy if one wants to infer potentially reactive types from the data. The proximate object of discovery cannot be types. It must be two-dimensional patterns, i.e., combinations of the development of experiences over time with the development of choices over time. The data can only inform the researcher about the distinct characteristics of these patterns. As the next step in the research process, she must attempt to rationalize these patterns.

How many patterns? From the foregoing, it is clear that one should expect more patterns than types and hence should impose a number of clusters that is larger than 5. But which is the optimal number? As we know the data generating process, with our simulated data, we can derive the maximum from theory. In the dataset, we have 5 types who interact in groups of 4. From (3.2) we know that this leads to 70 different group compositions. One might think that the number of environments that a player may face is smaller, as there are only three others in the group. Yet others are potentially themselves reactive. Then the choices the individual in question has made in the past may have influenced the experiences she has herself made in subsequent periods. Theoretically there are consequently 5 types $\cdot 70$ environments $= 350$ different patterns.

In the simulated data set, we have 1120 pairs of time-series for experiences and individual choices. If it were necessary to estimate 350 clusters, there would be little more than 3 participants per cluster, on average. The simulated dataset would be too small for the purpose. More disturbingly, it would be challenging to compile a set of experimental data that is big enough for this study's ultimate goal: to find out whether there are untheorized types. In the following, we show that a more parsimonious approach is both feasible and adequate. It has the additional important advantage that it can be applied to datasets where ground truth is unknown.

In machine learning language, one then faces a problem of unsupervised learning. "Internal" cluster validation indices have been developed for this task. They strike a balance between underfitting and overfitting the data. One does not want to miss information in the sample that is informative about the population one wants to understand. On the other hand one is concerned that one might assign meaning to noise. The following validation indices have been developed for time series data:

- Silhouette index (`Sil`)[3]
- Dunn index (`D`)
- COP index (`COP`)
- Davies-Bouldin index (`DB`)
- modified Davies-Bouldin index (`DB*`)
- Calinski-Harabasz index (`CH`)
- Score Function (`SF`)

These CVIs differ by the emphasis they put on cluster cohesion over cluster separation; whether they combine both parameters by way of summation or division; whether or not they rely on normalization (for detail see Arbelaitz et al. 2013). As we have no strong conceptual reasons to prefer one CVI over the other, we employ all methods and aggregate over the outcomes to find the best number of clusters.[4]

Results are visualized in Figure C.1 in the Appendix.[5] Visibly the indices do not all come to the same conclusion. The CH index has an early maximum. The DB index declines with few clusters but stabilizes from 14 clusters on. The SF index steadily grows until it reaches 16 clusters and then levels off. DBstar is relatively stable, but peaks at 32 clusters. The Sil index exhibits a few occasional drops but otherwise is fairly stable. Except for a slight peak at 10 clusters, the COP indexes also very stable. The D index has local ups and downs, but a positive trend.

To find the optimal number of clusters in the face of this heterogeneity, we proceed as follows: independently for each index, we rank the preferred number of clusters. Subsequently, separately for each number of clusters, we aggregate the ranks. It turns out that a specification with 39 clusters ranks highest. It is followed by 36 and 40 clusters, which gives us confidence that the optimal number of clusters is in this range.

[3] Letters refer to the code in `R` package `dtwclust`.

[4] As the clustering algorithm has a random starting point, we repeat the comparison with ten different starting points and use the mean index per CVI.

[5] The Dunn index, the COP index, and the Davies-Bouldin index are to be minimized, while the remaining indices require maximization. For comparability, the indices to be minimized are recoded.

How do types translate into patterns? As we have explained, we cannot use the mapping between clusters and types to validate our estimation procedure. "Ground truth" consists of pairs of experience and choice patterns. We are not aware of the formalizations of such pairs. Yet we can use simulated types to understand how the procedure works. This will help with interpretation if, with experimental data, we do not know reaction functions. This step of the analysis serves two functions. At a more general level, it helps understand pairs of experiences and choices as the object of investigation. We can use the patterns that we observe in the simulation as a blueprint at a more specific level. If in the experimental data, we find patterns that are dissimilar to all the 39 patterns shown in Figure 3.2, we know that extant theory does not exhaust the type space. If such patterns are sufficiently frequent in the experimental data, we know the gaps in the present understanding of the type space.

Figure 3.2 shows how types translate into two-dimensional patterns. The translation is easiest to understand for types we know not to be reactive themselves, i.e., for altruists and short-sighted freeriders. Altruists are dispersed over the first 7 clusters. Unsurprisingly, their contributions (left panel) are always near the top; for them, only the noise terms create variation. Yet the algorithm splits them up into clusters depending on the experiences they make, i.e., depending on the characteristics of the environment. As we know the data generating process, we can reconstruct how these environments originate.

In clusters 1 and 2, altruists are together with conditional cooperators. Experiences start at a lower level but gradually move to the top. This pattern would not emerge in the presence of a far-sighted freerider: she would start cashing in and draw down experiences from then on. The pattern could also not emerge in the presence of hump-shaped players: they would react perversely right from the start. Finally, the pattern would be impossible in a group with short-sighted freeriders: experiences could not reach the top. In clusters 3, 4, and 5, the level of experiences is less good, but on average, experiences are relatively stable. This pattern can result if an altruist is alone in a group with others who are willing to sustain a medium level of cooperation. They could be conditional cooperators or hump shaped players. In all three clusters, a few times series have a kink in the middle. We know that this kink is triggered by the presence of at least one far-sighted free rider. Yet as long as the kink is not pronounced, the algorithm does not use it to characterize the cluster. This is different with clusters 6 and 7. In these clusters, the kink dominates the shape of the environment. These are groups with more than one far-sighted freerider.

On the lower end of the spectrum, short-sighted freeriders are spread over clusters 29–39. Clusters 34–39 are exclusively populated by participants whom we know to be short-sighted freeriders. We find the mirror image to altruists. Clusters do not differ by participants' own choices: they are always near the bottom. The algorithm distributes participants over different clusters since they live in characteristically different environments. In clusters 36–39, the average contributions of the remaining group members start somewhere in the middle and gradually deteriorate. This pattern results if sufficiently many of the group members are reactive, be they conditional cooperators, hump-shaped, or far-sighted freeriders. In clusters 34 and 35, contributions of others do not fall to the bottom and are reasonably stable. This pattern requires that others sustain cooperation in the face of exploitation. By the design of the game, this is only possible if there is no more than one short-sighted freerider in the group. In clusters 29–33, the algorithm mixes short-sighted freeriders with different types. These different types affect the level and shape of participants' contributions: they are slightly more positive than in clusters 33–39. Among themselves, clusters 29–33 differ from the experiences participants make. These experiences are poor and declining in clusters 30–32, but very favorable in cluster 29.

There are only 5 clusters that exclusively consist of conditional cooperators, clusters 8, 9, 11, 15, and 22. In clusters 1, 2, and 7, they are put together with altruists. In clusters 12–14 and 16, they are classified together with hump-shaped players. In cluster 21, they are together with far-sighted freeriders. In cluster 33, they are together with short-sighted freeriders. Clusters 10, 17, and 23–25 mix them with even more types. This forceful shows: an entirely reactive type is most challenging to infer from the choices she makes, combined with the choices to which she reacts. Yet another characteristic is worth noting. If there are conditional cooperators in a cluster, contributions and experiences resemble each other very closely. Like chameleons, participants adapt to the local environment. If this environment is cooperation friendly (clusters 1–2), so are the conditional cooperators. If the level of cooperation is not perfect, but reasonably high (cluster 8, 14, 15, and 16), this is what they match. If far-sighted freeriders start cashing in (clusters 11, 12, 21, and 22), they follow suit. If cooperation quickly fades out (clusters 13, 17, 23–25), this is how they behave as well. The near perfect match between choices and experiences is how this type can be inferred from the data.

Whenever there is a kink in experiences, there is at least one far-sighted freerider in the group. But the inverse does not hold. The fact that there is a kink in contributions does not imply that the participant in question is a far-sighted freerider. She may instead be a conditional cooperator (clusters 7, 8, 10, 11, 12, 21, 22) or a hump-shaped player (clusters 10, 12, 21), who reacts to the choices of one or more far-sighted freeriders. As these two types are reactive as well, experiences may also exhibit a kink when the player in question is indeed a far-sighted freerider (clusters 18 and 19). There is, however, a way to identify far-sighted freeriders from the data. This is straightforward if experiences are stable, and only the participant in question reduces contributions (cluster 20). More frequently, the cue is more subtle: the participant's contributions decay earlier than experiences (clusters 10, 18, 19).

Hump-shaped players are most easily identified if they react inversely to experiences. This happens if the mean contributions of others fall below the threshold (which we have simulated to be at 10). Then they swap strategies and start (re-)stabilizing cooperation (clusters 26 and 27). Yet if the cooperation level is poor in the first place, hump-shaped players just behave like conditional cooperators, and are hard to distinguish from them (clusters 23–25), or from short-sighted freeriders if cooperativeness is very high (cluster 29) or very low (clusters 30–32).

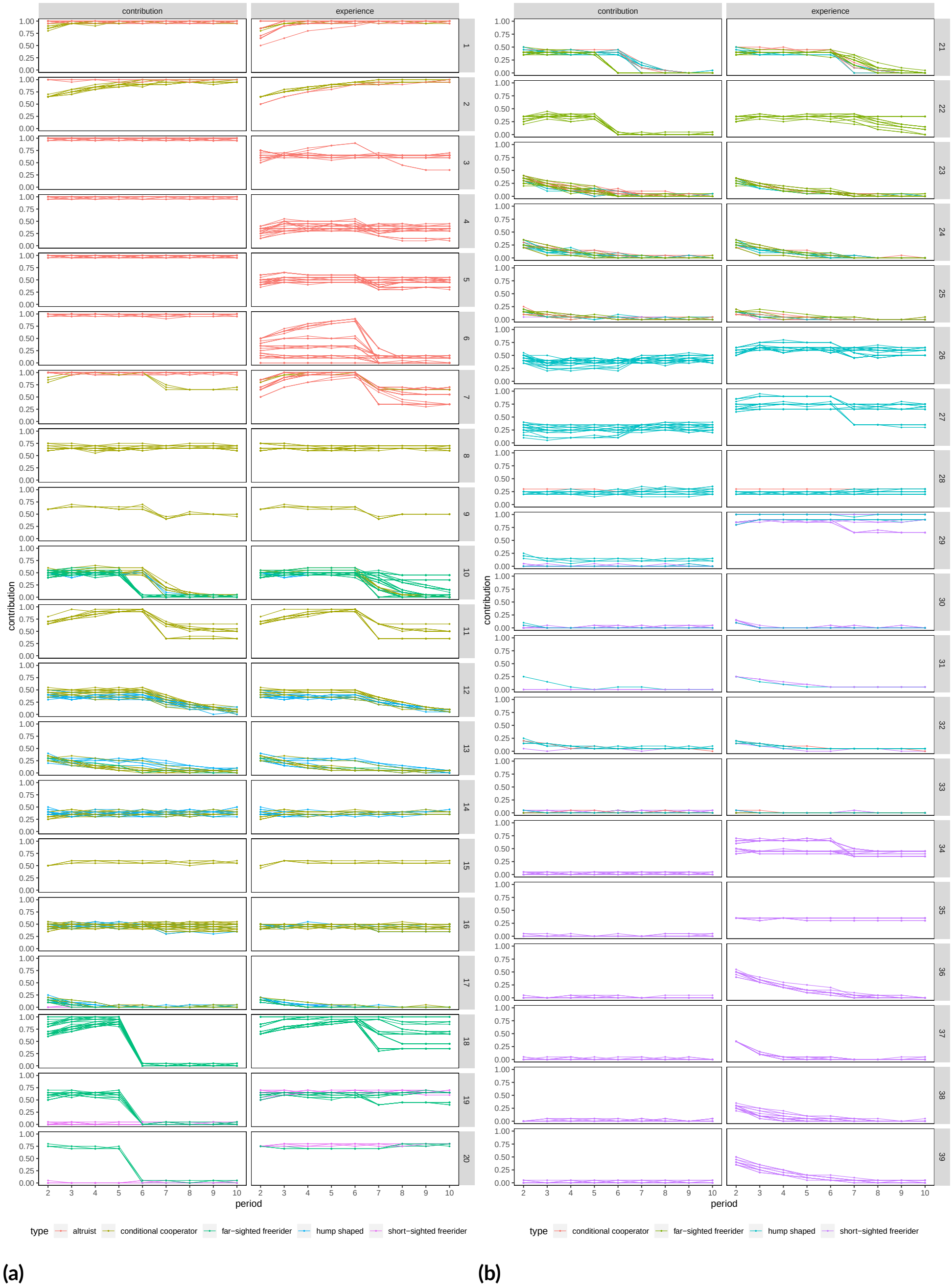
## 3.6 Experimental Data

Section 3.5 has demonstrated in which ways, in a linear public good, a pair of two timeseries is related to the reaction function of a participant. The development of choices over time must be seen in the light of the development of experiences this participant has made. As we have explained, there is no one-to-one mapping between this two-dimensional times series and the reaction function, and hence the participant's type. Yet we have shown in which indirect ways the type can be inferred. As we expect the type space to be limited, we use clustering (of two-dimensional time series data) to organize the evidence. This gives us a methodology for the ultimate purpose of writing this paper: we want to infer from clustering real, experimental data whether the true type space differs from, or is richer than, the five types that have already been established and theorized.

**Table 3.3:** Information on Experimental Studies Included

| study | period | endowment | group size | MPCR | subjects |
|---|---|---|---|---|---|
| Diederich, Goeschl, and Waichman (2016) | 7 | 40 | 10 | 0.3 | 360 |
| Diederich, Goeschl, and Waichman (2016) | 7 | 40 | 40 | 0.3 | 200 |
| Diederich, Goeschl, and Waichman (2016) | 7 | 40 | 100 | 0.3 | 500 |
| Diederich, Goeschl, and Waichman (2016) | 7 | 1, 000 | 10 | 0.3 | 50 |
| Engel and Kurschilgen (2013) | 30 | 20 | 4 | 0.4 | 44 |
| Engel and Kurschilgen (2014) | 30 | 20 | 4 | 0.4 | 48 |
| Engel and Kurschilgen (2020) | 30 | 20 | 4 | 0.4 | 48 |
| Engel and Rockenbach (2020) | 20 | 20 | 3 | 0.4 | 30 |
| Engel, Kube, and Kurschilgen (2020) | 10 | 20 | 4 | 0.4 | 96 |
| Kosfeld, Okada, and Riedl (2009) | 20 | 20 | 4 | 0.4 | 40 |
| Kosfeld, Okada, and Riedl (2009) | 20 | 20 | 4 | 0.6 | 176 |
| Nikiforakis and Normann (2008) | 10 | 20 | 4 | 0.5 | 24 |

Data Table 3.3 defines the dataset. We only use data from linear public goods without any experimental intervention, i.e. data from voluntary contribution mechanisms. We have a total of 18,090 observations from 1,616 participants. As participants have not yet made any experiences in the first round, we only use data from the second round on, which gives us 16,474 usable datapoints. Figure 3.3 visually represents the dataset. On average, all experiments featured in the dataset exhibit the characteristic negative time trend. Yet there is considerable variance. The level of cooperativeness is differently high. The decay in cooperation is differently steep. In one experiment, contributions are even almost stable over time. We see this variance as an advantage. It gives us more scope for finding unknown reaction functions, in particular, due to variance in the experiences participants have made.



(a)

(b)

**Figure 3.2:** 39 Clusters for 5 Types on Simulated Data

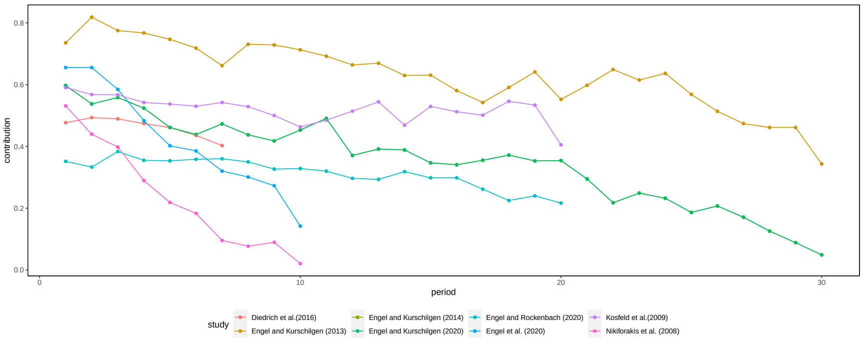


**Figure 3.3:** Means of Participants' Contributions by Study

Clustering Section 3.5 shows that, for finding reactive types, one needs a sufficiently large number of clusters. With simulated data, 39 clusters have proven the best amount. Our strategy for finding hitherto unknown types relies on a comparison between simulated data patterns and the patterns generated by the otherwise same methodology with experimental data. This is why we use the same algorithm, with the same parameters, and have it organize the data into 40 clusters.

**Table 3.4:** Size of Experimental Clusters

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| cluster | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| size | 44 | 54 | 43 | 69 | 20 | 28 | 20 | 15 | 6 | 15 |
| cluster | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| size | 20 | 21 | 18 | 17 | 4 | 14 | 16 | 50 | 52 | 32 |
| cluster | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 |
| size | 75 | 104 | 35 | 89 | 30 | 45 | 18 | 56 | 25 | 18 |
| cluster | 31 | 32 | 33 | 34 | 35 | 36 | 37 | 38 | 39 | 40 |
| size | 99 | 31 | 55 | 87 | 26 | 91 | 61 | 21 | 58 | 34 |

As a comparison of Figure 3.2 with Figure 3.5 shows, in the experimental data there is considerably more variance between individuals than in the simulated data. In the interest of facilitating interpretation, we, therefore, exploit that the clustering algorithm not only assigns participants to clusters. It also, separately for each cluster, defines a prototype, using the centroid. These prototypes (of choices combined with experiences) are collected in Figure 3.4. We focus on this graph.

A few of the experimental clusters resemble clusters in the simulated data. Cluster 1 (experimental) is similar to clusters 1 and 2 (simulated). Cluster 2 (experimental) is similar to Cluster 3, 4, and 5 (simulated). Cluster 6 (experimental) is related to Cluster 11 (simulated). Clusters 37 and 40 (experimental) are comparable to clusters 34 and 35 (simulated). Yet, the prototypes of many experimental clusters are completely different from the simulated clusters. Moreover, even if a cluster is comparable in some respect, it is dissimilar in an important other respect.

Interpreting the five biggest clusters We start discussing the true type space with the five most frequent clusters. In the biggest cluster 22 (104 groups), groups successfully coordinate in the middle of the action space. This also holds for the player in question. Yet her choices exhibit considerably more variance (Figure 3.5). In the most characteristic time series of individual choices (the centroid), contributions increase slightly in the second half of the game but go back to the midpoint in the end. Different deviations are present in the raw data (Figure 3.5), but deviations from the middle tend to be small. In principle, this pattern could result from conditional cooperation. However, the centroid suggests that deviations to higher contributions are not just noise, but systematic. Such deviations cannot be explained by pure conditional cooperation. The choice program must go beyond merely matching (reasonably favorable) experiences. A potential motive is signaling. Participants who are willing to cooperate themselves signal others that the whole group could do even better, by shifting to a higher level of contributions. This would be participants who are unwilling to be exploited and therefore go back to the way how others react to the initiative. But in their perspective, this would only be the second best outcome. They give in to the implicit message they receive from the rest of the group: we are not willing to invest even more in the joint project.

The second biggest cluster 31 (99 groups) looks like the classic pattern from VCM experiments: choices and experiences decay over several initial periods, and almost reach the bottom. Clusters 13, 23, 24 of the simulated data exhibit a comparable pattern. Yet one difference is worth noting. Individual choices stay in the area of experiences but are systematically lower. There are two possible interpretations. Fischbacher and Gachter (2010) have suggested the possibility of conditional cooperation being imperfect. In their reanalysis of Fischbacher & Gächter's data, Engel and Rockenbach (2020) argue that the seeming imperfection results from the choices made by far-sighted freeriders. In the simulations, this type has been modeled as participants perfectly mimicking conditional cooperation until a specific period. For rationalizing cluster 31, one would need an alternative definition of the type. Selfish participants would never perfectly condition their contributions to experiences. They would just stay close enough so that truly cooperative group members can assign a lower overall level of contributions to noise.

Participants assigned to the third biggest cluster 36 (91 groups) exploit more cooperative group members: their contributions are systematically below experiences. Nevertheless, their choices are not at or near the bottom. In the beginning, they even increase contributions. At some later point (at period 7 in the centroid), they increase contributions even further to decrease them. This pattern can be rationalized by the intention to exploit the group while being wary that others could lose faith if the overall level of contributions looks too poor. This would be super smart free riders. They invest in sustaining the cooperativeness of the remaining group members that they intend to exploit continuously.

In the fourth biggest cluster 24 (89 groups), experiences are almost entirely stable, at an intermediate level. However, the participant reacts to this stability with pronounced instability. In the centroid, the participant reduces contributions for two periods and then goes back to the group's cooperation level. None of the five simulated types could rationalize this behavior. Two plausible choice programs could generate this behavior. One option is a far-sighted freerider who tries to exploit the group but is afraid that further exploitation would not be tolerated. Alternatively, participants who in principle are willing to cooperate want to test the waters, or fall into temptation for that matter, but lose courage.

The fifth biggest cluster 34 (87 groups) is also best explained by modified free-riding. For most of the experiments, experiences are almost entirely stable, at an intermediate level. However, choices deteriorate in the first periods, remain stable at a level below experiences, and further decay by the end of the game. It is only at this point that experiences slightly decay as well. This could be the behavior of a person intending to exploit the group. She might hold the opinion that contributions slightly below experiences will not induce cooperative participants to reduce contributions. In this perspective, the initial decay results from learning. A selfish participant believes that she can somewhat reduce her contributions without putting the cooperativeness of the truly cooperative participants at risk. By the end of the game, she sees no point in further sustaining cooperativeness. A slight decay in the contributions of others would no longer outweigh the additional cost of their contributions.

Families of clusters In Figure 3.4 and Figure 3.5 we have ordered the 40 clusters such that family resemblance becomes visible. In clusters 1–5, the participant is very cooperative. In cluster 1, individual and social cooperativeness match nearly perfectly. Now the participant in question increases her contributions over the first periods. This excludes that she is genuinely cooperative; she is not an altruist in the sense of the simulation. The pattern can be rationalized by perfect conditional cooperation. By contrast, choices in clusters 2–5 are substantially higher than experiences. However, choices are only very close to the top throughout in cluster 2. In clusters 3–5, choices increase for the initial periods. They decrease in the end. In the centroid of cluster 5 there is also a dip at an intermediate period. In cluster 4, choices never reach the top. None of clusters 2–5 could result from conditional cooperation. Yet choices can also not be explained by altruism. Otherwise, choices would have to be at the top continuously. These choice patterns suggest a missing hero attitude. Participants want to save the entire group by inducing all others to make higher contributions. The lower contributions in the first rounds would result from doubt whether cooperation is possible in this group. The participant in question would conclude that the overall level of cooperation is high enough to invest in sustaining a high level of cooperation worth making. The decay of contributions in the final periods, observed in clusters 3–5, supports this explanation. The later in the game, the lower the expected payoff from any further investment.

Choices in the two clusters of the second family (clusters 6 and 7) look very similar. Choices quickly move to the top (cluster 6) or to a very high level (cluster 7), and stay high until close to the end. Yet experiences are very different. In cluster 6, they are very high, while they are rather low in cluster 7. The pattern of choices and experiences in cluster 6 could result from either conditional cooperation or a very patient version of a far-sighted freerider. By contrast, in cluster 7, the participant in question contributes high above experiences throughout the game. This pattern can only be generated by a participant who wants to save her group, maybe also intends to educate the other group members. We have put both clusters into the same family to signal that the same behavioral program might also be at work in cluster 6, with the only difference being the susceptibility of the remaining group members to the benevolent intervention.

The characteristic feature of the third family of clusters (clusters 7–9) is contributions that start high, go low, and recover. In clusters 8–9, these choices are observed when experiences are stable at an intermediate level. In cluster 10, experiences slightly decay over time. None of the simulated types can rationalize these choice patterns. All three clusters are relatively small. We can, therefore, not exclude that participants did not follow a consistent choice program. Yet the behavior might be rationalized by exploration. Participants might want to find out, at a potential cost to themselves, whether they can make even more money by reducing contributions. Exploration would not be irrational as participants have no direct information about the choice programs followed by the remaining group members.

In the fourth family (clusters 11 and 12), experiences are stable at an intermediate level. The characteristic feature of this family is contributions that start high, fall reasonably low, and recover. Again none of the simulated types can generate this behavior. It could, as in the previous family of clusters, result from exploration. But there is also a motivational option for rationalizing these choices. Participants would initially be falling for the selfish temptation. But when they realize that the remaining members do not retaliate, they repent.

In the fifth family (clusters 13–17), not only choices zigzag; so do experiences. This behavior may result from a very narrow definition of conditional cooperation in a noisy environment. While we cannot exclude this interpretation, we deem an alternative explanation more likely. Participants refrain from formulating a choice program, and instead adopt the heuristic "follow the crowd".

The sixth family is fairly large. It encompasses clusters 18–25, and 467 participants. Experiences are fairly stable at an intermediate level. In most of these clusters, experiences are also relatively uniform, except for cluster 25 and, to a lesser extent, cluster 18. Clusters differ by the most characteristic individual choice pattern, i.e., the centroid. In cluster 25, one may argue that choices mirror experiences. Therefore, it is conceivable that choices have been motivated by a narrow reading of conditional cooperation, or by the heuristic "follow the crowd". Yet neither explanation works for clusters 18–24. We have already offered explanations for the choices in clusters 22 and 24. There are differences in the beginning. Contributions go down in clusters 18, 21, 23 and 25, and up in clusters 19, 20, and 22. An explanation for choices at the beginning of the game might be a stepwise adaptation to experiences. This would not be irrational. When they decide about their contributions, participants do not know how experiences will develop over time. They might cautiously move into the direction of experiences but would not do so entirely until they have received repeated information about the level of cooperation in their group. There are even more pronounced differences in the end. Contributions go down in clusters 18, 20 and 22, and up in clusters 19 and 23. In the end of cluster 24, contributions go first down and then up again. In the end of cluster 25, contributions go first up and then down. The fact that choices deteriorate in the end could be explained if participants interpret positive contributions as an investment in their group's cooperativeness. This would be the same motive as in the first family of clusters. The investment would only be much more moderate. The participant only refrains from exploiting her peers. The choice pattern in cluster 24 could result from a far-sighted freerider who repents. The upward trend in the end of cluster 19 is most peculiar. In the final period, choices even go above experiences. The raw data demonstrate that this upward trend is actually the characteristic feature of the entire cluster and not just noise. It could be explained by the intention to reward group members for their loyalty.

We have put cluster 26 in a family on its own as it has little resemblance to any other cluster. Experiences are reasonably stable, at an intermediate level. Yet choices first go up, almost from bottom to top. They stay above experiences for a while, start falling, and almost reach the bottom in the end. Forty-five participants are sorted into this cluster. The raw data show variance in the exact shape of contributions. But there is always a low beginning, a high plateau, and contributions gradually falling in the end. As choices stay high above experiences for quite some periods, the pattern is unlikely to result from some variant of far-sighted free-riding. One would need a very contrived subjective theory of the reactions of the remaining group members to explain why the participant in question deems it necessary to contribute that much, only to make sure that the contributions of the remaining group members do not fall below an intermediate level. Repent is also not a likely explanation. It could motivate why a participant overcompensates early exploitation. But would not explain why contributions steadily fall, and are way below experiences, by the end of the game. One consistent explanation would be this: initially, the participant holds the belief that, in this game, everybody will be selfish and hence contribute 0. She learns that this is not how others in her group behave. She needs the initial periods to be convinced that the cooperative behavior of others is indeed steadfast. She considers it unfair that, due to her error, she has exploited the remaining group members in early rounds, and wants to compensate them by even contributing more than others for several periods. Yet the motive for making high contributions would not be to save the group or to be overly generous. She just wants to make the rest of the group whole. Once this has been achieved, she starts increasing her payoff.

The eighth family comprises clusters 27–31. The family resemblance stems from the fact that contributions start high but fall over time, even to the bottom, in some clusters. There are clear differences between the clusters in this family, though, in the experiences these participants make. We have already discussed the biggest cluster in the family, cluster 31. It can be rationalised with a variant of far-sighted free-riding. The same explanation might hold a fortiori for cluster 30: in this cluster, individual contributions are even further below experiences from the fourth period on. It is rather unlikely that conditional cooperation would be that imperfect, while the pattern of choices fits a free rider who has come to the conclusion that sufficiently many loyal members will tolerate exploitation. Cluster 29 looks more like conditional cooperators losing faith, after an initial attempt at raising the contribution level. Participants put into cluster 27 are willing to exploit the group. But they do not contribute 0 right away. Instead, they gradually reduce contributions. This suggests yet another variant of five-sighted free-riding: participants test what they can get away with, without killing the cow they want to milk. Although choices in clusters 27 and 28 look similar, the level of experiences in cluster 28 calls for a different explanation. As participants initially contribute substantially above experiences, their beliefs about the reaction functions of the remaining group members would have to be reasonably contrived to make this choice pattern optimal for a participant who wants to maximise income. A more likely explanation is a participant who tries to raise the contribution level, realises that the group is not responsive, and stops investing in cooperativeness once she considers the cost of investment (by not maximising her own payoff) to be out of proportion, given the low number of periods to come.

In the ninth family, consisting of clusters 32–34, contributions also decay over time. We have already discussed the most significant cluster in this family, cluster 34, and have offered a free-riding cum learning story. This interpretation could also hold for cluster 33. But the behavioral pattern could alternatively result from participants who are cooperation-minded themselves and try to make others cooperate as well. In cluster 32, choices initially drop faster than experiences. But then the participant changes gears, to later again sharply reduce contributions. The initial reduction in contributions suggests selfish motives. The sizeable increase in intermediate periods would then have to be motivated by an attempt at convincing others to contribute more. The participant would reason: if the decay continues, I will make less money than if I create the impression for others that cooperation is worth the while. This reasoning would be consistent if the participant in question expects others to be reactive, be they pure conditional cooperators, or other far-sighted freeriders.

Cluster 35 is very different from all remaining clusters. Participants contribute nothing up to some point in time. Then contributions raise to the top, as do experiences. A consistent explanation is a group of conditional cooperators who are all overly pessimistic. As soon as one of them gives it a try, the remaining group members immediately jump on.

The penultimate family consists of clusters 36–38. We have already discussed cluster 36 and offered a prevoyant form of free-riding: participants invest in sustaining cooperation to expect a steady stream of income slightly above average. The same explanation might hold for cluster 38, while cluster 37 looks more like classic, short-sighted free-riding.

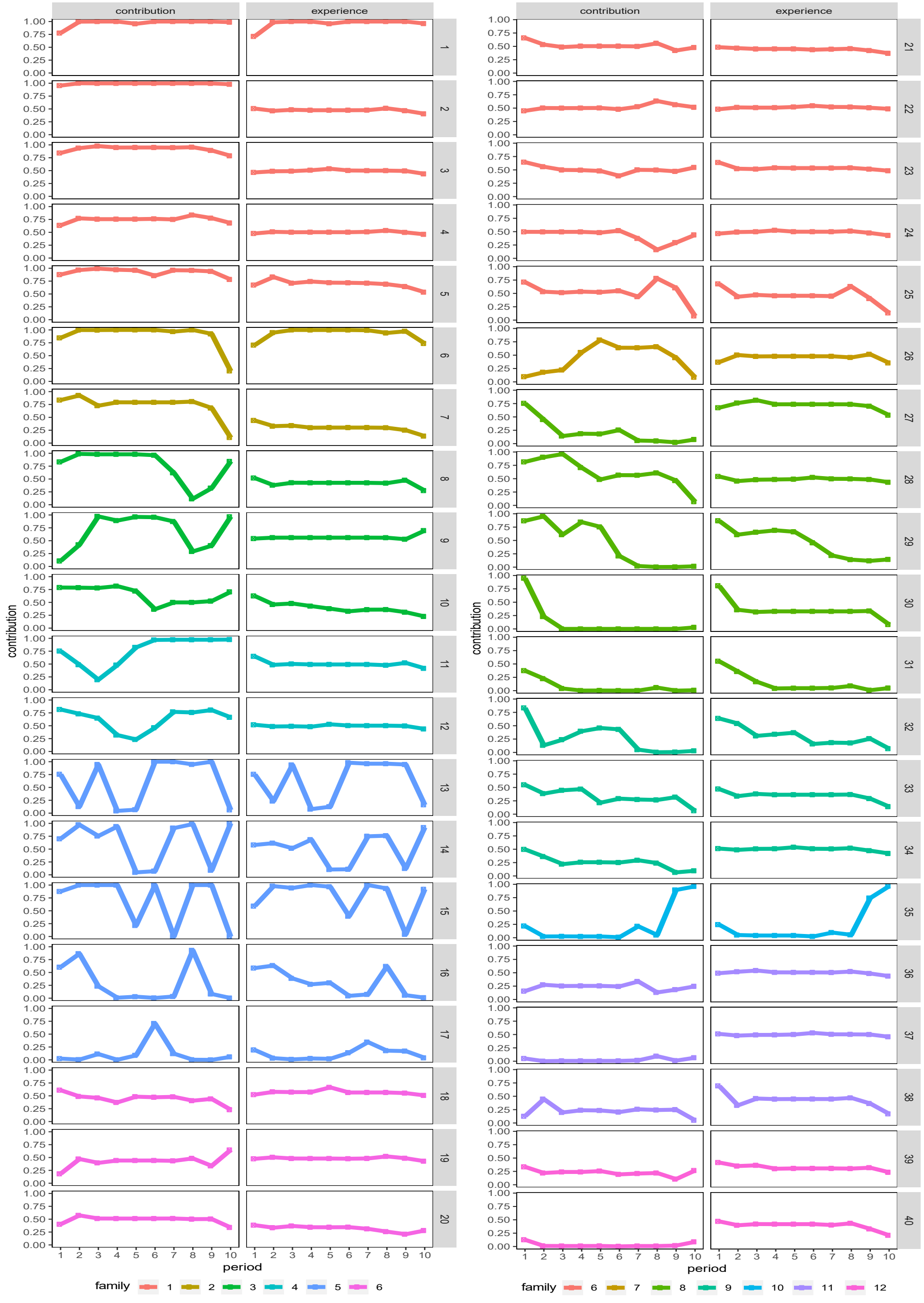
In the final family, consisting of clusters 39 and 40, experiences are fairly poor, and choices are even weaker. Except for the slight decay in the beginning, cluster 40 looks like straightforward, short-sighted free riding. Recall that this is the prediction of standard theory. Actually, the algorithm only puts 34 participants, or little more than 2% of the sample, into this cluster. Already cluster 39 requires a modified behavioral program. It could be imperfect conditional cooperation, or a variant of far-sighted free-riding.

## 3.7 Conclusion

The linear public good is one of the workhorses of behavioral economics. Hundreds of experiments have been run with this paradigm. The design is appealing as, in a stylized way, it captures what arguably is the essence of many conflicts of life, running from the degradation of the environment over the instability of a cartel to the precarious nature of any constraining institutional framework. The design implements a multi-person, multi-period prisoners dilemma with a known end. If one assumes that actors exclusively maximize individual profit, the repeated game has a unique solution. In the final period, all group members will contribute nothing to the joint project. Through unraveling, this is also the prediction for any earlier period.

The first experiments undertaken with this design have already refuted this prediction. On average, contributions start at some higher level but decay over time. Per se, social preferences can rationalize positive contributions, but they do not predict the decay. Interestingly, per se the prominent concept of conditional cooperation cannot predict the decay either. If all group members are perfect conditional cooperators and expect all others to follow the same behavioral program, any level of cooperation can be sustained, depending on initial beliefs. Fischbacher and Gachter (2010) propose a consistent explanation: the decay could result from conditional cooperation being imperfect. Participants would be willing to let themselves be guided by the level of cooperativeness in their group. But they would always try to undercut slightly. In their reanalysis of Fischbacher's and Gächter's data, Engel and Rockenbach (2020) have shown that true conditional cooperation is near-perfect. The decay results from heterogeneity. The combination of choice data with belief data shows that the decay results from the presence of short- and far-sighted freeriders. This is where the present project starts. It uses machine learning methods to cast light on this heterogeneity, and chart the type space.

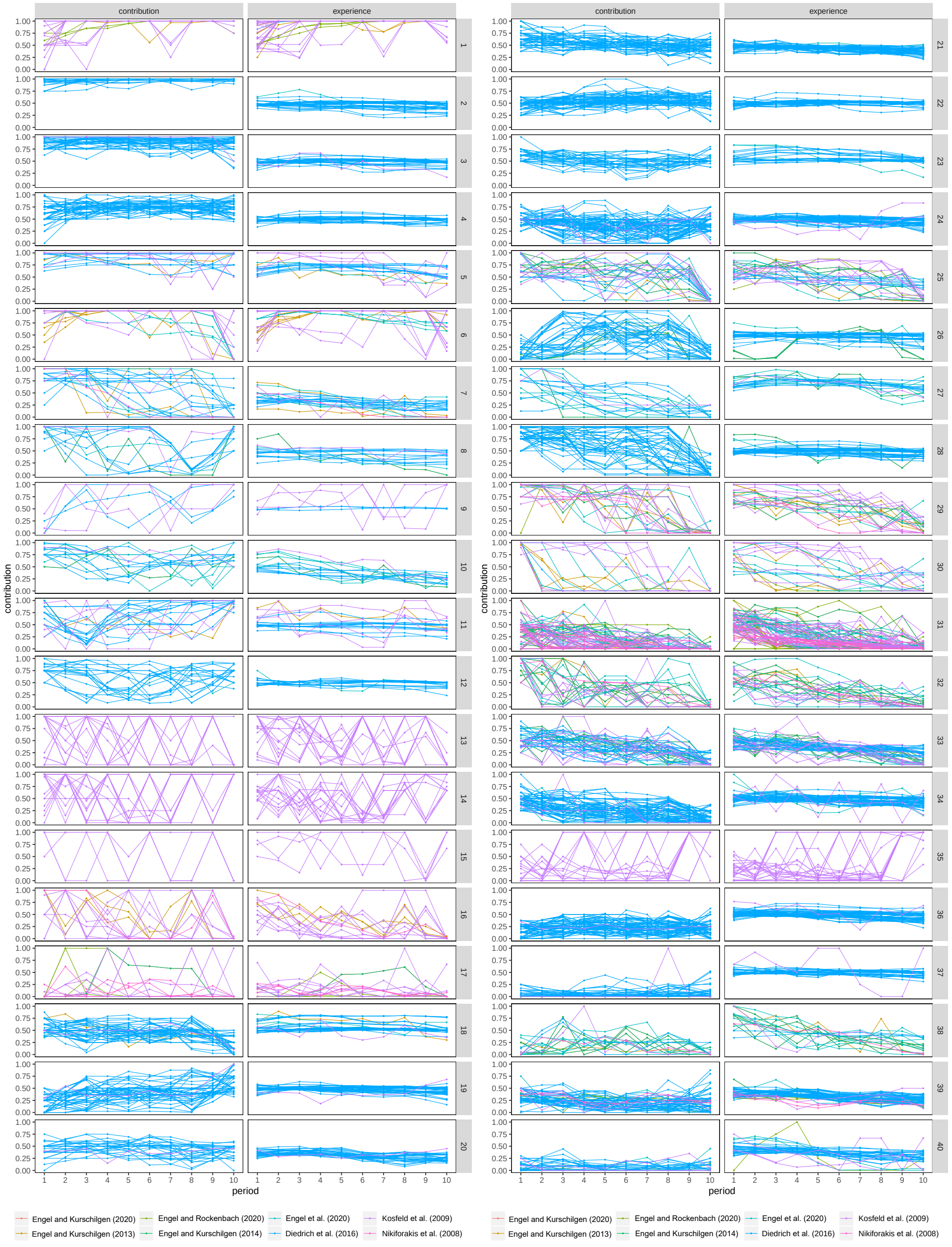
The paper makes a methodological and a substantive contribution. On the methodology side, it shows in which ways clustering can be used to infer the composition of the type space. On the substantive side, it shows that existing theories about behavioral types can



(a)

(b)

**Figure 3.4:** Clusters' Centroids in the Experimental Data



**(a)** **(b)**

**Figure 3.5:** Experimental Data Clustered

only explain a very narrow fraction of the data.

Repeated experiments generate time series data. In principle, the large family of algorithms for clustering time series data is therefore appropriate. Based on a pilot, we select an algorithmic configuration known to have excellent performance. Yet, contributions could not exhibit a downward trend unless at least some participants hold a reactive choice program. If we were to deprive the algorithm of the experiences that participants make, it would lump together choice patterns generated by entirely different behavioral programs. This is why we use multivariate clustering and feed the algorithm with pairs of experiences and choices.

One might naïvely think that the algorithm will find as many clusters as there are distinct behavioral programs. With simulation, we show why this approach must fail. We simulate all combinations of five behavioral programs that have been theorized in the literature: altruists, conditional cooperators, far-sighted freeriders, hump-shaped contributors, and short-sighted freeriders. For investigating these five behavioral programs, we need many more clusters. Yet we also show that we do not need the theoretical maximum of 350 clusters; this would approach next to unusable for real data, as one would need a considerable amount of data for that many clusters to be credible. We use internal cluster validation indices to determine that the appropriate trade-off between underusing and overusing the evidence is reached with approximately 40 clusters. Some of these clusters are indeed pure, in the sense that all combinations of experiences and choices stem from participants of the same simulated type. Yet many are not, as different behavioral problems may generate indistinguishable behavior. We do not see this as a limitation of the approach. Instead, it demonstrates what can be achieved with clustering: one is informed about distinct patterns and must discuss whether they could be generated with alternative behavioral programs.

We apply this methodology to a large dataset consisting of 16,474 observations. Results clearly show that the true type space is much more abundant than thus far assumed by the literature. Only very few of the clusters that we find in the experimental data can be rationalized with any of the five theoretical behavioral programs used to simulate data. We find practically no altruists and only very few outright selfish participants. Very few participants are near-perfectly reactive, as assumed by the canonical model of conditional cooperation. No cluster requires the assumption that players are hump-shaped, and perversely react to good experiences.

On the motivational side, the observed clusters suggest that some participants try to educate the remaining group members. Relatively many clusters can be rationalized by the intention to invest in the cooperativeness of the group (and to stop doing that if the opportunity cost no longer outweighs the expected benefit). Some participants seem to repent that they have exploited their peers, compensate them for the harm they have inflicted on them, or even want to reward them for their cooperativeness.

On the cognitive side, quite some choice patterns are consistent with exploration. Initially, participants are unsure about the behavioral programs implemented by their peers.

They only gradually adapt as they become more confident about the inferences they make. It seems that participants intending to exploit the group cautiously test which degree of exploitation will still be tolerated by the remaining group members. Other clusters suggest that participants choose by heuristic and mimic the experiences they have made.

The type space is considerably more abundant than typically assumed in the behavioral literature. Further work is needed to understand the observed behavioral patterns. In the next step, behavioral programs that could rationalize the observed patterns should be theorized and rigorously tested. This investigation is urgent if one hopes to learn from experimental data about the behavioral determinants of social dilemmata, in the interest of designing more powerful interventions.

*What is a moderate interpretation of the text? Halfway between what it really means and what you'd like it to mean?*

U.S. Supreme Court Justice Antonin Scalia

# 4

# Text Classification of Ideological Direction in Judicial Opinions

This paper draws on machine learning methods for text classification to predict the ideological direction of decisions from the associated text. Using a 5% hand-coded sample of cases from U.S. circuit courts, we explore and evaluate a variety of machine classifiers to predict "conservative decision" or "liberal decision" in held-out data. Our best classifier is highly predictive (F1 = .67) and allows us to extrapolate ideological direction to the full sample. We then use these predictions to replicate and extend Landes and Posner (2011) analysis of how the party of the nominating president influences circuit judges' votes.

## 4.1 Introduction

In the United States, judges wield significant power due to the common law system (Dainow 1966). The extent of U.S. judges' influence is a motivation for the extensive research into the determinants of judicial decision-making. In particular, there is a large literature on how opinions are affected by the ideology of the respective judge (Martin and Quinn 2002; Martin, Quinn, and Epstein 2005; Segal et al. 1995).

A leading paper in this literature is Landes and Posner (2011). This paper looks at how the party affiliation of U.S. circuit court judges affects the political ideology of their votes (conservative or liberal) on the court. While judges are nominally non-partisan, party affiliation can be proxied by the party of the appointing president or the party share in the Senate at the time of appointment. Landes and Posner (2011) show that judge party affiliation is statistically related to the ideological direction of votes.

For their empirical analysis, Landes and Posner (2011) draw upon the Songer database of U.S. circuit courts,[1] which provides rich metadata, e.g., the political ideology of votes for each judge in each case. The classification of votes by ideological direction was a labor-intensive exercise which has led to frequent use in the empirical legal studies and political science literatures (Ginn, Searles, and Jones 2015; Landes and Posner 2011; Reid and Randazzo 2016).

Notwithstanding its broad use in the literature, the Songer database has some limitations. First, the political ideology classification has been assigned by human coders, which could be error-prone. These errors add noise to regressions and complicate replicability. In particular, as noted by Landes and Posner (2011), the political positions of conservative/liberal are not constant over time. Therefore, data coded in the past may not be categorized correctly, and Songer Project ideology labels for older circuit court opinions may be systematically incorrect.

Another problem with the database is the sampling approach. First, the database is only available for 1925–2002, so that empirical analysis of vote ideology is only possible for that time period. Second, only a small set of cases was labeled (just 5% of the cases for those years). Finally, the authors used stratified sampling to get labels for similar numbers of opinions across courts and time. Therefore, the dataset is not representative of the full distribution of circuit court cases.

The goal of this paper is to address these shortcomings using machine learning and natural language processing techniques. The idea is to treat a machine to code the ideological direction of the votes. Within the set of labeled cases, we can check how well the algorithm replicates human labels.

The classifier would provide a number of benefits. As soon as the classifier is trained, predictions even for an extremely large sample cost very little relative to hand-labeling (which require a human to read an opinion). We could potentially take the classifier to cases before

[1] The original, as well as the extended versions, are available at songerproject.org.

1925 and after 2002. Within the 1925–2002 period, we could classify the other 95% of unlabeled cases. Besides producing new labels, the classifier could be used to audit and check existing labels for probable errors.

In this paper, we produce such a model. For the sake of interpretability, we focus on linear models. The model which worked best in our setting is a Ridge Classifier. Our model is trained on the complete opinion text in combination with the circuit, year, and case type data. After optimization it achieves a cross-validated accuracy of 61.5% on the three-label input and 66.5% on the two-label subset. The final calibrated classifier working on the two-label subset achieves the same accuracy score while increasing its precision as well its recall on the test set to 71.1% and 72.4% respectively.

With a validated dataset in hand, we use it to undertake an extended replication of Landes and Posner (2011). First, we do our best to replicate the original paper and, despite some problems in replicating the original dataset, we could replicate significance as well as the direction of the most important coefficients. We extend the results and probe their robustness to multi-way clustering, group, and additional covariates. Finally, we show that the results hold partly when using our machine-predicted ideological labels as the outcome.

This paper contributes to the emerging literature applying data science techniques to empirical legal research questions. We review some of that literature in section 4.2. After that, in section 4.3, we describe the supervised learning task to predict ideological labels in circuit court decisions. Next, section 4.4 reports the results of our replication study. Section 4.5 concludes.

## 4.2 Literature

This research sits at the intersection of two literatures. On one side, our paper is related to the research on judge ideology, which is focused on the positioning of judges, mostly for the U.S. Supreme Court (Epstein et al. 2012; Giles, Hettinger, and Peppers 2001; Ginn, Searles, and Jones 2015; Johnson, Songer, and Jilani 2011; Kassow, Songer, and Fix 2012; Martin and Quinn 2001; Masood and Songer 2013; Randazzo, Waterman, and Fix 2011; Reid and Randazzo 2016; Sturm and Pritchett 1949).

The judge ideology literature has taken two main approaches. The first approach is to hand-code cases by ideological direction. These include the Spaeth database for the Supreme Court and the Songer database for the circuit courts (Epstein et al. 2012; Giles, Hettinger, and Peppers 2001; Martin and Quinn 2001; Sturm and Pritchett 1949). The second approach is to use a latent factor model based on the voting behavior, to estimate a latent dimension for ideology based on judge agreement. This approach can identify median judges and the relative judge positioning on a scale over time (Martin and Quinn 2002).

The advantage of the first approach is that the scale is interpretable, exists on the case level, and relies on expert judgment. However, it is costly, and there are errors in coding. The advantage of the second approach is that it is cheap to compute for all judges, but the scale is not directly interpretable and does not exist at the case level. The scale also requires that judges vote in panels.

Our approach is something of a compromise, as we can form predictions for all cases and judges cheaply. It requires at least some hand-coding, but then can be applied to all cases. Methodologically, it is different because it uses the directly interpretable ideological labels of the hand-coded database. It does not assume a latent factor model, like Martin-Quinn. It also does not rely on contrasting votes of judges in a panel. This is relevant in our context because the large majority of decisions on the appellate courts do not have dissents. Voting behavior is not necessary, only some hand labels and the original opinion text.

The second literature to which we contribute is that on using texts as data for social science research. In particular, we contribute to that literature which uses texts to produce measures of ideology or partisanship. In law, an old study in this vein is Segal et al. (1995), who use texts from newspaper editorials as a proxy for the ideology of newly appointed Supreme Court judges. More recently, popular methods in political science for scoring ideology in text include Wordscores (Laver, Benoit, and Garry 2003), Wordfish (Slapin and Proksch 2008), and Wordshoal (Lauderdale and Herzog 2016). These tools use statistical differences in word frequencies by topic. They are most useful for text corpora for which differences in ideology come through in different words. As opinions of (lower) judicial courts are constrained in their (permitted) wording opinion texts may only satisfy that criterion in a very limited fashion.

In the legal domain, our paper is most closely related to literature predicting case type (Boella et al. 2012; Sulea et al. 2017; Undavia, Meyers, and Ortega 2018) and that concerned with dimensions in judicial texts (Ash and Chen 2019; Ash, Chen, and Lu 2018). The three papers closest to ours, in goal as well as methodological approach, are by Lauderdale and Clark (2014), Aletras et al. (2016), and Cao, Ash, and Chen (2018). In Lauderdale and Clark (2014), the authors use an LDA model to estimate how different issues at stake in cases are related to Supreme Court judges' voting behavior. The paper by Aletras et al. (2016) looks at decision direction of the European Court of Human Rights (ECHR) in regard to the violation of specific articles. The third paper, Cao, Ash, and Chen (2018), separates opinion texts into ideological and fact-driven parts, and look at how well these different paragraphs predict case directionality. However, none of the approaches in those three papers are viable for our goal or dataset. Lauderdale and Clark (2014) use the underlying text, but their focus on votes means that the approach is not applicable. In the case of Aletras et al. (2016), in a modelling perspective the approach is similar. However, their results rely on very clean data resulting in a very homogeneous directionality criterion. As a consequence, it is more than a simple question of transferring their results. Last, the paper by Cao, Ash, and Chen (2018) does look at ideological directionality. The focus on paragraphs, however, means that an additional labeling effort is needed while we seek to minimize the costs of classification.

To recap, our paper contributes in the technical literature to the understanding how to best implement a machine learning approach in the domain of judicial opinions. We aim to decrease labeling cost and increase scalability and reproducibility compared to the hand-labeling approach while at the same time improving explainability relative to the latent modeling approach.

## 4.3 Supervised Classification

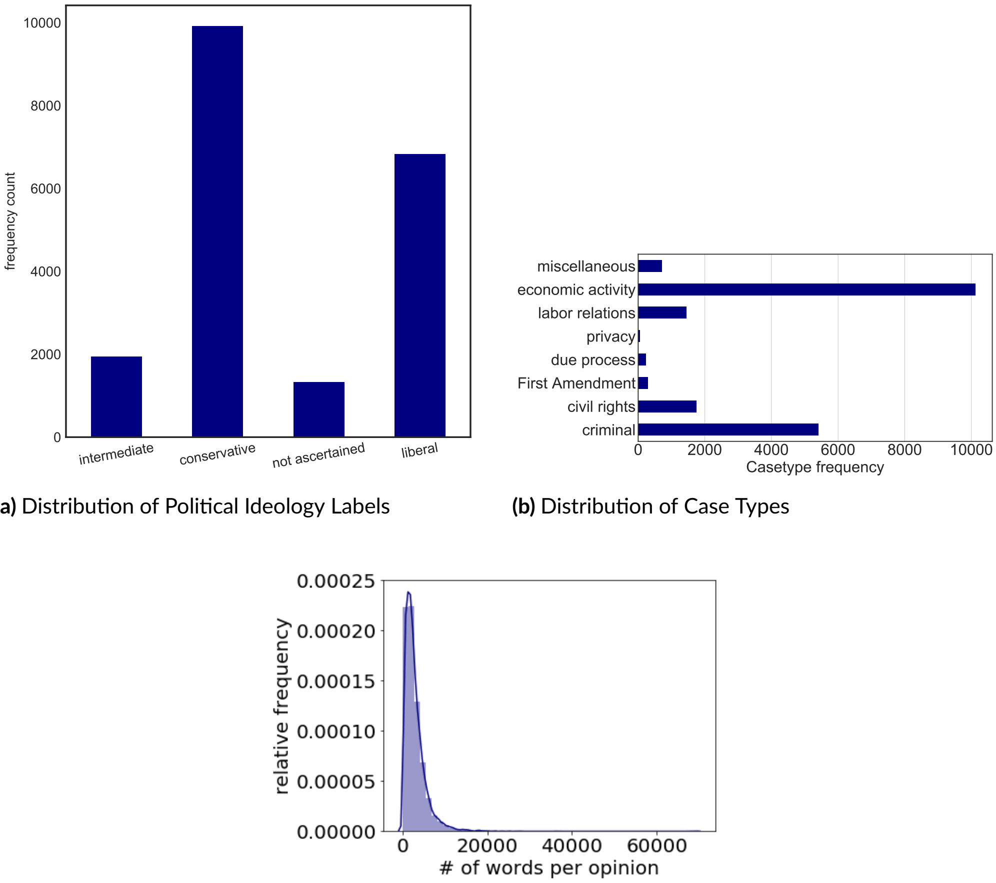
This section focuses on the classification algorithm which can reliably predict the political ideology of circuit court judges' written opinions. After training the algorithm on existing ideology labels, it can predict labels for unseen opinions. The beginning of this section provides information about the data necessary for classification. What follows is a detailed description of how the classifier is trained. Finally, the classification performance is evaluated.

### 4.3.1 Data

Broadly speaking, a supervised machine learning classifier maps an input to output. This section enumerates the datasets used for the inputs and outputs in our context. For our classification problem, we use the hand-coded ideology labels for these cases, provided by the Songer Project, as output. As input we use the U.S. circuit court judges' written opinions.

Songer Data on Decision Direction. The output or label of our classifier is the ideological direction of the opinion. As the number of circuit court judges' opinions is over 300 thousand, the Songer Project has annotated political ideology labels for only a small sample of opinions, equalling less than 2.6% of the total published opinions available. The total is $769,986$ when only taking those not decided per curiam into account. The cases were decided between 1925 and 2002 and the database contains a total of $20,355$ cases. Overall, four directionality codes are available: "liberal“, "conservative“, "mixed“ and "not ascertained“. While "mixed“ refers to the opinion of the case being of unclear directionality, "not ascertained“ signals that the coders were unable to assign a label according to the codebook's instructions. Please note that directionality is defined for each particular case type, with "conservative" and "liberal" being exactly opposite outcomes. Figure 4.1a shows the distribution of labels for the complete dataset. The categories "conservative" and "liberal" dominate, whereas the other two categories are underrepresented.

The Songer coders assigned the directionality of a case according to specific rules within case type. The case type of an opinion identifies the nature of the conflict between the litigants. Over 220 case type categories are organized into eight major categories: criminal, civil rights, First Amendment, due process, privacy, labor relations, economic activity and regulation, as well as miscellaneous. Figure 4.1b shows the distribution of the eight major categories for our dataset. "Civil rights" and "economic activity and regulation" are the two case types most frequent in the data.



**(a)** Distribution of Political Ideology Labels

**(b)** Distribution of Case Types

**(c)** Distribution of Opinion Length (in Words)

**Figure 4.1:** Summary Statistics

Landes and Posner (2011) mention in their paper that they applied substantial corrections to the raw Songer data, but those are not laid out in sufficient detail to reproduce. We approached the authors with the request to provide us with their version of the dataset. Unfortunately, they were not able to provide it yet.

Judicial Opinion Corpus. We matched the Songer dataset with the Lexis dataset, containing the full opinion text. With this approach, we could match 20, 052 opinion texts to the 20, 355 entries that the Songer database comprises. Regarding the non-matchable cases there is no clear pattern visible as these cases span nearly the complete time period as well as nearly all circuits. The distribution across time and circuits does not reveal any peculiarities either.

In terms of the matching itself, we subset the data according to the different circuits. That was only done for speed, as matching is a linear searching process which has to be repeated for each query. The actual matching was then done on either Federal Reporter citation or docket number. First, we tried to match via the normalized Lexis ID, i.e., the Federal Reporter citation when the opinion spanned more than one page in the Federal Reporter (to avoid confusion with other opinions). If such a match was not possible, we matched via the circuit court and the docket number. The reason why we preferred the Federal Reporter citation over the docket number is that the Songer database uses only encoded docket numbers, which are less prone to errors.

Figure 4.1c shows the distribution of opinions' word counts in our dataset. The shortest opinion consists of one word, the longest of 69, 320 words. The average opinion consists of 2, 809 words. As we use data from Lexis, each opinion had a specific structure. We extracted the text and split it into parts when encountering more than a single new-line character. Special characters such as "new-line" characters and Roman numbers were removed.

If a potential heading was found within the text, we excluded it, the reason being that such a heading would potentially include biasing information such as judge names. It is especially important to exclude those, as the model could focus on judge names as a proxy for the directionality since most cases were decided without dissent. This is an issue in our empirical context because we would like to use the predicted data to analyze judge characteristics. Including the judges in the prediction would induce mechanical correlation.

In a second step, we applied regular expressions trying to capture the part of the opinion in which judges might dissent from the majority. Including a dissenting part which by its nature goes against the directionality of the majority in the input would not only add noise but may also lead the classifier to average over the different directions, leading to an overall worse performance. When we found a dissent, we split off the relevant paragraph and saved it as an extra entry in the database, marking it as dissent. We excluded those entries and did not use them as input.

### 4.3.2 Model

This section describes how we deploy a supervised learning approach to predict the ideological direction of decisions from the association opinion text. Our approach, as outlined by Table 4.1, is quite uncommon in the literature of classifying a legal text's ideology. More traditional approaches, mainly used for ideology detection in political speeches, include word scores, word fish, or word shoal models. These approaches are either dictionary-based or require a reference text to which all other instances are compared. Our approach, by contrast, does not require one reference text to be selected and deploys more sophisticated selection mechanisms than naïve word counts.

One characteristic of machine learning approaches is their exploratory nature. We, too, test multiple combinations of data subsets, feature sets, models, and evaluation methods to find the best performing one. The instances to test are either selected by theoretical considerations, such as choosing only judicial quotations as predictive features; or they are chosen based on popularity, such as choosing support vector machines because they are known for their excellent performance on a broad range of NLP classification tasks.

All calculations were performed on the Max Planck Computing and Data Facility's high-performance cluster Draco, using one node of the type Broadwell with up to 40 CPUs and 256 GB memory. Moreover, each step relying on randomness was initialized with a pseudo-random seed for replicability. Our code most heavily draws upon functionalities provided by the python package sci-kit learn (Fabian Pedregosa et al. 2011).

**Table 4.1:** Construction of the Methodological Approach

| Subset of Data | Input | Text Preprocessing | Feature Engineering | Model | Model Evaluation |
|---|---|---|---|---|---|
| All case types, all 3 labels | Whole opinion | Remove: capitalization, punctiation, stopwords | bag of words | Passive Aggressive Classifier, Support Vector Machine, Logistic Regression, Ridge Classifier | Weighted scores: accuracy, precision, recall, f1-score |
| Case type "economic activity", all 3 labels | Quotations | | tf-idf | | |
| Case type "criminal", all 3 labels | Citations | Stemming | uni-grams | Multinomial Naïve Bayes | |
| All case types, 2 labels | | | bi-grams | | Alternative evaluation: regression results |

Subset of Data. In order to investigate how different categories or a differing number of labels affects a prediction, we constructed different subsets of the data for analysis. Four subsets constructed from the original data and used for this analysis are listed in the first column of Table 4.1. A naïve approach predicts political ideology labels regardless of case type. However, the naïve approach ignores the fact that directionality in the Songer data is assigned dependant on case type according to explicit rules differing for each case type. Subsetting the data by case type factors in this aspect of the coding scheme.

However, as Figure 4.1b shows, the dataset is heavily imbalanced in favor of the case types "economic activity" and "criminal". As the remaining case types are only marginally represented, we restrict the subset to these two case types, as only for them enough labeled observations to train the classifier are available.

Moreover, not only case type but also the labels are imbalanced. As Figure 4.1a shows, there is only a limited amount of observations available for the political ideology labels "not ascertained" and "mixed". We therefore derive two additional subsets. The subset "two labels" only includes the labels "conservative" and "liberal" as those two are not only the most frequent ones but also those we are most interested in. Especially if the remaining two labels ("not ascertained" and "mixed") are either considered as noise or wrongly classified, this subset should improve the classifier's performance. In particular, the exclusion of the label "not ascertained" is likely to not be problematic in any case: The number of cases labeled such is relatively low when compared to the other three labels. Moreover, the codebook shows that this label may be used in any case where it was not possible to assign one of the other three labels. This may either be due to the fact that the case truly fits into no other category or merely due to a lack in inter-coder agreement. However, past results show that such a sparsely represented, miscellaneous category decreases classification performance. For this reason, the final subset excludes this category altogether.

Input. We experimented with four different representations of the input. The most straightforward approach is to feed the complete preprocessed opinion text into the model. After screening a sample of randomly drawn opinions and cross referencing them according to the labeling instructions from the codebook, we identified two additional representations.

First, we separately extracted the citations from the cases. The topic, as well as the political directionality of a case, might be captured already by citations. Citation networks, e.g. used by the *Supreme Court Mapping Project*, is one example using this reasoning (Ash, Chen, and Lu 2018; Chandler 2007).[2]

Second, we extracted quotations from the text to serve as input. Many quotations immediately preceded citations. It is in the nature of a quotation that it represents the most relevant aspects to a matter at hand. As judges quote legal concepts from statutes and precedents relevant to the matter discussed, quotations, in turn, may be associated with either a "conservative" or a "liberal" leaning of the opinion.

The advantage of the whole opinion text as input is that no information is lost. Its downside, however, is that it may include more noise than only citations or quotations.

Text Preprocessing. For any data subset, the raw text needs to be preprocessed. We applied the prevalent practice of removing capitalization, punctuation as well as stop words. Furthermore, we reduced the words to their word stem, base, or root form (stemming).

Feature Engineering. The preprocessed text was tokenized, and the tokens were then used to form lists of n-grams (phrases) up to length three. N-grams extract information from text through local word order (Sidorov et al. 2014; Suen 1979). In the next step, these tokens were mapped to a numerical representation. We computed counts and frequencies over n-grams. The second specification is to weight the counts (tf) by inverse document frequency (idf), which upweights relatively rare words that could be more informative of topic or ideology.

Apart from converting opinion texts to vectors, we included the year the case was decided, the circuit at which the case was heard, and the case type as assigned by the authors of the Songer database to the feature set. Via grid search, we established which input and preprocessing combinations worked best, especially regarding single words versus n-grams.

Model. After vectorization, the next step is the actual classification of the text input, listed in the second last column of Table 4.1. In general, the classifiers may be grouped

[2] see SCOTUS Mapper Library by the University of Baltimore.

into two families, with the first being statistical methods. The advantages of this family are high explainability as well as being well-researched and understood (Ribeiro, Singh, and Guestrin 2016). The second family are deep learning algorithms mostly comprising some form of neuronal network architecture. In common NLP tasks, these algorithms outperform traditional algorithms (Kim 2014; Vaswani et al. 2017).

However, a downside to these models is that feature introspection, as well as explainability, are difficult. While there are attempts to develop methods for feature introspection, such as Shrestha et al. (2017) or Ribeiro, Singh, and Guestrin (2016), results so far have been preliminary. Consequently, we focus on well-researched statistical classifiers, maximizing thereby the explainability of the results. The classifiers we deploy are a passive aggressive classifier (Crammer et al. 2006), a logistic regression (Schmidt, Le Roux, and Bach 2017), a Ridge classifier (Rifkin, Yeo, and Poggio 2003), and a support vector machine with stochastic gradient descent (SGD) learning (Zhang 2004). All models are trained on a train-test split, stratified with respect to case type.

Model Evaluation. For model evaluation, we use standard performance metrics for machine learning, namely accuracy, precision, recall and F1 score (last column of Table 4.1).[3] The F1 score is the harmonic mean of precision and recall. As compared to accuracy, for example, it is more stable with respect to unbalanced datasets like ours.

Furthermore, in the context of this paper we consider precision as more important as recall because our dataset contains much less liberal than conservative cases. Thereby, we consider it as more important to actually find these few liberal cases and risk to classify some conservative cases as liberal.

As all performance measures are 5-fold-cross-validated, the scores reported are weighted averages. As the label space per category is heavily imbalanced in the validation set, accuracy has to be interpreted with care, and therefore the best performing classifier is selected by referring to the weighted F1 score. In our case, an additional model evaluation is the use of the predictions in the replication analysis below.

### 4.3.3 Evaluation of Results

In the following, we provide in-depth analysis across the different classification models introduced by Table 4.1.

Performance Metrics. Figure 4.2 depicts the performance metrics F1 score, accuracy, precision and recall for all models tested. Figure 4.2 shows that the scores depend more heavily on the subset-input combination than on the specific classifier used.

[3] While in traditional statistics measures such as the p-value are more prevalent, that measure is not appropriate in machine learning because we are trying to form accurate test-set predictions rather than to test for treatment effects. Moreover, the features in machine learning are often highly correlated, so the estimated coefficients for them are difficult to tease apart.

Based on this observation, we select four models to analyze and compare them in detail. Figure 4.2 depicts the model for each of the four subsets tested which reaches the highest F1 score. We report the accuracy, precision, recall, and F1 score respectively (coded by color, see legend). Each of the four groups of bars refers to a different subset of the data, for which we explored different modeling approaches. The top row looks only at the liberal and conservative votes, dropping the category "other". Second, we classify the full dataset with all three categories. Third, we limit the dataset to criminal cases. In the bottom row, we limit the dataset to economic cases.

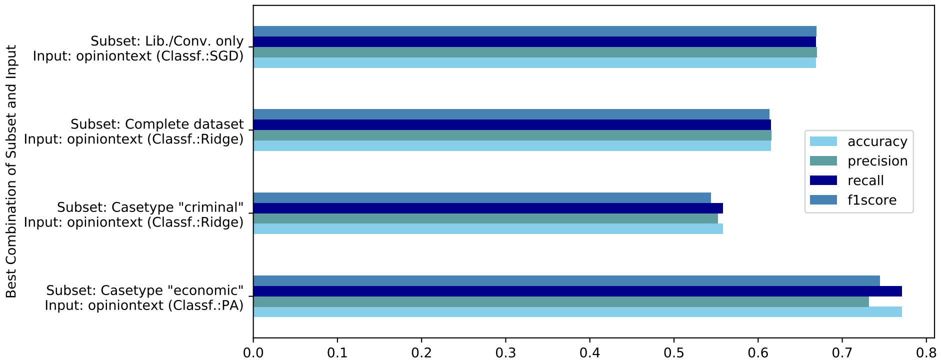


**Figure 4.2:** Best Performing Combinations by Subset

On the y axis, we indicate a feature that all four models have in common: They perform best on the input opinion text, rather than on citations or quotations. While additional calibration and tweaking of the model parameters would improve the performance of the classifier using either citations or quotations as input, the result is consistently outperformed when using the complete opinion text as input. This observation contrasts with the idea that citations or quotations would summarize the information in a meaningful way. However, instead of subtracting what we considered as noise, it seems that these input variations subtract important information.

As mentioned, the four subsets differ with respect to the subset of cases. Comparing the subsets concerning label, we differentiate between two- or three- label classification. The subset displayed at the top of Figure 4.2 takes two labels into account. A random guess, assuming a random distribution of labels, should yield an accuracy of approximately 50%. The model reaches an accuracy of 67.04%, lying clearly above this threshold.

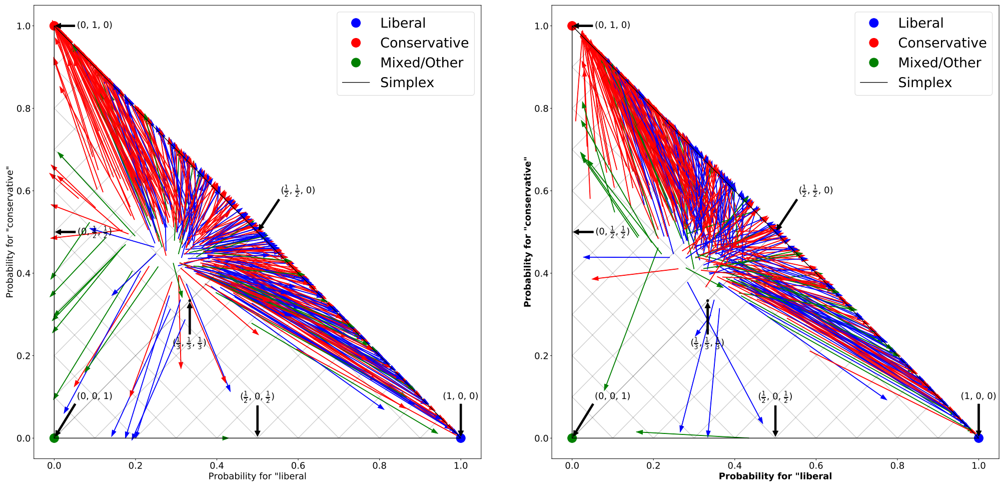
The second group of statistics are from the three-labels model. How much performance do we gain when predicting two instead of three labels? The two models at the top of Figure 4.2 show – only these two take all case types into account – an increase in accuracy from 62.00% to 67.04%. We believe that this increase in performance may offset the loss of information by excluding the label "mixed/other" as less than 1/7 of all cases fall into this category. This opinion is shared by other authors as well: Most studies drawing upon the

Songer/Auburn do exclude the “mixed/other” cases. However, for the sake of thoroughness we undertake the calibration presented in the following section for both the two- and three- label subset.

In the third and fourth groups of performance metrics, we show the three-label model but subset on case type. Interestingly, performance depends strongly on the case type. As mentioned in subsection 4.3.1, directionality is defined within case type while the number and quality of rules are quite distinct. Additionally, as Figure 4.1 shows, case type is heavily imbalanced in favor of economic rather than criminal. These two facts help to explain why the subset “criminal” only reaches an accuracy of 55.80% and, by contrast, why the subset “economic” achieves an accuracy of 77.10%. However, in order to increase generalizability, we instead opt to focus on classifiers trained on data containing all case types, as some results from, e.g., the case type “economic” may carry over to the case type “criminal”.

Probability Calibration. In the following, we analyze our classifiers’ calibration: Predicting a judicial opinion to either be conservative or liberal, we not only want to know the label but how confident the classifier is in assigning one particular label versus the other. In order to boost calibration, the classifiers were re calibrated using either a sigmoid or an isotonic calibration function.

The sigmoid function rests on a parametric approach based on the sigmoid model by Platt, Cristianini, and Shawe-Taylor (2000). The non-parametric isotonic variant is based on an isotonic regression.



(a) Ridge (Isotonic Calibration)

(b) SGD (Sigmoid Calibration)

**Figure 4.3:** Drift-Plots Showing the Change of Predicted Probabilities after Calibration

Figure 4.3 depicts the Ridge and SGD classifier respectively. For both classifiers, the calibration methods were applied for visualization purposes.[4] The three corners of Figure 4.3 correspond to the three classes: conservative, liberal, and mixed/other. Arrows point from the probabilities predicted by an uncalibrated classifier to the probabilities predicted by a calibrated classifier. For clarity of presentation, only each fiftieth data point from the test set is depicted.

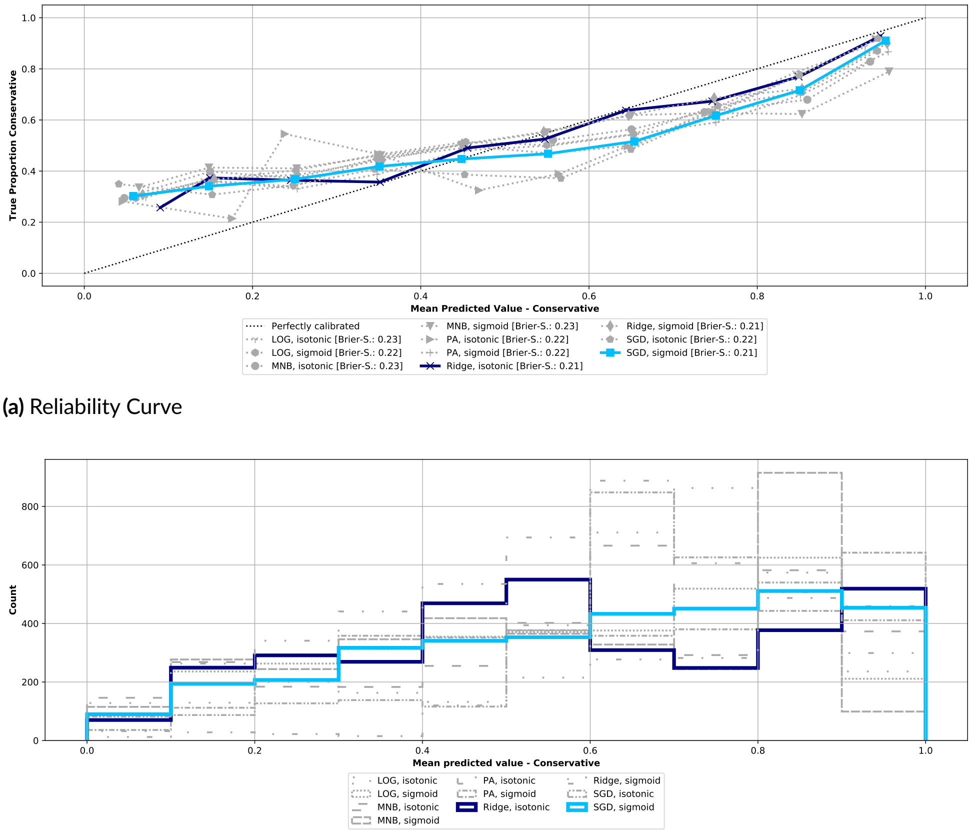
Figure 4.3 shows that calibration results in both classifiers shifting from under-confident to over-confident predictions, as the mass of predicted points moves away from the center of $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ towards the edges. This means that the classifier is likely to categorize similar cases very differently, as the predicted label is further away from the decision boundary for all cases. On the other hand, it also means that the classifier gets more confident about cases which are hard to classify. However, we accept this change, as both the absolute accuracy and the F1 score increase, although there may be an additional error for boundary cases.

While the two classifiers do not majorly differ in their confidence, they do differ in their error rate of assigning the label "liberal" to liberal cases. If one looks at the blue arrows, which depict cases for which the true label is "liberal", one can see that for the Ridge classifier (left panel) the mass of the blue arrows falls into the simplex spanned by the corner points $(\frac{1}{2}, \frac{1}{2}, 0)$, $(\frac{1}{2}, 0, \frac{1}{2})$,$(1, 0, 0)$. Every arrow point found within this simplex is classified as "liberal". Consequently, as the mass of blue arrows falls into that area, the mass of them is categorized correctly. In contrast, for the SGD classifier (right panel) a lower amount of the blue arrows falls into that area, meaning that the misclassification rate for "liberal" is higher. This means the precision for "liberal" is lower for SGD compared to the Ridge classifier. On the other hand, the inverse is true for the recall. As the original dataset features fewer liberal cases than conservative ones, on balance we might prefer to misclassify conservative cases as liberal instead of liberal ones as conservative. At this point, this speaks in favor of the Ridge classifier vs. the SGD classifier.

When looking at the "mixed/other" cases, we can see that the Ridge classifier classifies the majority of them correctly. However, that seems to come at the expense of misclassifying a disproportionally high amount of liberal cases. For the reasons stated above, we consequently exclude the "mixed/other" to gain performance in predicting only the labels "conservative" and "liberal".

Figure 4.4a provides another visualization to assess how well the probabilistic predictions of different classifiers are calibrated: It displays reliability curves which show the correct proportion of conservative cases (vertical axis) against the bins of predicted probabilities that a case is conservative (horizontal axis). The closer the reliability curve is to the 45-degree line, the better is the classification model's performance in terms of reproducing the original distribution. The Ridge classifier with isotonic calibration, as well as the SGD classifier with sigmoid calibration, are highlighted in shades of blue.

[4] Probability calibration was performed on data not used for model fitting. To this end, the training set consisting of 80% of the Songer data was cut in thirds and the model was then trained with 3-fold cross-validation. During this, 2/3 of the data were used for training and 1/3 was used for calibration. For each classifier the calibration algorithm yielding the best results was chosen.



**(a)** Reliability Curve

**(b)** SGD (Sigmoid Calibration)

**Figure 4.4:** Reliability Curves and Distribution Diagram

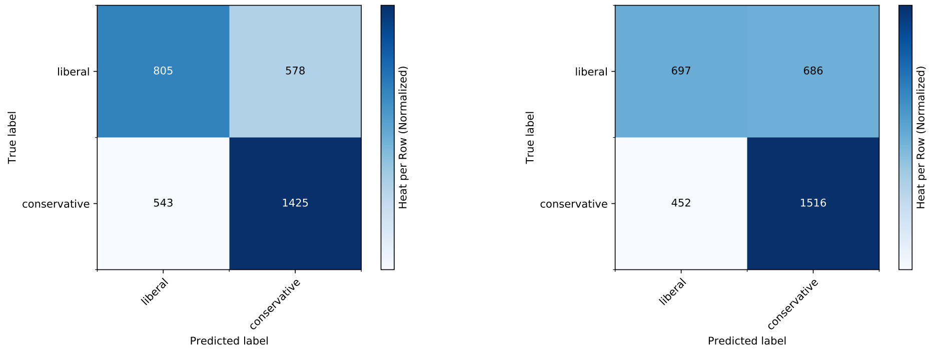
Consider the Ridge classifier: For all cases which it predicts to be conservative with a 20% probability, about 40% are actually conservative. In other words, it underestimates conservativeness. However, for cases close to the hyperplane (probability of .5 for either directionality), Ridge approximates the directionality distribution very well.[5] Finally, at around 70% likelihood, the classifier begins to overestimate the number of conservative cases.

Alongside Figure 4.4a, Figure 4.4b shows that despite calibrating the classifiers, a significant part of the predicted directionality's mass lies close to the decision boundary of .5. This, in turn, means that the classifiers have to be relatively precisely close to the decision boundary and be able to shift away mass from the decision boundary. Figure 4.4b shows that the two classifiers most successful in this are the Ridge classifier, calibrated with the isotonic algorithm, and the SGD support vector machine, calibrated with the sigmoid

[5] This is an important aspect as the Ridge classifier is similar to a support vector machine in that it uses the instances closest to the hyperplane for the separation of the data points.

algorithm.

Heatmaps. In the previous paragraph, we concluded that a two- label classifier for all case types will be the basis for predicting political ideology labels. In terms of performance metrics, the SGD classifier reaches the highest F1 score. However, the decision for the final model should not just take the F1 score but rather the types of errors that the classifier makes into account, as well. Therefore, Figure 4.5 plots normalized[6] confusion matrices for those two models deploying the best F1 score: the Ridge as well as the SGD classifier.



**(a)** Ridge Classifier

**(b)** SGD Classifier

**Figure 4.5:** Confusion Matrices for the SGD and Ridge Classifiers

As mentioned in subsection 4.3.2, we consider it as crucial to correctly predict as many liberal cases as possible even if some conservative cases are wrongly predicted as liberal. Figure 4.5b shows that as far as liberal cases are concerned, the SGD classifier predicts 697 cases correctly as liberal but almost as many cases (686) wrongly as conservative. The Ridge classifier displayed by Figure 4.5a, by contrast, predicts 805 liberal cases correctly as liberal and only 578 liberal cases wrongly as conservative.

Best Classifier. Based on performance metrics, heatmaps, and calibration results, we can select the classifier most suited for the task set out in this paper. The F1 score – our preferred performance metric – peaks both for the Ridge classifier, calibrated with an isotonic function, and for the SGD-classifier, calibrated with a sigmoid function. The second performance metric we consider as critical is precision, for which the Ridge classifier shows better results than SGD. In the same vein, the reliability curves show that Ridge is closer to the 45-degree line than SGD, which makes the former preferable. The only aspect where the SGD support vector machine slightly outperforms the Ridge classifier is in terms of mass, as shown in Figure 4.4b. However, overall, the difference in this regard is negligible.

[6] The normalized heat is calculated by dividing each value by the row mean.

Given this reasoning, we chose the Ridge classifier calibrated with the isotonic algorithm as model to perform out-of-sample predictions.[7]

### 4.3.4 Analysis

This section analyzes and interprets the predictions of the best two-label classifier. We look at predictions over time and by judge. We also interpret the model by examining predictive features.

Prediction of the Time Series in Decision Direction. Landes and Posner (2011) point out that the accuracy of the original Songer data is susceptible to the year in which a judge decided a case. Coders had more trouble coding older cases as compared to newer ones. We would like to see if this is reflected in differential performance of our classifier over time.

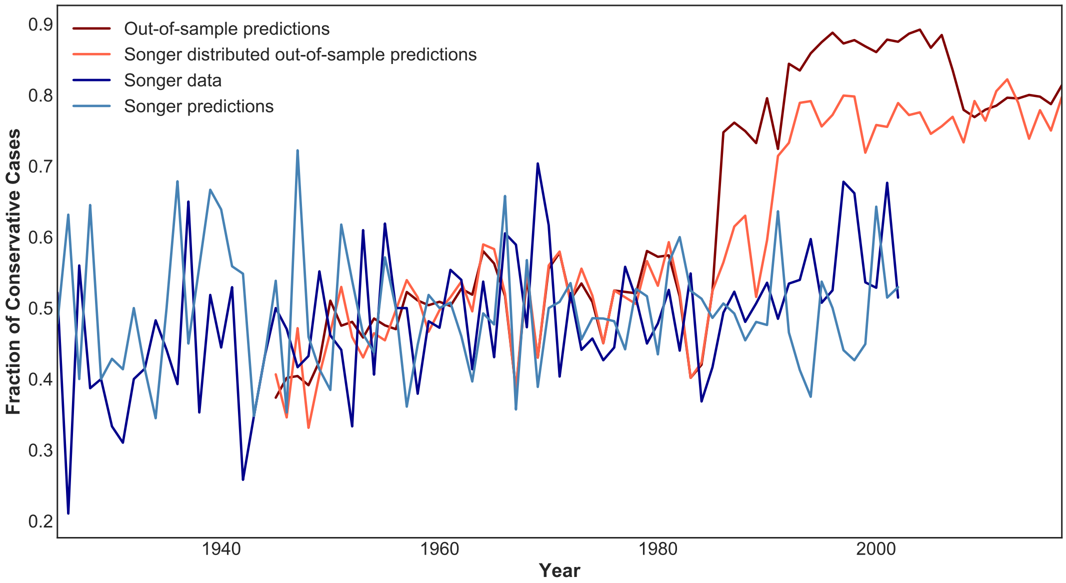


**Figure 4.6:** Fraction of Conservative and Liberal Cases, each Calculated for Actual as well as Predicted Case Directionality, Plotted by Year

Figure 4.6 shows the fraction of conservative and liberal cases by year for all circuits.[8] We include out-of-sample data which is made up of scraped Lexis data without the cases

[7] The final specifications of the classifier are as follows: We preprocess the text by excluding all stop words as well as punctuation. Following that, a lemmatizer is applied. This input transformed into bi-grams and then fed to a tf-idf-vectorizer. That vectorizer calculates the distance based on the “l2”-norm. It also makes uses the three additional features of year, circuit and case type. The regularization strength parameter $\alpha$ for the Ridge classifier is 2.0

[8] The cases categorized as “mixed” or “other” are excluded.

already within the Songer dataset. The original scraped dataset holds more than 1 million cases. As our classifier uses the year of the case, the circuit, and the case type as laid out by Songer,[9] these features have to be available for all out-of-sample cases as well. Especially the last one constrains the Lexis dataset because the case type was only available for cases of the years 1930 and later. Consequently, Figure 4.6 shows out-of-sample predictions only for those years.[10]

Figure 4.6 shows that for the in-sample predictions on the test set of the Songer data (20% hold-out data), the predictions closely approximate the original labels. This is also reflected in the high correlation of .73 ($\alpha < 1\%$). Especially for the years 1950 to 1980, the classifier performs very well. The out-of-sample predictions for that time period approximate the trend observed in the Songer data. Only for the years of 1980 onwards, the out-of-sample data (red line) is predicted to be considerably more conservative.

This spread may be caused, amongst others, by the classification error. Another reason could be the sampling process used by Songer and its team to construct the database.[11] To test this presumption, we plot a subset of the Lexis data constructed according to Songer's rules ("Songer-distributed out-of-sample", the orange line).

Indeed, we find that the orange and red lines diverge after 1980, with the orange line being closer to the original Songer data. This illustrates that indeed the sampling process heavily influences the distribution of decision directionality: As soon as the total amount of cases increases[12] by a significant amount, a spread appears. As the absolute number of court cases increased over time (Casper and Posner 1974), at least for cases after 1980, the Songer data may not be a good sample for the full set of cases. Consequently, the difference in out-of-sample predictions as compared to Songer predictions may simply stem from the fact that there is a structural shift in conservativeness (either in variation or trend) from 1980 onwards which is not represented by the Songer sample.

Directed Votes per Judge. Next we zoom in on particular judges. We look at performance for the ten judges who cast most of the votes in the Songer dataset, analyzing performance in civil and criminal cases separately. Those judges who did not hear both civil and criminal cases were excluded. The horizontal axis of Figure 4.7 indicates the true proportion of conservative votes while the vertical axis indicates the predicted proportion

[9] We matched the Lexis case types to the one laid out in the Songer database. However, the match has no bijective property. In order to get a reasonable good match, the subcategory case types of both, the Lexis data base as well as the Songer data base were used. This match is surjective with the Lexis subcategory case types as a base set. Then the matched Songer sub categories are aggregated to a Songer top category. Except for very few cases ($<$ 1000) this aggregation is unequivocal.

[10] If one is willing to forgo the performance gain introduced by the case type feature (about 2.5% points in the current configuration), one can predict directionality for all Lexis cases.

[11] For the original Songer database, at maximum 30 cases per year per circuit were sampled from all available cases after 1961. Before 1961, only 15 cases per year per circuit were selected.

[12] Where for the year 1945 only slightly more than 100 cases per year per circuit were coded with a usable case type in the out-of-sample dataset, for the year 2000 there are more than 2000 per year per circuit.

of conservative votes. Each point indicates these statistics for a single judge. If a judge's predicted behavior is the same as the truth, then his/her data point would lie on the dotted 45-degree line.

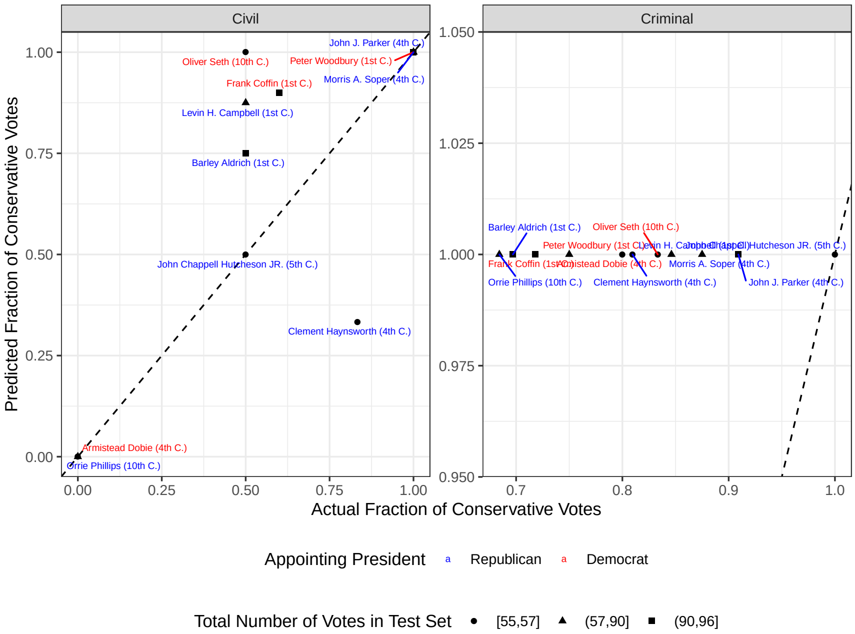


**Figure 4.7:** Fraction of Directed Votes per Judge – Comparison of Actual and Predicted Votes

Figure 4.7 shows that for civil cases, predicted and actual fractions are quite close. A $\chi^2$-test shows that the distribution of predicted fractions is not statistically different from the distribution of actual fractions (p-value $> .1$). For case type "criminal", however, the distributions of true and predicted fractions across judges are statistically different. The reason for this might be that the majority of criminal cases is labeled as conservative. Consequently, as the classifier uses the case type as feature, it can increase performance on criminal cases by labeling it as conservative. In other words, the classifier tends to overpredict the number of conservative cases in criminal law.

Feature Inspection. To further understand the two-label classifier, we investigate the features that are most important in driving our predictions. For this purpose, let *feature* be a feature, *value* be a value it could take, and *label* one of the ideological directions (conservative or liberal). We ranked the informativeness of each feature by the highest value of $P(feature = value | label = conservative)$ divided by $P(feature = value | label = liberal)$. Note that these are equivalent to coefficients from a Naïve Bayes Classifier.

The coefficients of the different features are represented by their standardized moments, meaning that normalization was performed by dividing through the standard deviation.

This means that each coefficient is on the same scale and therefore comparable. The hyperplane separating "conservative" from "liberal" lies at 0, meaning, a hypothetical case for which all the decision results would be zero falls into neither category. The higher the coefficient of a feature, the further away does a single feature move the case instance from the hyperplane when the feature is present within the case.

Table 4.2 lists the most informative features used by our best performing classifier. Please note that the *most informative* features for the label "liberal" are constructed such that they are *least informative* for the label "conservative". The features are either opinion-text phrases, quotation phrases, or citations.

Table 4.2 shows that the coefficients differ vastly in absolute size across the three different input variations. This corroborates the results of the metric scores. Especially for citation as input, the range of the coefficients' values is very narrow, with $-7.49$ being the minimum and 10.16 being the maximum. Consequently, many features loading clearly either the "liberal" or the "conservative" side are needed in order to have the case fall into a category. By contrast, the range of the coefficients' values for opinion text is much wider, with a minimum of $-57.67$ and a maximum of 189.96. A case including the words "reverse remand", for example, would be classified immediately as liberal. In essence, this means that features for the opinion text or quotations as input are more informative than for the citations.

The first column of Table 4.2a and Table 4.2b have the most predictive quotations. Quotations loading heavily on the label "conservative" are "knowingly" or "unique circumstances". The court quotes these phrases, i.e., they are singled out as relevant to the case at hand. Both phrases indicate a possible conviction. As the code book by the authors of the Songer database very often label a conviction as "conservative", this seems to be in line with the data provided. On the other side, the quotations for "liberal" are not as easily interpreted.

The second column of Table 4.2a displays those citations loading on the label "conservative". For the most heavily conservative citation, Humphrey v. Moore, the court limited the power of unions from infringing too far on employees of a company who were not part of the union. In Dandridge v. Williams, the court found that the state has some right to interpret how it puts into practice federal welfare laws. In consequence, Maryland was found not to be in violation of the Anti-Discrimination Act. Another conservative example would be United States v. Robinson, in which the court strengthened the police powers for searches during lawful arrests under the Fourth Amendment.

In comparison, the second column of Table 4.2b features citations which the classifier finds to be indicative of a liberal case. The most indicative citation would be United States v. Taylor, a case in which the bar for conviction on charges of conspiracy was raised. Coppedge v. United States dealt with the fact that the sentenced petitioner had not received the plenary review of his conviction to which he is entitled, and all his appeals against his conviction against this ground were dismissed. The Supreme Court reversed the decision to dismiss his appeal and generally strengthened defendants, rights in this regard. In the same vein, Green v. United States reversed the sentencing of the defendant under the Fifth Amendment as he was put in jeopardy twice for the same offense. Consequently, while absolute size of the coefficients for citations hint at only a limited quality for the overall classification into either “liberal” or “conservative”, the cases as such seem to fall into the right domain.

The last column shows the predictive phrases from the full opinion text. Features such as “judgment affirm” or “plaintiff appeal” are predictive of the label “conservative”. In line with those, but not shown here, are the features “affirm judgment” and “appeal dismiss” on place 11 and 14 respectively. This is in line with labeling rules as set out by the Songer team for criminal cases, where the coding rules state that affirming the decision against an appellant is to be coded as conservative. Conversely, within the most predictive features for “liberal”, one can find “reverse remand”, “remand proceeding”, or “reverse case”, reflecting that predictive features seem to be driven by criminal cases.

**Table 4.2:** Best Predictive Features

**(a)** Best Predictive Features for Label “Conservative”

| | quotations (Ridge) | | citation (Ridge) | | opiniontext (Ridge) | |
|---|---|---|---|---|---|---|
| | coef | feature | coef | feature | coef | feature |
| 1 | -17.13 | knowingly | -7.49 | Humphrey_v_Moore | -57.67 | motion new |
| 2 | -13.18 | John_Doe | -7.43 | Dandridge_v_Williams | -53.71 | plaintiff argue |
| 3 | -11.97 | unique_circumstances | -6.59 | SEC_v_Chenery_Corp | -51.91 | prior art |
| 4 | -11.47 | X | -6.42 | Co_v_Zenith_Radio_Corp | -50.86 | appellant claim |
| 5 | -11.40 | No | -6.19 | Dalehite_v_United_States | -50.78 | grant motion |
| 6 | -11.03 | minor | -6.06 | Brady_v_Maryland | -49.45 | plaintiff appellant |
| 7 | -10.85 | search | -5.60 | United_States_v_Robinson | -48.85 | plaintiff contend |
| 8 | -10.63 | attractive_nuisance | -5.55 | Mal_v_Riddell | -45.70 | fiduciary duty |
| 9 | -10.09 | may | -5.38 | Port_Gardner_Investment_Co_v_U | -45.62 | plaintiff appeal |
| 10 | -10.04 | overhead | -5.25 | Olim_v_Wakinekona | -44.01 | judgment affirm |

**(b)** Best Predictive Features for Label “Liberal”

| | quotations (Ridge) | | citation (Ridge) | | opiniontext (Ridge) | |
|---|---|---|---|---|---|---|
| | coef | feature | coef | feature | coef | feature |
| 1 | 19.98 | that_where_the_State_has_provided_an_opportuni... | 10.16 | Yes_v_United_States | 189.96 | reverse remand |
| 2 | 19.86 | Motion_for_Judgment | 9.18 | United_States_v_Taylor | 133.90 | remand proceeding |
| 3 | 19.57 | fairer_to_those_adversely_affected_by_a_bond_f... | 9.11 | ...Inc_v_Commissioner | 103.28 | case remand |
| 4 | 19.16 | take_care | 9.09 | Townsend_v_Sain | 98.70 | remand district |
| 5 | 18.30 | urge_that_the_indictment_charged_the_maintenan... | 8.88 | United_States_v_Young | 89.69 | government argue |
| 6 | 17.32 | good_faith | 8.43 | Dennis_v_United_States | 85.99 | remand new |
| 7 | 17.30 | anything_of_value | 8.21 | Coppedge_v_United_States | 84.05 | proceeding consistent |
| 8 | 16.76 | crack a_little_bit_of_time_to_research_on_the_... | 8.15 | ...Inc__v_United_States | 75.33 | consiStatent opinion |
| 9 | 16.76 | a_little_bit_of_time_to_research_on_the_backgr... | 8.00 | Green_v_United_States | 74.29 | new trial |
| 10 | 15.49 | clear_and_convincing | 7.97 | Brown_v_Board | 60.13 | reverse case |

## 4.4 Replication and Robustness Checks

This section focuses on the replication aspect of Landes and Posner (2011). For comparison, all tables and figures that Landes and Posner (2011) produced with data of circuit courts are listed in Table D.1, section D.1. The most relevant tables for our purposes are Tables 11 and 13 as numbered in the original paper.

Summary Statistics. This paragraph compares our summary statistics listed in Table 4.3b to those by Landes and Posner (p. 803, 2011) listed in Table 4.3a. As can be seen, the statistics differ. We count a total of 56, 602 cases; Landes and Posner (2011) count 55, 041 cases. Furthermore, we count more opinions classified as "conservative" or "other" than Landes and Posner (2011) do.

One possible explanation for these diverging results is that not all of the corrections that Landes and Posner (2011) applied in the original paper were described in sufficient detail so that they could be reproduced. We were able to apply the corrections concerning political ideology, but we were unable to apply judge-related corrections. Landes and Posner (2011) briefly mention judge-related corrections and refer to a website for a detailed description. This website, however, is no longer available online.

**Table 4.3:** Court of Appeals Votes by Subject Matter and Ideology for 538 Court of Appeals Judges Only: 1925–2002

**(a)** Original by Landes and Posner (2011)

| | Crim | Civ Rts | First | Due Proc | Priv | Labor | Econ | Misc | Total |
|---|---|---|---|---|---|---|---|---|---|
| Conservative | 6823 | 2721 | 566 | 461 | 117 | 1351 | 9361 | 525 | 21925 |
| Liberal | 1876 | 1766 | 477 | 201 | 67 | 1922 | 9884 | 559 | 16752 |
| Mixed | 635 | 460 | 89 | 51 | 13 | 420 | 1775 | 22 | 3465 |
| Other | 5321 | 210 | 102 | 79 | 3 | 179 | 6047 | 958 | 12899 |
| Total | 14655 | 5157 | 1234 | 792 | 200 | 3872 | 27067 | 2064 | 55041 |

**(b)** Replication

| | Crim | Civ Rts | First | Due Proc | Priv | Labor | Econ | Misc | Total |
|---|---|---|---|---|---|---|---|---|---|
| Conservative | 7217 | 2647 | 397 | 412 | 83 | 1397 | 11084 | 478 | 23715 |
| Liberal | 1911 | 1755 | 379 | 176 | 38 | 0 | 10375 | 596 | 15230 |
| Mixed | 613 | 473 | 86 | 48 | 9 | 423 | 1689 | 31 | 3372 |
| Other | 5652 | 212 | 40 | 24 | 3 | 2232 | 5177 | 945 | 14285 |
| Total | 15393 | 5087 | 902 | 660 | 133 | 4052 | 28325 | 2050 | 56602 |

Regression. Next we replicate the primary regression analysis of circuit court judges in Landes and Posner (2011), focusing only on the essential part of their analysis. For Table 13, we replicate the regressions focusing on the fraction of conservative votes and only taking the period from 1925 to 2002 into account.[13] Regarding the baseline regression, Landes and Posner (2011) specify their regression model as follows:

$$FrCon_{ij} = \beta_0 + \beta_1 X_i + w \tag{4.1}$$

where $FrCon_{ij}$ denotes the fraction of conservative votes, calculated as votes per judge over the sample period. $X_i$ encompasses several judge characteristics such as the party of the appointing president, share of Republican senators at the time of nomination, year of appointment, gender, race[14], prior experience as a district judge, and judge-circuit fixed effects[15]

According to Landes and Posner (p. 810, 2011), their regressions are weighted either by the judge's total votes in civil cases or the total votes in criminal cases. Furthermore, Landes and Posner (2011) do not specify how they compute their standard errors, but we assume that they use heteroskedasticity-robust standard errors (treating each judge as an observation) and therefore use errors of that type for the replication.

Civil Cases. With Table 4.4, we provide our first replication table, dealing with civil cases only. Column (1) corresponds to Landes and Posner (2011) Table 13, column (6).[16] As in the original paper, we report the t-statistics, rather than standard errors or p-values, for all coefficients in parentheses. Landes and Posner (2011) do not specify how they computed standard errors for their regression Table 13, but we inferred that they used heteroskedasticity-robust errors.

The main research interest of Landes and Posner (2011) was whether judges follow their party affiliation in their decisions. They find a significant influence of being appointed by a Republican president *(RepPres)* on the fraction of conservative votes for civil cases (Table 4.4, column 1). Our result for civil cases (Table 4.4, column 2), is quite similar when compared to Landes and Posner (2011)'s; in our data, being appointed by a Republican is associated with a positive and significant effect of voting conservatively in civil cases. The evidence for a relationship between party and ideology actually appears to be stronger in our replication than implied by the original study.

[13] In turn, this means that we do not display results for the fraction of liberal votes, as displayed in columns (2) and (4) of Landes and Posner (2011) Table 13, nor do we report results for the period of 1960–2002 as reported in Table 14.

[14] Race is a dummy for Black = 1, 0 else

[15] The judge specific data was acquired from the Auburn database by Gary Zuk, Deborah J. Barrow and Gerard Gryski onhttp://www.songerproject.org and then matched to the Songer data by a judge identifier code.

[16] These are the columns with the "uncorrected" data. We only compare uncorrected data as Table 4.3 showed that we were not able to replicate even summary statistics for the corrected version.

Apart from deploying heteroskedasticity-robust errors, we propose a model specification with multi-way clustering (non-nested) as recommended by Badir Alnidawy (2015). Based on the advice from Abadie (2020), we add two-way clustering by circuit and year. This allows for correlation in the error term across judges within court over time, and across courts in the same year as well. Clustering leaves coefficients unchanged, and a comparison of columns (2) and (3) reveals that t-statistics only differ slightly as a result of the two-way clustering.[17]

While Landes and Posner (2011) grouped the data on judge level, we additionally run the empirical analysis with data at the vote level. This specification allows us to control for case characteristics with circuit-year fixed effects. For getting at the effect of party affiliation on ideology, this is an important step econometrically because the number of Republican-appointed judges and the proportion of conservatively decided cases could be correlated over time due to unobserved confounding factors.

The dependent variable is now binary. It equals one for conservative decisions and zero for liberal decisions (cases with the mixed/other category are dropped). The vote level regression model both includes circuit-year fixed effects, and clustered standard errors by judge and year. This specification successfully replicates the significant positive effect of a conservative appointing president (*RepPres*) on the fraction of conservative votes.

Model specifications (5) and (6) are estimated not only with hand-labeled but also with predicted data. The predictions on which estimation results of columns (5) and (6) are based, were generated with a calibrated Ridge classifier.

These re-estimations serve as an alternative way to assess the performance of the classifier. The rationale behind this procedure is that generating labels is not the ultimate goal, but using these labels in an empirical model is. Therefore, even if the classifier cannot predict political ideology with an accuracy of 100%, its performance can be viewed as appropriate if the results of the empirical model do not change drastically when estimated with the classifier's predictions.

As far as column (5) is concerned, using predicted instead of hand-labeled data does not change the results for coefficients *RepPres*. Estimating the vote-level fixed effects model with predicted labels instead of hand-labeled (column 6) results in estimates for RepPres that are no longer statistically significant.

Criminal Cases. With Table 4.5, we provide our second replication table; it deals with criminal cases only. Landes and Posner (2011) found a positive and significant influence of being appointed by a Republican president *(RepPres)* on the fraction of conservative votes. Our result for criminal cases is quite similar to Landes and Posner's, our coefficient being slightly larger.

[17] We provide regression results with errors clustered on the year of appointment, circuit court, and the party of appointing president in Table 4.4, column (3).

Furthermore, for criminal cases, Landes and Posner (2011) found a negative effect of appointment year. However, we do not find such an effect. They also report a negative impact of being black (*Black*) on crime conservatism, which we replicate. Two-way clustering changes t-statistics only slightly. This leads to no change in significance level for the coefficient *(RepPres)*, but it left the coefficient *(Black)* to no longer be significant.

The fixed effects multi-way clustering model on vote level data replicates the significant and positive effect of the party of the appointing president (*RepPres*) as well as of being black (*Black*) on the fraction of conservative votes.

The multi-way error component model using predicted data could not reproduce the significance of the coefficient *RepPres*. Instead, being male turned to have a significant negative impact on criminal conservatism. The fixed effects multi-way clustering model on vote level with predicted data could neither reproduce the significance for coefficient *RepPres* nor *Black*.

Extreme Bounds Analysis. The extreme bounds analysis (EBA) is a sensitivity test that examines how robustly the dependent variable of a regression model is associated with a variety of possible determinants (Hlavac 2016). We estimate an EBA, including all possible combinations of independent variables that Landes and Posner (2011) specified. To limit the influence of coefficient estimates with high multicollinearity, we follow the recommendations by Hlavac (2016) and specify the maximum acceptable variance inflation factor to be 7. Next, we increase the weights of those regression models that better fit the data – that is, by its likelihood ratio index according to Mcfadden (1974).

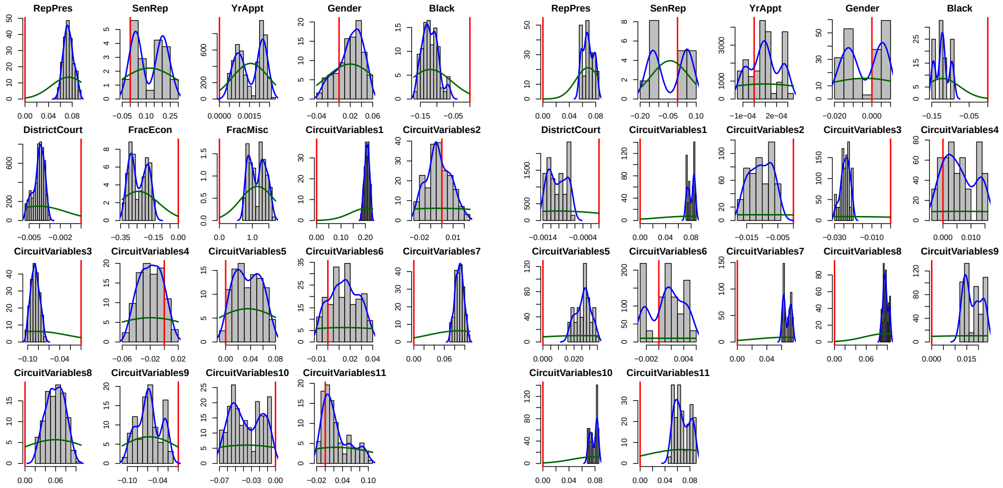
Figure 4.8 shows histograms for each of the independent variables included in the model. The green curve displayed in each histogram is a density curve which approximates the coefficients' distribution with a normal distribution.

A positive coefficient indicates that holding all else equal, a higher value of the examined variable is associated with a higher fraction of conservative votes. On the other hand, if most of the area of the histogram's bins lies to the left of zero, higher values of the corresponding variable are associated with a lower fraction of conservative votes.

For the civil cases, Figure 4.8a suggests that when the appointing president *(RepPres)* is Republican (rather than Democrat), when the judge was appointed in later years *(YrAppt)*, and when the specific judge participated in a higher fraction of miscellaneous votes *(FracMisc)*, a judge's fraction of conservative votes increases. Furthermore, circuits 1 and 7 are consistently associated with a higher fraction of conservative votes.

Being black *(Black)*, having served more years as a district judge *(DistrictCourt)*, and an increasing fraction of economic votes *(FracEcon)*, are associated with a lower fraction of conservative votes. Furthermore, circuits 3, 9, and 10 have a lower fraction of conservative votes.

To conclude the visual inspection as well as the interpretation of the statistics, found in section D.5, the EBA for civil cases suggests that the variables *RepPres*, *FracMisc* and circuit



**(a)** Civil Cases

**(b)** Criminal Cases

**Figure 4.8:** Histograms Extreme Bounds Analysis, for Civil and Criminal Cases

1 are very strongly associated with the dependent variable.

For criminal cases, Figure 4.8b shows that being appointed by a Republican (rather than Democrat) president (*RepPres*) is consistently associated with a higher fraction of conservative votes for all regression models estimated. Furthermore, circuits 1, 5, 7, 8, 9, 10, and 11 are associated with a higher fraction of conservative votes.

By contrast, being black (*Black*) and having served more years as a district court judge (*DistrictCourt*) decrease the fraction of conservative votes. Furthermore, circuits 2 and 3 are associated with a lower fraction of conservative votes.

To conclude the visual inspection, EBA results for criminal cases suggest that the variables *Pres*, *Black* as well as circuit 8 and 10 are robustly associated with the fraction of conservative votes.

## 4.5 Conclusion and Outlook

This paper had two main goals. Our first goal was to replicate the analysis on circuit courts proposed by Landes and Posner (2011), and to add multiple robustness checks to assess the validity of the regression model initially specified. Second, we show an approach for extending the dataset used in the original study via machine learning, especially in regards to the input used for any future algorithm.

As far as replication of the empirical analysis of Landes and Posner (2011) is concerned, we were able to reproduce the most critical findings. The robustness checks found, just as Landes and Posner (2011) did, that the party of the appointing president and being black influences the fraction of conservative votes. We find that the result for party affiliation is actually stronger than the original article found, as it extends to both civil and criminal cases.

What explains our different results? We paid particular attention to the code generating the fraction of conservative votes. As multiple reshaping and grouping operations as well as joining different datasets were necessary in order to obtain this variable, its calculation is not exactly trivial. We can imagine that a small mistake in the original code by Landes and Posner (2011), such as an inner instead of an outer join, could change the fraction. In turn, its association with the dependent variable may also change.

However, we could not replicate the exact summary statistics of the dataset Landes and Posner (2011) used because they did not provide replication code and did not sufficiently specify their corrections in the original paper. That, in particular, may affect the rest of their findings.

In order to extend the dataset, we experimented with different classifying algorithms, where the best one was a passive-aggressive classifier for economic cases, reaching a F1 score of 74.49%.

In order to assess the validity of the classification, we compared the regression results obtained by using predicted data to those obtained by using only hand-labeled data. Coefficients found to be significant with the replication as well as with the robustness checks were not replicated with the predicted data, suggesting that that 1) the classifier still needs improvement, or 2) researchers should be careful with using predictions as data in downstream empirical analysis. Future research should, therefore, take into account that the distribution of the Songer data in regards to cases per circuit per year does not mirror the distribution of the universe, and as such it may skew the predictions of any classifier. Oversampling is only an imperfect correction for this issue, as is the inclusion of the circuit or year as a feature. Otherwise, the consistency of results may not be guaranteed.

One aspect that we neglected thus far is that predictions cannot be directly plugged into a regression without correcting for the classification error. Fong and Tyler (2020) proposed one approach to do so. However, Fong and Tyler (2020) describe a case in which one or more independent variables are predicted. In our case, however, we predict the dependent variable. Therefore, we propose to develop a correction approach in order to prevent forward propagation of the prediction error used within a dependent variable which at this point may be of the main reason for failure.

Furthermore, the distributions of the enlarged dataset and that one of the original data are significantly distinct. Overall, the classifier was trained on roughly 5% as compared to the number of labels that were predicted. As soon as such a considerable disbalance is present, non-random draws or the lack of stratification is very problematic. Lack of stratification is the case with the original Songer database, i.e., Songer (1993) does not keep the original distribution of cases per circuit as they focused on preserving other aspects such as the presence of all circuits in each year.

Taking the above into account, our results provide a concise groundwork for future research in this area. First, in order to establish a ground truth that goes beyond mere statistical significance and also looks at distributional aspects, more than just regression results are needed. Here, we suggest that multiway error component modeling as well as an extreme bounds analysis should be used on any prior results before trying to take them as a baseline for any extension of the Songer database.

Secondly, in regards to machine learning, we show quite clearly that any input which does not include the complete opinion text in some form cannot result in a good overall performance. That is important as it shows that other aspects which are otherwise very useful in the domain of law, such as citations for citation networks, do not contain enough information for this specific task. This holds despite the fact that when using citations as input, the classifier uses many citations to which it assigns the correct ideology label if one were to label them by hand. However, when taken as an aggregation, neither citations nor quotations are distinctive enough.

Moreover, while the Songer database features four labels, our results show that the error the classifier makes on the label "mixed" is nearly equally split between "conservative" and "liberal". As the label "other" is negligible in terms of occurrence, we can, therefore, conclude that training a classifier only on the two labels "conservative" and "liberal" does not introduce any systematic bias. Due to the increase in performance, such a setup should consequently be preferred. Lastly, looking at the regression results, it may be that text alone is not enough. Future research should therefore also think about taking meta-information, such as the circuit court it was heard at, into account.

Moreover, looking at the literature of the median judge (Martin, Quinn, and Epstein 2005), it may also be important with which other judges a judge sits on a panel. This may be another important aspect, a machine learning classifier may have to take into account.

We hope that our work acts as a baseline on which future work can build on. The obvious next step is to scale back on the interpretability of the model in favor of sophistication: Specifically, we propose a modified doc2vec model in combination with an attention mechanism. Furthermore, future work could stack multiple classification algorithms tailored more closely to the rules of the coding book that the Songer database provides.

Another exciting avenue for future work is to compare in depth the differences, advantages, and disadvantages of various methodological approaches. A particular exciting comparison is a Bayesian framework, as proposed by Martin and Quinn (2002), compared to machine learning approaches, as suggested by this paper.

Apart from methodological extensions, a more content-related one is particularly interesting: Most of the literature is targeted towards high ranking courts, such that the Supreme Court or circuit courts. This lack of attention towards lower courts might stem from the fact that the universe of cases to code is vast. Consequently, not even a partially coded dataset, as far as political ideology labels are concerned, is available for lower courts. A classifier, trained on circuit courts' opinions could predict the label for opinions of lower courts and, by that, help to close this particular gap in the literature.

**Table 4.4:** Regression Analysis of Court of Appeals Votes: 1925–2002, Civil Cases

| | Dep. Variable: Fraction of Conservative Votes | | | | | |
|---|---|---|---|---|---|---|
| | *true data* | | | | *predicted data* | |
| | Landes (2009) | replicated | multi.clus | vote | multi.clus.pred | vote.pred |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| RepPres | 0.035$^{***}$ | 0.069$^{*}$ | 0.069$^{***}$ | 0.092$^{***}$ | 0.032$^{**}$ | 0.031 |
| | (3.860) | (2.125) | (4.136) | (3.821) | (2.942) | (1.417) |
| SenRep | 0.072 | −0.017 | −0.017 | 0.095 | 0.004 | 0.219 |
| | (1.710) | (−0.090) | (−0.347) | (0.647) | | (1.677) |
| YrAppt | 0.0003 | 0.001 | 0.001 | 0.0003 | 0.001 | 0.001 |
| | (0.790) | (0.665) | (1.237) | (0.202) | (0.796) | (0.431) |
| Gender | −0.006 | 0.015 | 0.015 | −0.026 | −0.0004 | −0.058 |
| | (0.260) | (0.344) | (0.318) | (−0.681) | (−0.011) | (−1.384) |
| Black | −0.028 | −0.105 | −0.105 | 0.007 | −0.125 | −0.001 |
| | (1.180) | (−1.505) | | (0.124) | | (−0.023) |
| DistrictCourt | 0.002 | −0.004 | −0.004 | −0.002 | −0.002 | −0.0005 |
| | (0.330) | (−1.455) | (−1.183) | (−1.712) | (−0.345) | (−0.417) |
| FracEcon | −0.090 | −0.230 | −0.230$^{**}$ | 0.355$^{**}$ | −0.249 | 0.451$^{***}$ |
| | (1.640) | (−1.506) | (−2.690) | (2.774) | (−1.918) | (3.531) |
| FracMisc | −0.049 | 1.345$^{*}$ | 1.345$^{*}$ | −0.920 | 1.464$^{***}$ | −0.324 |
| | (0.350) | (2.442) | (2.107) | (−1.842) | (6.118) | (−0.673) |
| circuit FE | yes | yes | no | no | no | no |
| circuit-year FE | no | no | yes | yes | yes | yes |
| Observations | 535 | 498 | 498 | 4169 | 498 | 4169 |
| R$^2$ | 0.240 | 0.119 | 0.119 | 0.047 | 0.123 | 0.066 |

$^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$. Linear regression with heteroscedasticity robust standard errors. Variables: *RepPres*: Party of the appointing president, conservative or liberal (omitted category); *SenRep*: Share of republican senators at the point of election; *Gender*: sex of the judge, male or female (omitted category). *Black*: dummy for the race of the judge; *DistrictCourt*: Years spent as a district judge; *FracEcon*: Fraction of economic votes; *FracMisc*: Fraction of miscellaneous votes; *Circuit Variables*: All regressions include 11 dummy circuit variables – circuits 1 to 11. The D.C. court is the omitted circuit variable.

**Table 4.5:** Regression Analysis of Court of Appeals Votes: 1925–2002, Criminal Cases

| | Dep. Variable: Fraction of Conservative Votes | | | | | |
|---|---|---|---|---|---|---|
| | *true data* | | | | *predicted data* | |
| | Landes (2009) | replicated | multi.clus | vote | multi.clus.pred | vote.pred |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| RepPres | 0.056$^{**}$ | 0.077$^{***}$ | 0.077$^{***}$ | 0.051$^{**}$ | 0.038 | 0.005 |
| | (4.220) | (3.634) | (3.811) | (3.022) | (1.734) | (0.829) |
| SenRep | −0.076 | −0.151 | −0.151 | 0.010 | −0.020 | 0.078$^{**}$ |
| | (1.090) | (−1.399) | | (0.141) | (−0.542) | (2.844) |
| YrAppt | −0.001$^{***}$ | −0.00001 | −0.00001 | −0.0003 | 0.001$^{**}$ | −0.001$^{**}$ |
| | (3.390) | (−0.023) | (−0.032) | (−0.601) | (2.876) | (−2.709) |
| Gender | −0.014 | −0.019 | −0.019 | 0.010 | −0.023$^{*}$ | −0.012 |
| | (0.710) | (−0.740) | (−0.876) | (0.545) | (−2.219) | (−1.750) |
| Black | −0.057$^{*}$ | −0.091$^{*}$ | −0.091 | −0.081$^{**}$ | −0.020 | −0.027 |
| | (2.060) | (−1.814) | (−1.047) | (−2.717) | (−0.257) | (−1.697) |
| DistrictCourt | 0.001 | −0.001 | −0.001 | 0.0003 | −0.001 | 0.001 |
| | (0.140) | (−0.817) | (−0.390) | (0.360) | (−0.346) | (1.917) |
| circuit FE | yes | yes | no | no | no | no |
| circuit-year FE | no | no | yes | yes | yes | yes |
| Observations | 523 | 498 | 498 | 13543 | 498 | 13543 |
| $R^2$ | 0.240 | 0.084 | 0.084 | 0.019 | 0.052 | 0.014 |

$^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

Linear regression with heteroscedasticity robust standard errors.

Variables: *RepPres*: Party of the appointing president, conservative or liberal (omitted category); *SenRep*: Share of republican senators at the point of election; *Gender*: sex of the judge, male or female (omitted category). *Black*: dummy for the race of the judge; *DistrictCourt*: Years spent as a district judge; *FracEcon*: Fraction of economic votes; *FracMisc*: Fraction of miscellaneous votes; *Circuit Variables*: All regressions include 11 dummy circuit variables – circuits 1 to 11. The D.C. court is the omitted circuit variable.

# A
# Appendix, Chapter 1

## A.1 Data Overview

We collected samples at the behavioral economics laboratory[1] located at the Friedrich Schiller University of Jena as well as at the Decision Lab[2] located at the Max Planck Institute for Research on Collective Goods in Bonn. Table A.1 compares the two samples collected.

**Table A.1:** Session Overview

| | Jena (2018–11) | Bonn (2019–06) | overall |
|---:|---:|---:|---:|
| observations | 88 | 146 | 229 |
| sessions | 8 | 12 | 20 |
| color-blind | 1 | 2 | 3 |
| failed | 1 | 2 | 3 |
| share females | 0.60 | 0.61 | 0.61 |
| mean age | 22.17 | 25.75 | 24.41 |

## A.2 Favorite Color

This sections exhibits that mouse movements are unlikely to be driven by other factors than the payoff related to the color. To control for possible other motives of a participant's

[1] https://experiment.wiwi.uni-jena.de/public/

[2] https://www.coll.mpg.de/124252/decision-lab

choice of color, we asked them to report their favorite color. Table A.2 shows that most participants stated not to have a favorite color. 12.93% of the participants actually reported their favorite color. A binomial test shows that the probability of successes (favorite colors reported) is not significantly different from a random choice (p-value = .32).

**Table A.2:** Count of Participants' Favorite Colors

| | blue | green | brown | pink | yellow | grey | none |
|---|---|---|---|---|---|---|---|
| count | 30 | 36 | 2 | 14 | 16 | 37 | 53 |

## A.3 Archetypes

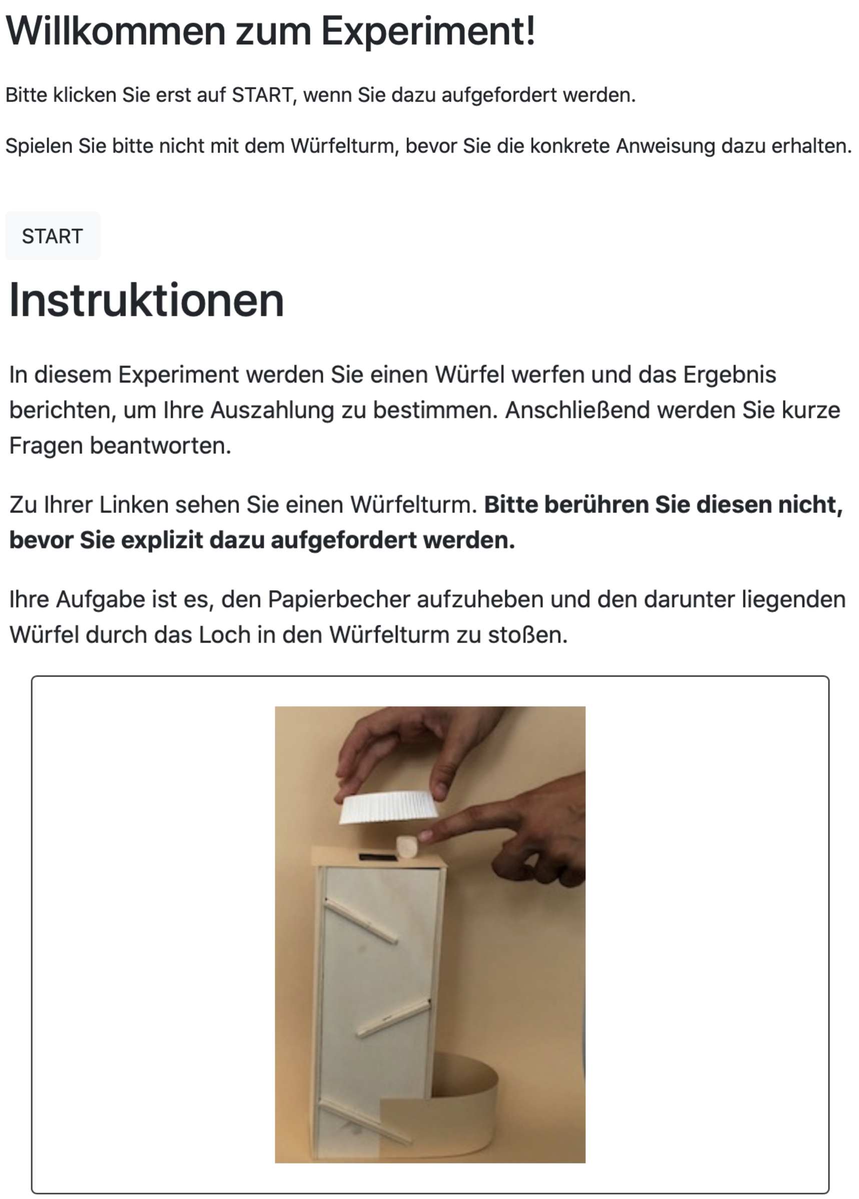
Table A.3 provides an overview of the interpretation of the four behavioral archetypes found.

**Table A.3:** Interpretation of Behavioral Archetypes by Type of Die Rolled

| | number | color |
|---|---|---|
| A | maximize payoff; report face rolled | report face rolled |
| $A = \Omega$ | report face rolled; temptation, guilt | report face rolled; temptation, curiosity |
| $A < \Omega$ | maximize payoff | maximize payoff |
| $A > \Omega$ | report face rolled; guilt | report face rolled; guilt |

## A.4 Screenshots of the Experiment

In the following, we provide screenshots of the experiment. Screenshots are taken for treatment $C^P$ which denotes rolling colored dice and reporting under time pressure.



# Willkommen zum Experiment!

Bitte klicken Sie erst auf START, wenn Sie dazu aufgefordert werden.

Spielen Sie bitte nicht mit dem Würfelturm, bevor Sie die konkrete Anweisung dazu erhalten.

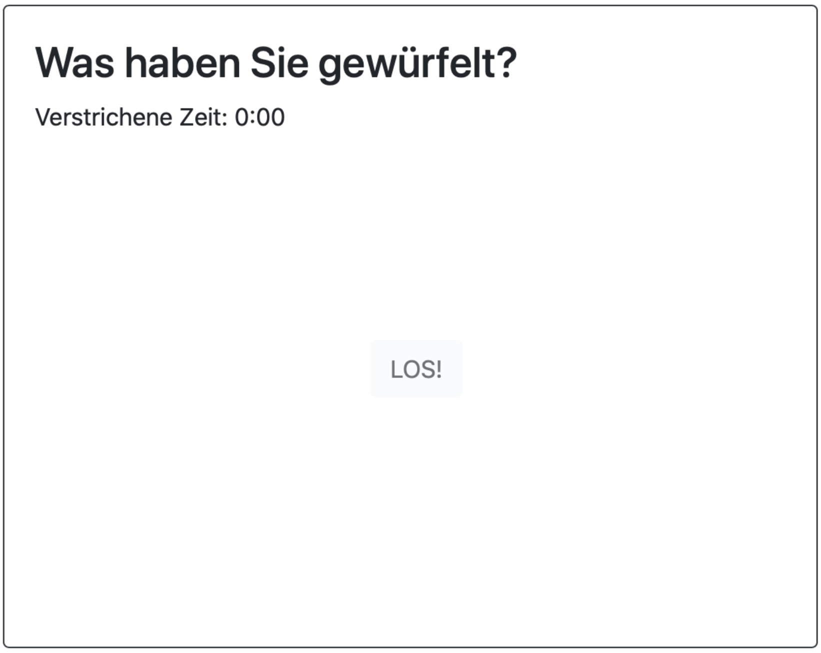
# Instruktionen

In diesem Experiment werden Sie einen Würfel werfen und das Ergebnis berichten, um Ihre Auszahlung zu bestimmen. Anschließend werden Sie kurze Fragen beantworten.

Zu Ihrer Linken sehen Sie einen Würfelturm. **Bitte berühren Sie diesen nicht, bevor Sie explizit dazu aufgefordert werden.**

Ihre Aufgabe ist es, den Papierbecher aufzuheben und den darunter liegenden Würfel durch das Loch in den Würfelturm zu stoßen.

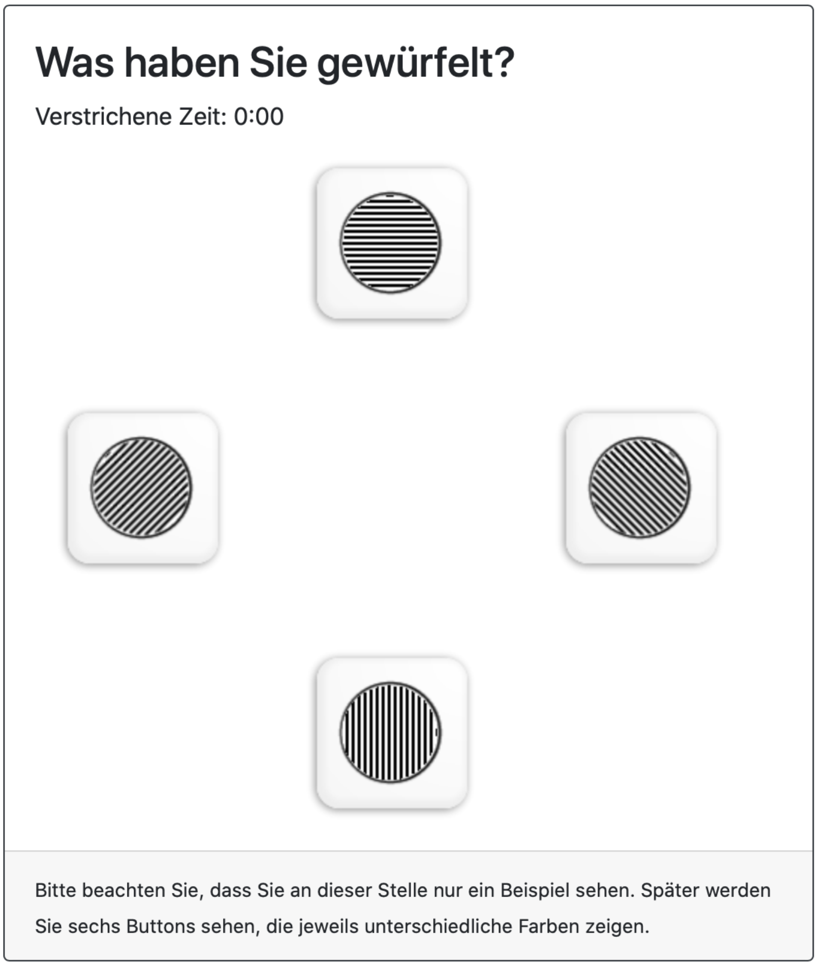
Nachdem Sie gewürfelt haben, erscheint folgender Bildschirm:



Was haben Sie gewürfelt?

Verstrichene Zeit: 0:00

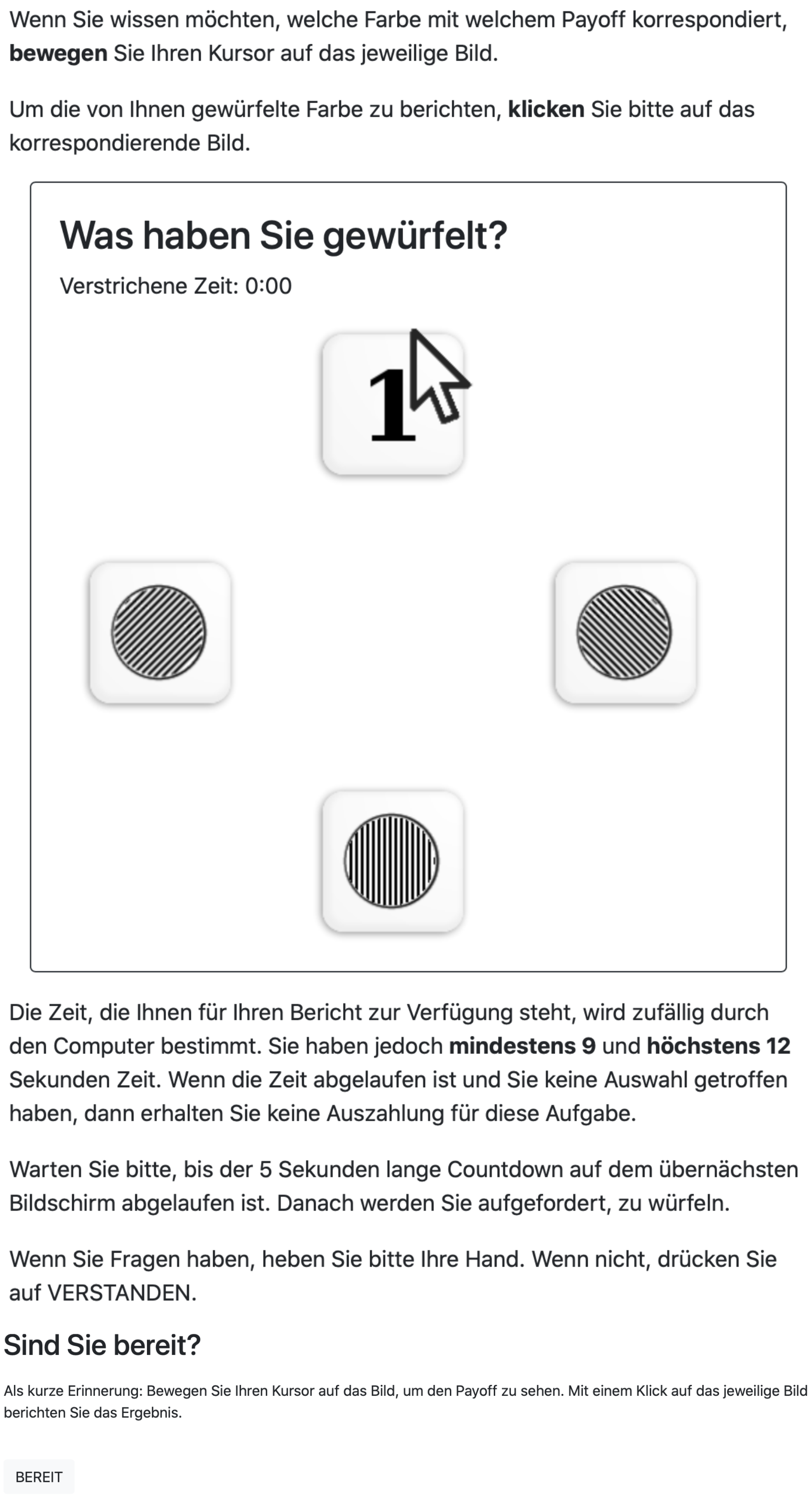
Drücken Sie auf LOS!, um zu folgendem Auswahl-Menü zu gelangen:



## Was haben Sie gewürfelt?

Verstrichene Zeit: 0:00

Bitte beachten Sie, dass Sie an dieser Stelle nur ein Beispiel sehen. Später werden Sie sechs Buttons sehen, die jeweils unterschiedliche Farben zeigen.



Wenn Sie wissen möchten, welche Farbe mit welchem Payoff korrespondiert, **bewegen** Sie Ihren Kursor auf das jeweilige Bild.

Um die von Ihnen gewürfelte Farbe zu berichten, **klicken** Sie bitte auf das korrespondierende Bild.

## Was haben Sie gewürfelt?

Verstrichene Zeit: 0:00

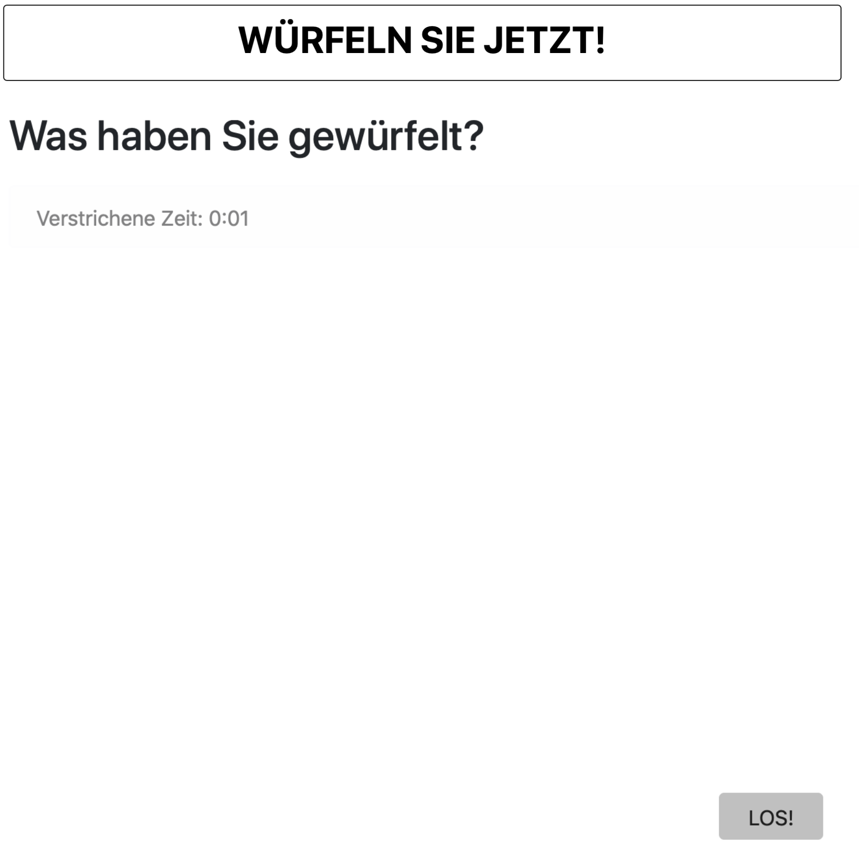
Die Zeit, die Ihnen für Ihren Bericht zur Verfügung steht, wird zufällig durch den Computer bestimmt. Sie haben jedoch **mindestens 9** und **höchstens 12** Sekunden Zeit. Wenn die Zeit abgelaufen ist und Sie keine Auswahl getroffen haben, dann erhalten Sie keine Auszahlung für diese Aufgabe.

Warten Sie bitte, bis der 5 Sekunden lange Countdown auf dem übernächsten Bildschirm abgelaufen ist. Danach werden Sie aufgefordert, zu würfeln.

Wenn Sie Fragen haben, heben Sie bitte Ihre Hand. Wenn nicht, drücken Sie auf VERSTANDEN.

## Sind Sie bereit?

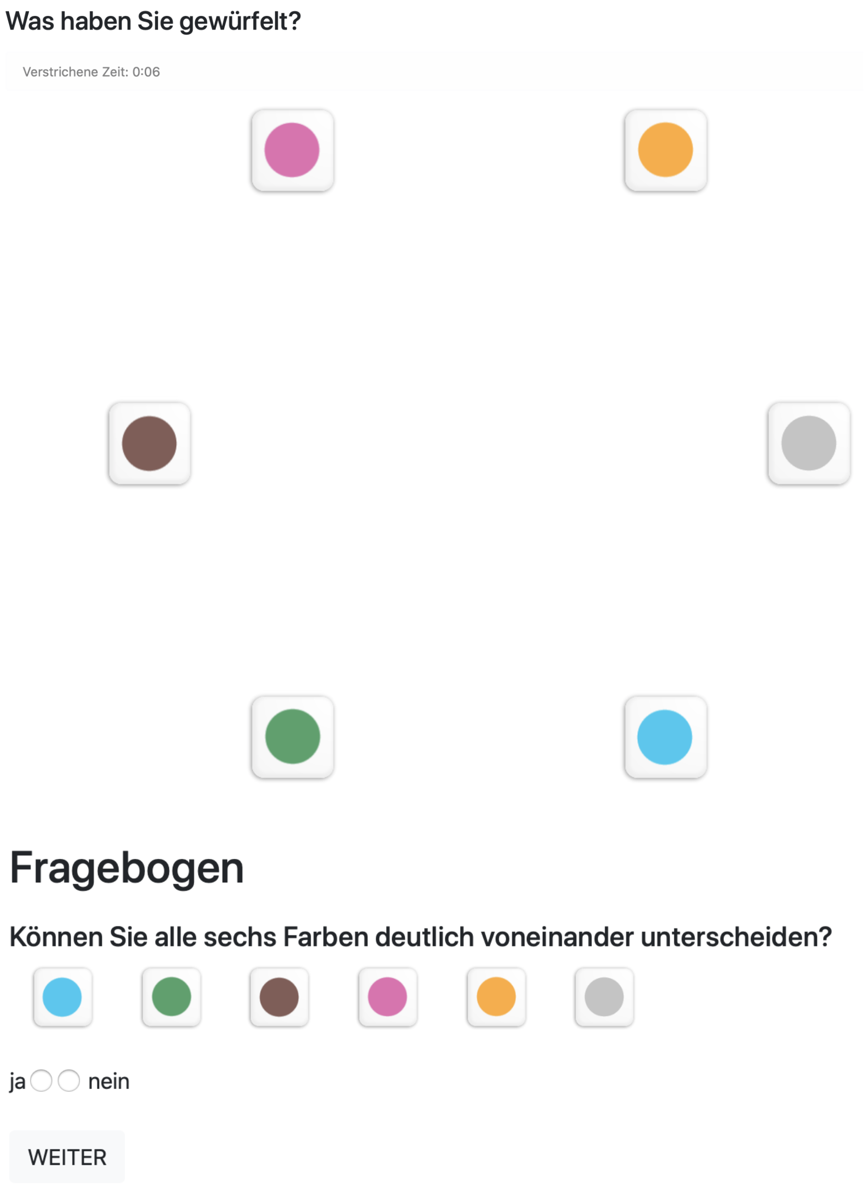
Als kurze Erinnerung: Bewegen Sie Ihren Kursor auf das Bild, um den Payoff zu sehen. Mit einem Klick auf das jeweilige Bild berichten Sie das Ergebnis.



WÜRFELN SIE JETZT!

Was haben Sie gewürfelt?

Verstrichene Zeit: 0:01



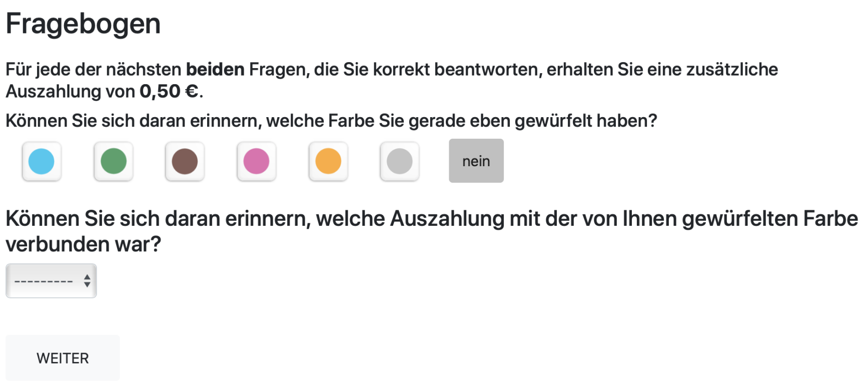
Was haben Sie gewürfelt?

Verstrichene Zeit: 0:06

## Fragebogen

Können Sie alle sechs Farben deutlich voneinander unterscheiden?

ja ○ ○ nein

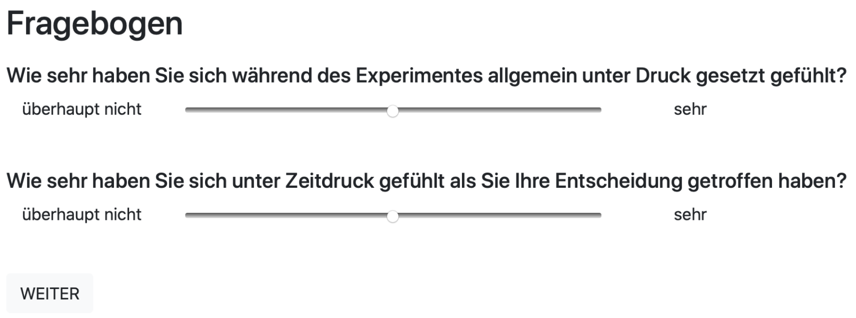


## Fragebogen

Für jede der nächsten **beiden** Fragen, die Sie korrekt beantworten, erhalten Sie eine zusätzliche Auszahlung von **0,50 €**.

Können Sie sich daran erinnern, welche Farbe Sie gerade eben gewürfelt haben?

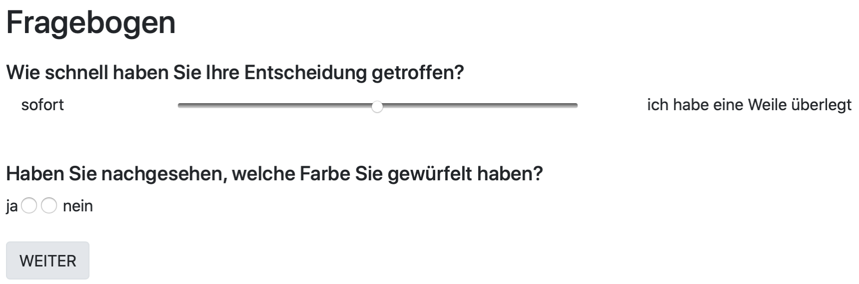
### Können Sie sich daran erinnern, welche Auszahlung mit der von Ihnen gewürfelten Farbe verbunden war?



## Fragebogen

### Wie sehr haben Sie sich während des Experimentes allgemein unter Druck gesetzt gefühlt?

### Wie sehr haben Sie sich unter Zeitdruck gefühlt als Sie Ihre Entscheidung getroffen haben?

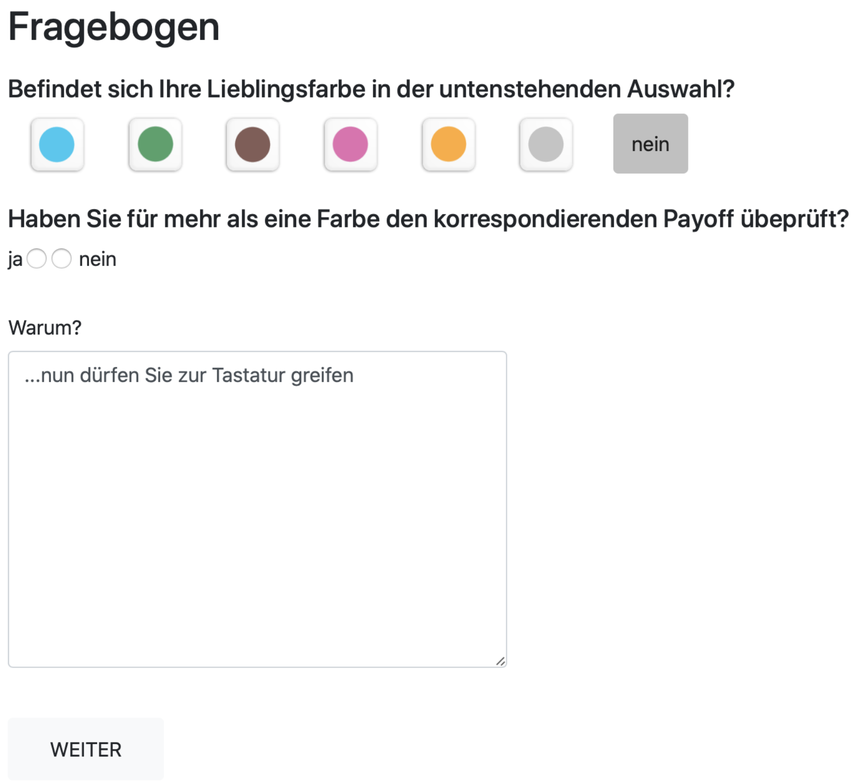


## Fragebogen

### Wie schnell haben Sie Ihre Entscheidung getroffen?

### Haben Sie nachgesehen, welche Farbe Sie gewürfelt haben?

ja ○ ○ nein



# Fragebogen

## Befindet sich Ihre Lieblingsfarbe in der untenstehenden Auswahl?

● ● ● ● ● ● nein

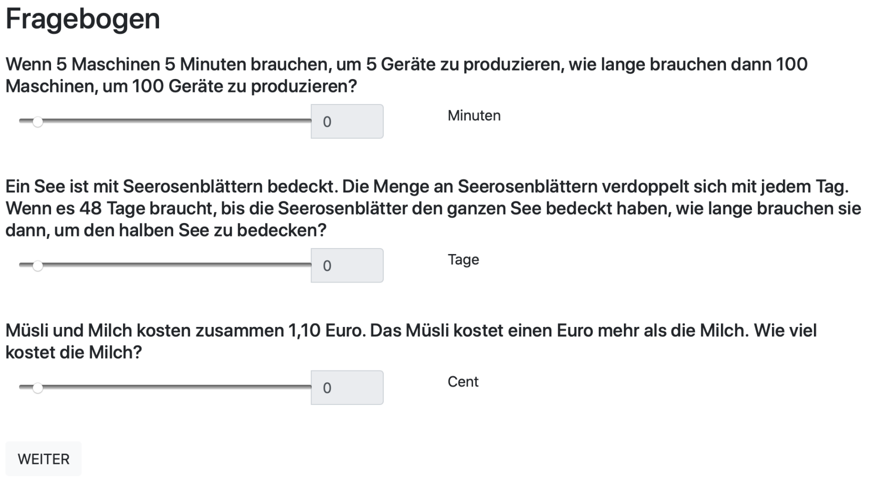
## Haben Sie für mehr als eine Farbe den korrespondierenden Payoff übeprüft?

ja ○ ○ nein

Warum?

...nun dürfen Sie zur Tastatur greifen

WEITER



## Fragebogen

Wenn 5 Maschinen 5 Minuten brauchen, um 5 Geräte zu produzieren, wie lange brauchen dann 100 Maschinen, um 100 Geräte zu produzieren?

0 Minuten

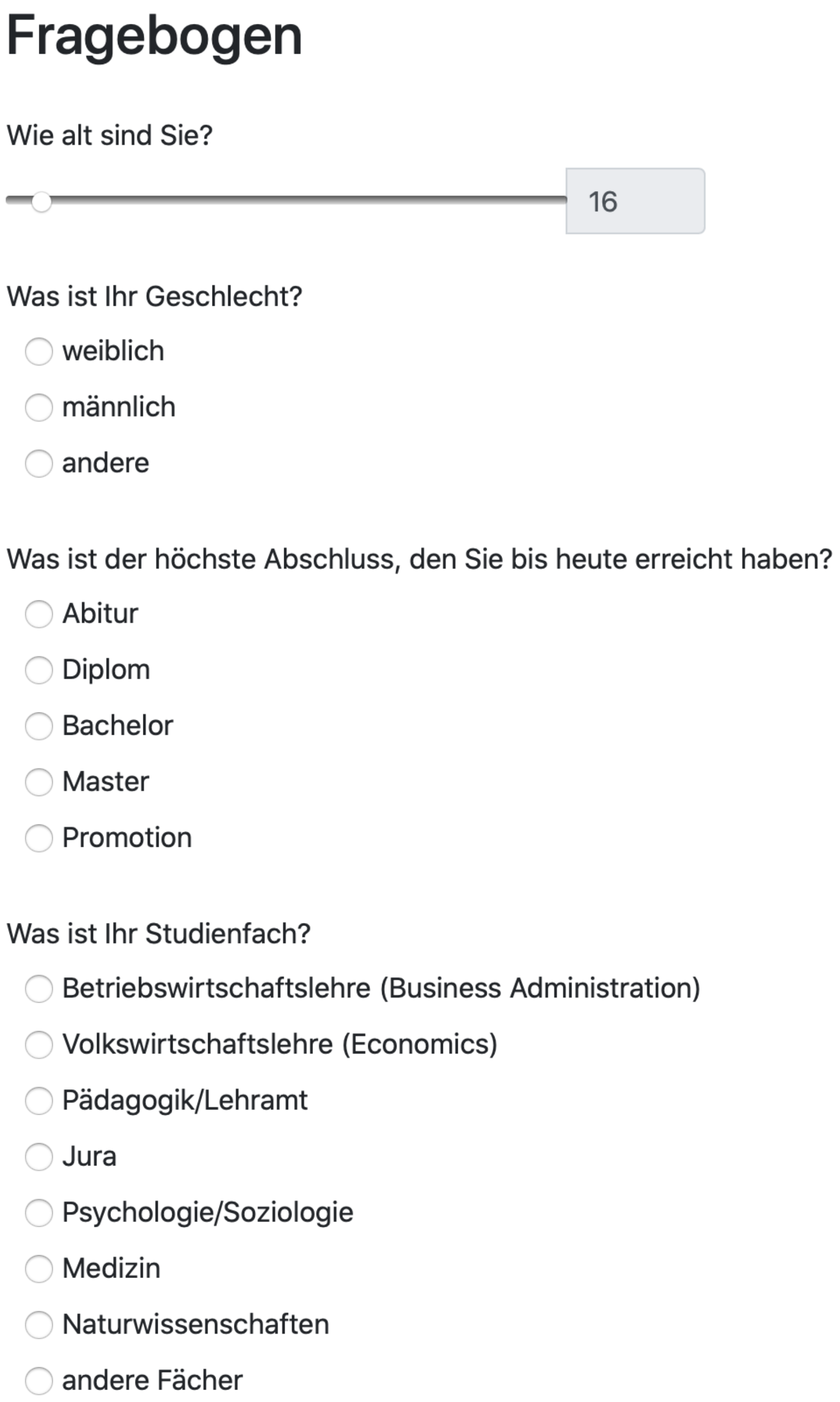
Ein See ist mit Seerosenblättern bedeckt. Die Menge an Seerosenblättern verdoppelt sich mit jedem Tag. Wenn es 48 Tage braucht, bis die Seerosenblätter den ganzen See bedeckt haben, wie lange brauchen sie dann, um den halben See zu bedecken?

0 Tage

Müsli und Milch kosten zusammen 1,10 Euro. Das Müsli kostet einen Euro mehr als die Milch. Wie viel kostet die Milch?

0 Cent

WEITER



# Fragebogen

Wie alt sind Sie?

16

Was ist Ihr Geschlecht?

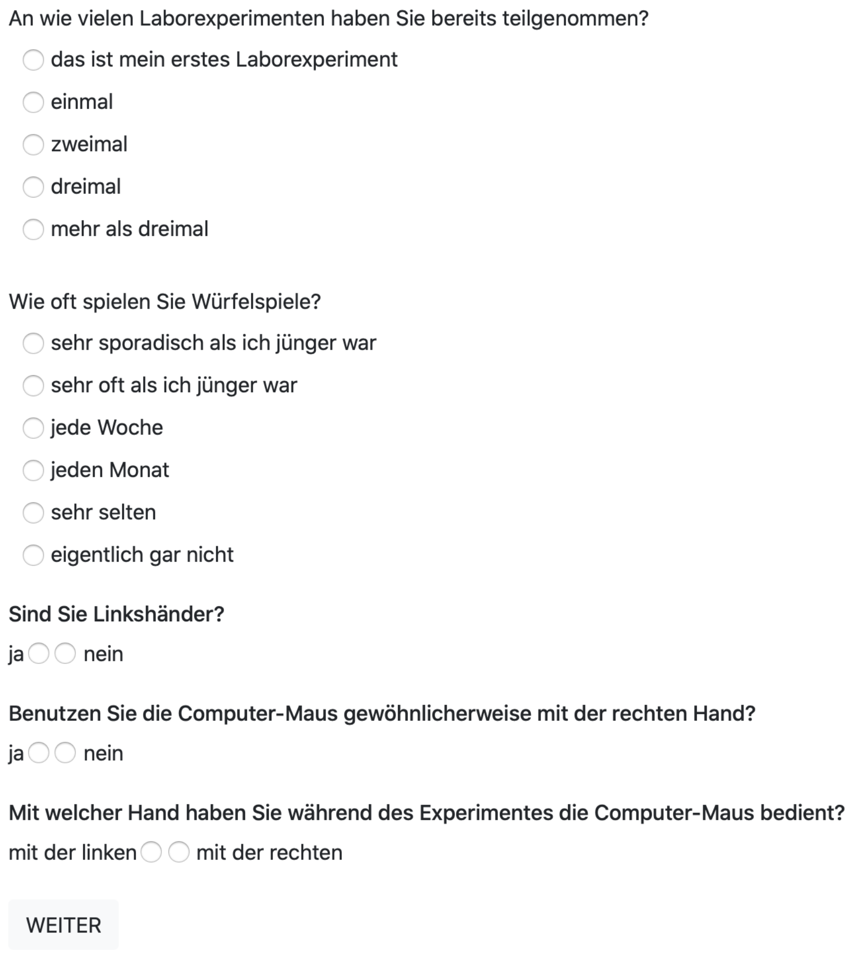
○ weiblich
○ männlich
○ andere

Was ist der höchste Abschluss, den Sie bis heute erreicht haben?

○ Abitur
○ Diplom
○ Bachelor
○ Master
○ Promotion

Was ist Ihr Studienfach?

○ Betriebswirtschaftslehre (Business Administration)
○ Volkswirtschaftslehre (Economics)
○ Pädagogik/Lehramt
○ Jura
○ Psychologie/Soziologie
○ Medizin
○ Naturwissenschaften
○ andere Fächer



An wie vielen Laborexperimenten haben Sie bereits teilgenommen?

○ das ist mein erstes Laborexperiment

○ einmal

○ zweimal

○ dreimal

○ mehr als dreimal

Wie oft spielen Sie Würfelspiele?

○ sehr sporadisch als ich jünger war

○ sehr oft als ich jünger war

○ jede Woche

○ jeden Monat

○ sehr selten

○ eigentlich gar nicht

**Sind Sie Linkshänder?**

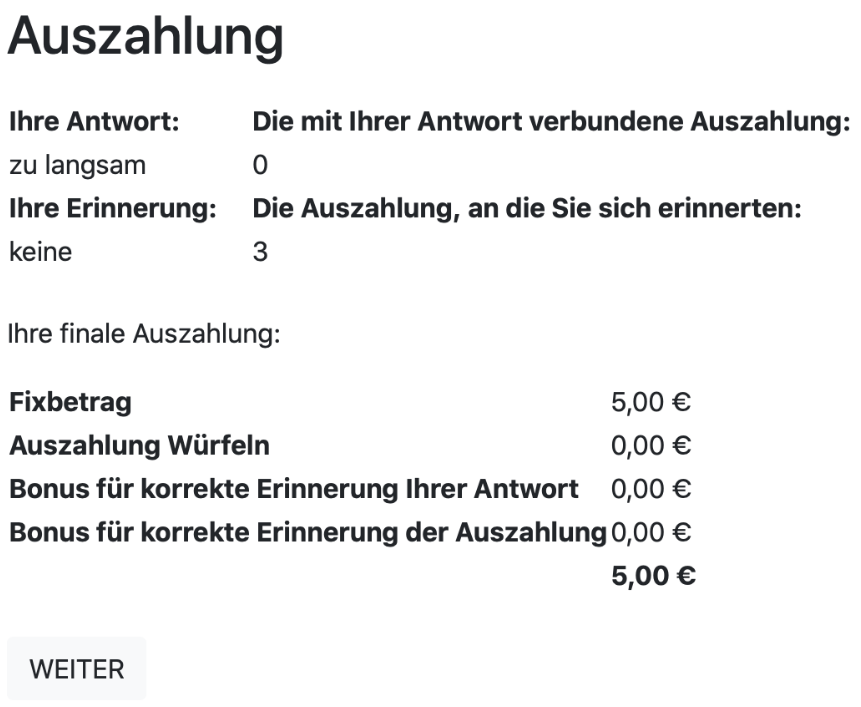
ja ○ ○ nein

**Benutzen Sie die Computer-Maus gewöhnlicherweise mit der rechten Hand?**

ja ○ ○ nein

**Mit welcher Hand haben Sie während des Experimentes die Computer-Maus bedient?**

mit der linken ○ ○ mit der rechten

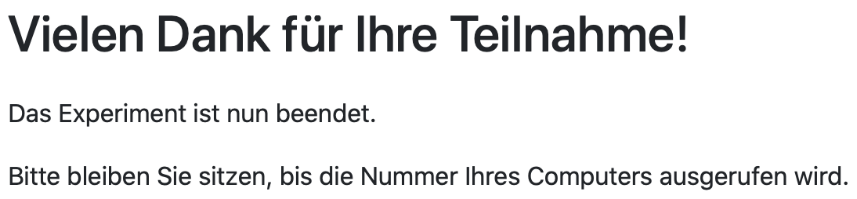


# Auszahlung

| **Ihre Antwort:** | **Die mit Ihrer Antwort verbundene Auszahlung:** |
|---|---|
| zu langsam | 0 |
| **Ihre Erinnerung:** | **Die Auszahlung, an die Sie sich erinnerten:** |
| keine | 3 |

Ihre finale Auszahlung:

| | |
|---|---|
| **Fixbetrag** | 5,00 € |
| **Auszahlung Würfeln** | 0,00 € |
| **Bonus für korrekte Erinnerung Ihrer Antwort** | 0,00 € |
| **Bonus für korrekte Erinnerung der Auszahlung** | 0,00 € |
| | **5,00 €** |



# Vielen Dank für Ihre Teilnahme!

Das Experiment ist nun beendet.

Bitte bleiben Sie sitzen, bis die Nummer Ihres Computers ausgerufen wird.

# B
# Appendix, Chapter 2

## B.1 List of Labels

The following list denotes the labels that were assigned manually by Fochmann et al. (2019) after reading the chat texts.

1. money, general
2. money, pro honesty
3. money, pro lying
4. tax, general
5. tax, pro
6. tax, against
7. risk, general
8. risk, pro honesty
9. risk, pro lying
10. honesty, general
11. honesty, pro honesty
12. honesty, pro lying
13. number, general
14. number, pro honesty
15. number, pro lying
16. keep strategy, pro honesty
17. keep strategy, pro lying
18. change strategy, pro honesty
19. change strategy, pro lying
20. insecurity, general
21. insecurity, honest
22. insecurity, lie
23. rules, general
24. rules, yes
25. rules, no
26. miscellaneous, general
27. miscellaneous, honest
28. miscellaneous, lie
29. consequences, general
30. consequences, positive

31. consequences, negative

32. reality vs. game

33. tax fairness

34. tax honesty

## B.2 Concepts to Potentially Influence Lying Behavior

After participants had stated their surplus hours, they reported sociodemographics (e.g., age, gender) and information on their risk attitudes. Furthermore, we extracted five items from the German Positive and Negative Affect Schedule (PANAS) (Janke and Glöckner-Rist 2014) with 10-item scales to measure affects concerning joy, anger, fear, guilt, and shame. Additionally, we collected individual data on lying morale and income. In the following, we analyze the distribution of answers concerning the questions asked.

Overall, we collected 351 observations; participants were 24.8 years old, and 60% of them were female. Experimental evidence shows that economics students lie more than students in other fields (López-Pérez and Spiegelman 2019). In the sample collected, 11% of the participants study economics, and 34% stated having taken more than one class in economics.

After reporting sociodemographic characteristics, participants stated the intensity of feelings and emotions experienced throughout the experiment. Figure B.1 shows that there is not much variation concerning anger, fear, shame, and guilt. There is, however, substantial variation in the experience of joy.

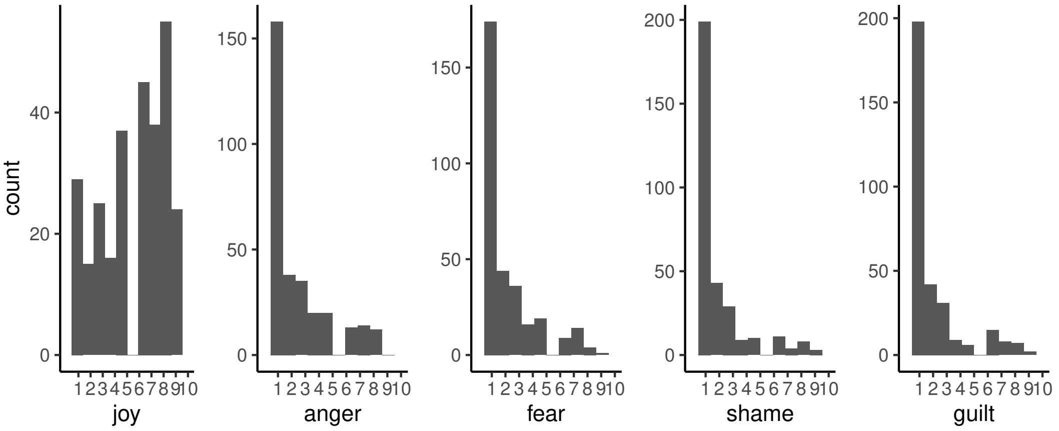


**Figure B.1:** Emotions Experienced During the Experiment

Apart from positive and negative affects, we asked about concepts potentially related to lying behavior such as income, religiosity, risk attitudes, lying attitudes, political orientation, and experience with laboratory experiments. Figure B.2 shows that there is not much variation in income stated and the number of times a participant prays. The experienced complexity of the experiment is also heavily skewed to the left. There is substantial variation in risk attitudes (1 = not at all prone to take risks; 9 = very prone to take risks), attitude towards lying (1 = one should never lie for one's advantage; 10 = one should lie for one's advantage), political orientation (1 = left-wing; 9 = right-wing), and the experience with laboratory experiments (1 = never; 5 = more than 20 times). Furthermore, there is a substantial variation in a participant's confidence in the other participants' honesty. The question asked: "Out of 100 participants, how many do you think state more than the true amount of surplus hours?"

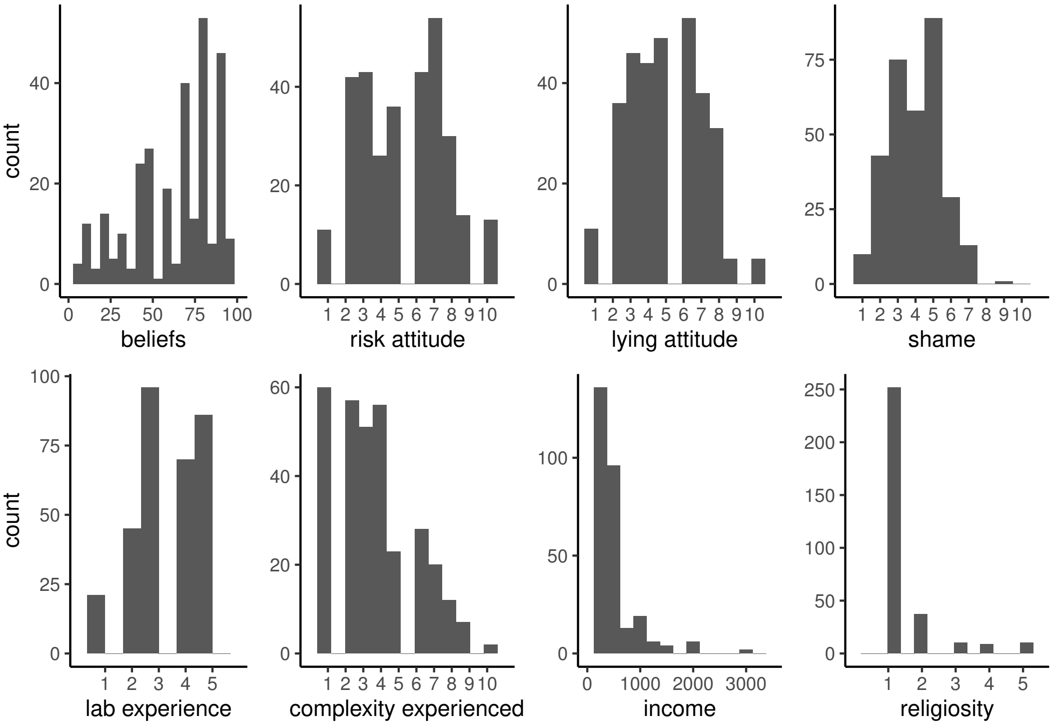
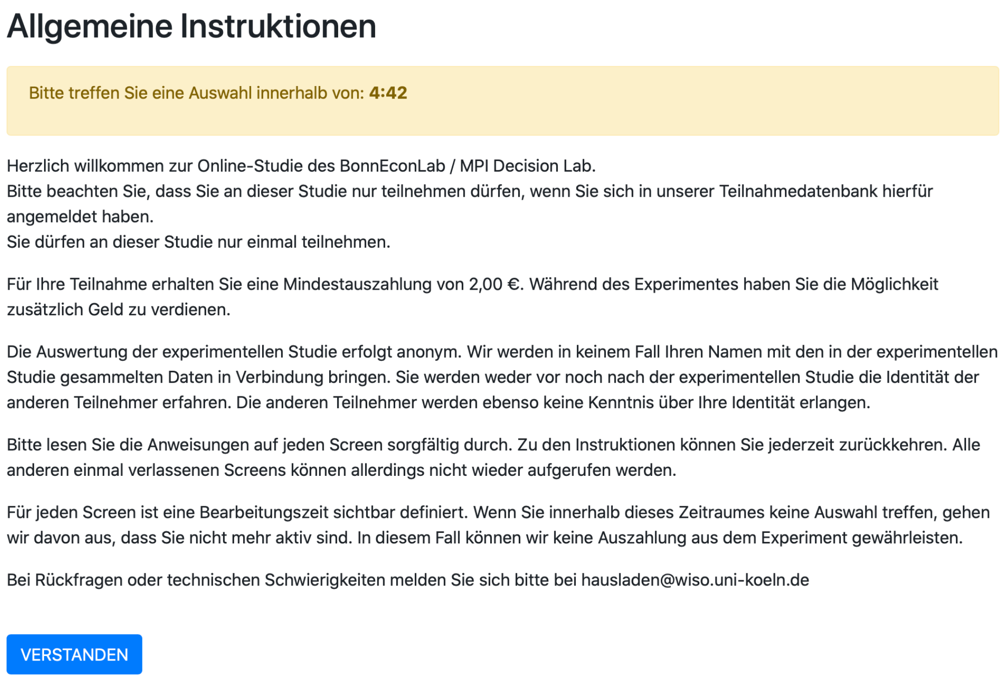


**Figure B.2:** Concepts Potentially Related to Lying Behavior

## B.3 Screenshots of the Experiment



Allgemeine Instruktionen

Bitte treffen Sie eine Auswahl innerhalb von: **4:42**

Herzlich willkommen zur Online-Studie des BonnEconLab / MPI Decision Lab.
Bitte beachten Sie, dass Sie an dieser Studie nur teilnehmen dürfen, wenn Sie sich in unserer Teilnahmedatenbank hierfür angemeldet haben.
Sie dürfen an dieser Studie nur einmal teilnehmen.

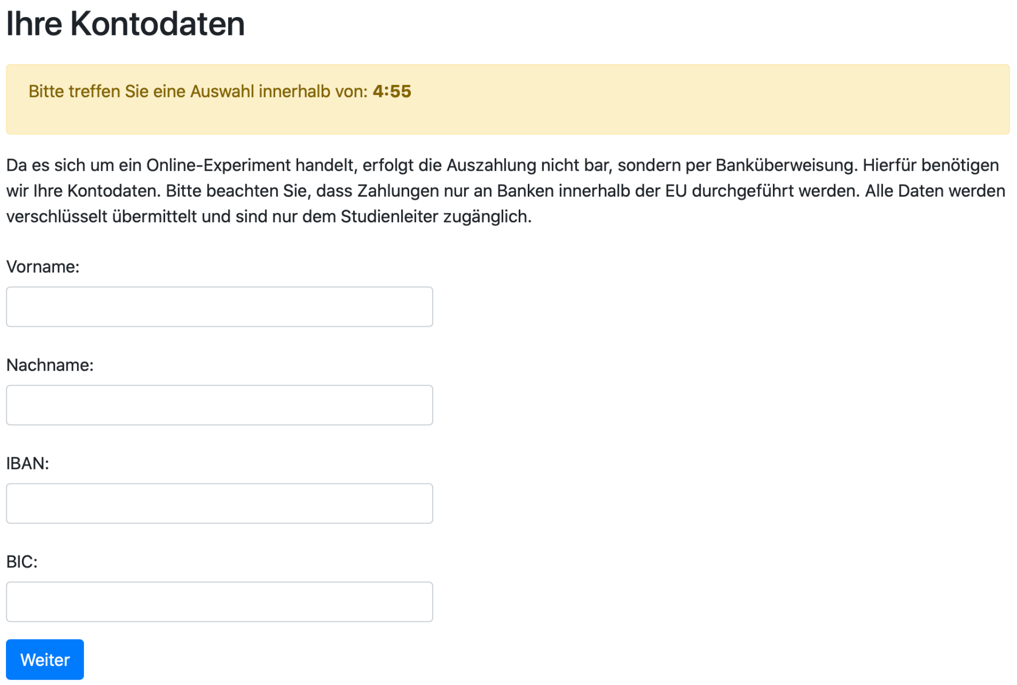
Für Ihre Teilnahme erhalten Sie eine Mindestauszahlung von 2,00 €. Während des Experimentes haben Sie die Möglichkeit zusätzlich Geld zu verdienen.

Die Auswertung der experimentellen Studie erfolgt anonym. Wir werden in keinem Fall Ihren Namen mit den in der experimentellen Studie gesammelten Daten in Verbindung bringen. Sie werden weder vor noch nach der experimentellen Studie die Identität der anderen Teilnehmer erfahren. Die anderen Teilnehmer werden ebenso keine Kenntnis über Ihre Identität erlangen.

Bitte lesen Sie die Anweisungen auf jeden Screen sorgfältig durch. Zu den Instruktionen können Sie jederzeit zurückkehren. Alle anderen einmal verlassenen Screens können allerdings nicht wieder aufgerufen werden.

Für jeden Screen ist eine Bearbeitungszeit sichtbar definiert. Wenn Sie innerhalb dieses Zeitraumes keine Auswahl treffen, gehen wir davon aus, dass Sie nicht mehr aktiv sind. In diesem Fall können wir keine Auszahlung aus dem Experiment gewährleisten.

Bei Rückfragen oder technischen Schwierigkeiten melden Sie sich bitte bei hausladen@wiso.uni-koeln.de



Ihre Kontodaten

Bitte treffen Sie eine Auswahl innerhalb von: **4:55**

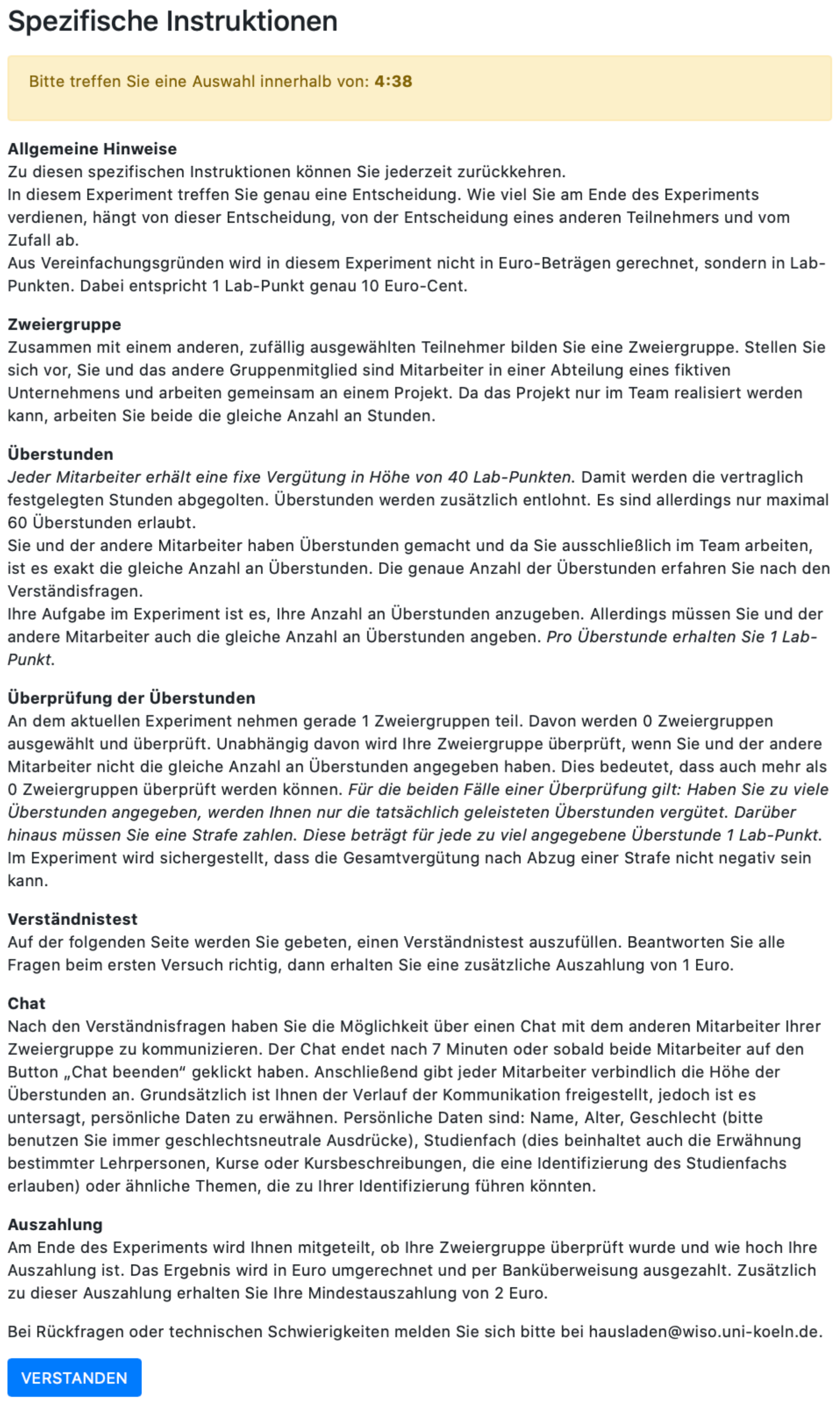
Da es sich um ein Online-Experiment handelt, erfolgt die Auszahlung nicht bar, sondern per Banküberweisung. Hierfür benötigen wir Ihre Kontodaten. Bitte beachten Sie, dass Zahlungen nur an Banken innerhalb der EU durchgeführt werden. Alle Daten werden verschlüsselt übermittelt und sind nur dem Studienleiter zugänglich.

Vorname:

Nachname:

IBAN:

BIC:



## Spezifische Instruktionen

Bitte treffen Sie eine Auswahl innerhalb von: **4:38**

**Allgemeine Hinweise**
Zu diesen spezifischen Instruktionen können Sie jederzeit zurückkehren.
In diesem Experiment treffen Sie genau eine Entscheidung. Wie viel Sie am Ende des Experiments verdienen, hängt von dieser Entscheidung, von der Entscheidung eines anderen Teilnehmers und vom Zufall ab.
Aus Vereinfachungsgründen wird in diesem Experiment nicht in Euro-Beträgen gerechnet, sondern in Lab-Punkten. Dabei entspricht 1 Lab-Punkt genau 10 Euro-Cent.

**Zweiergruppe**
Zusammen mit einem anderen, zufällig ausgewählten Teilnehmer bilden Sie eine Zweiergruppe. Stellen Sie sich vor, Sie und das andere Gruppenmitglied sind Mitarbeiter in einer Abteilung eines fiktiven Unternehmens und arbeiten gemeinsam an einem Projekt. Da das Projekt nur im Team realisiert werden kann, arbeiten Sie beide die gleiche Anzahl an Stunden.

**Überstunden**
*Jeder Mitarbeiter erhält eine fixe Vergütung in Höhe von 40 Lab-Punkten.* Damit werden die vertraglich festgelegten Stunden abgegolten. Überstunden werden zusätzlich entlohnt. Es sind allerdings nur maximal 60 Überstunden erlaubt.
Sie und der andere Mitarbeiter haben Überstunden gemacht und da Sie ausschließlich im Team arbeiten, ist es exakt die gleiche Anzahl an Überstunden. Die genaue Anzahl der Überstunden erfahren Sie nach den Verständisfragen.
Ihre Aufgabe im Experiment ist es, Ihre Anzahl an Überstunden anzugeben. Allerdings müssen Sie und der andere Mitarbeiter auch die gleiche Anzahl an Überstunden angeben. *Pro Überstunde erhalten Sie 1 Lab-Punkt.*

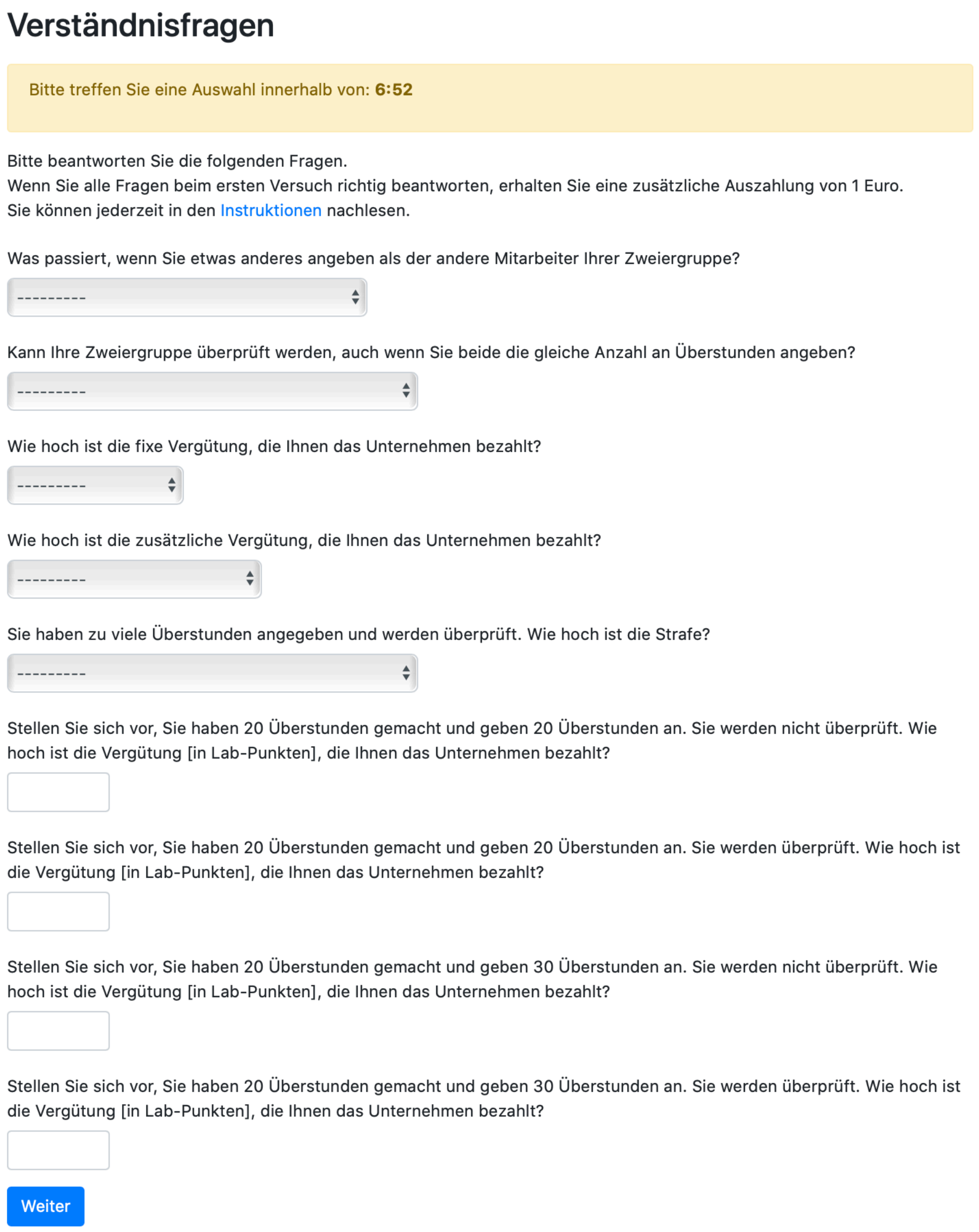
**Überprüfung der Überstunden**
An dem aktuellen Experiment nehmen gerade 1 Zweiergruppen teil. Davon werden 0 Zweiergruppen ausgewählt und überprüft. Unabhängig davon wird Ihre Zweiergruppe überprüft, wenn Sie und der andere Mitarbeiter nicht die gleiche Anzahl an Überstunden angegeben haben. Dies bedeutet, dass auch mehr als 0 Zweiergruppen überprüft werden können. *Für die beiden Fälle einer Überprüfung gilt: Haben Sie zu viele Überstunden angegeben, werden Ihnen nur die tatsächlich geleisteten Überstunden vergütet. Darüber hinaus müssen Sie eine Strafe zahlen. Diese beträgt für jede zu viel angegebene Überstunde 1 Lab-Punkt.* Im Experiment wird sichergestellt, dass die Gesamtvergütung nach Abzug einer Strafe nicht negativ sein kann.

**Verständnistest**
Auf der folgenden Seite werden Sie gebeten, einen Verständnistest auszufüllen. Beantworten Sie alle Fragen beim ersten Versuch richtig, dann erhalten Sie eine zusätzliche Auszahlung von 1 Euro.

**Chat**
Nach den Verständnisfragen haben Sie die Möglichkeit über einen Chat mit dem anderen Mitarbeiter Ihrer Zweiergruppe zu kommunizieren. Der Chat endet nach 7 Minuten oder sobald beide Mitarbeiter auf den Button „Chat beenden" geklickt haben. Anschließend gibt jeder Mitarbeiter verbindlich die Höhe der Überstunden an. Grundsätzlich ist Ihnen der Verlauf der Kommunikation freigestellt, jedoch ist es untersagt, persönliche Daten zu erwähnen. Persönliche Daten sind: Name, Alter, Geschlecht (bitte benutzen Sie immer geschlechtsneutrale Ausdrücke), Studienfach (dies beinhaltet auch die Erwähnung bestimmter Lehrpersonen, Kurse oder Kursbeschreibungen, die eine Identifizierung des Studienfachs erlauben) oder ähnliche Themen, die zu Ihrer Identifizierung führen könnten.

**Auszahlung**
Am Ende des Experiments wird Ihnen mitgeteilt, ob Ihre Zweiergruppe überprüft wurde und wie hoch Ihre Auszahlung ist. Das Ergebnis wird in Euro umgerechnet und per Banküberweisung ausgezahlt. Zusätzlich zu dieser Auszahlung erhalten Sie Ihre Mindestauszahlung von 2 Euro.

Bei Rückfragen oder technischen Schwierigkeiten melden Sie sich bitte bei hausladen@wiso.uni-koeln.de.



## Verständnisfragen

Bitte treffen Sie eine Auswahl innerhalb von: **6:52**

Bitte beantworten Sie die folgenden Fragen.
Wenn Sie alle Fragen beim ersten Versuch richtig beantworten, erhalten Sie eine zusätzliche Auszahlung von 1 Euro.
Sie können jederzeit in den Instruktionen nachlesen.

Was passiert, wenn Sie etwas anderes angeben als der andere Mitarbeiter Ihrer Zweiergruppe?

---------

Kann Ihre Zweiergruppe überprüft werden, auch wenn Sie beide die gleiche Anzahl an Überstunden angeben?

---------

Wie hoch ist die fixe Vergütung, die Ihnen das Unternehmen bezahlt?

---------

Wie hoch ist die zusätzliche Vergütung, die Ihnen das Unternehmen bezahlt?

---------

Sie haben zu viele Überstunden angegeben und werden überprüft. Wie hoch ist die Strafe?

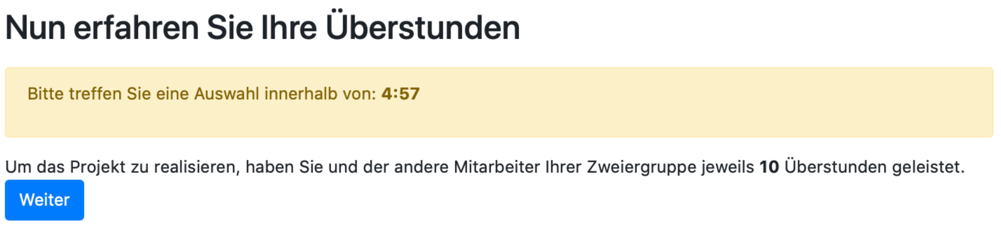
---------

Stellen Sie sich vor, Sie haben 20 Überstunden gemacht und geben 20 Überstunden an. Sie werden nicht überprüft. Wie hoch ist die Vergütung [in Lab-Punkten], die Ihnen das Unternehmen bezahlt?

Stellen Sie sich vor, Sie haben 20 Überstunden gemacht und geben 20 Überstunden an. Sie werden überprüft. Wie hoch ist die Vergütung [in Lab-Punkten], die Ihnen das Unternehmen bezahlt?

Stellen Sie sich vor, Sie haben 20 Überstunden gemacht und geben 30 Überstunden an. Sie werden nicht überprüft. Wie hoch ist die Vergütung [in Lab-Punkten], die Ihnen das Unternehmen bezahlt?

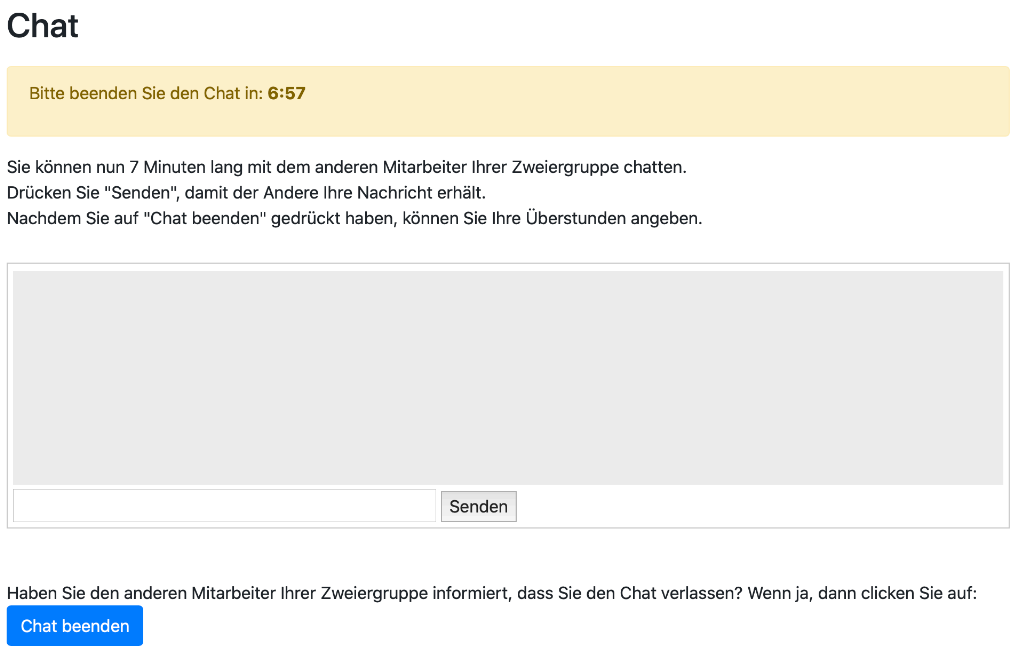
Stellen Sie sich vor, Sie haben 20 Überstunden gemacht und geben 30 Überstunden an. Sie werden überprüft. Wie hoch ist die Vergütung [in Lab-Punkten], die Ihnen das Unternehmen bezahlt?



## Nun erfahren Sie Ihre Überstunden

Bitte treffen Sie eine Auswahl innerhalb von: **4:57**

Um das Projekt zu realisieren, haben Sie und der andere Mitarbeiter Ihrer Zweiergruppe jeweils **10** Überstunden geleistet.

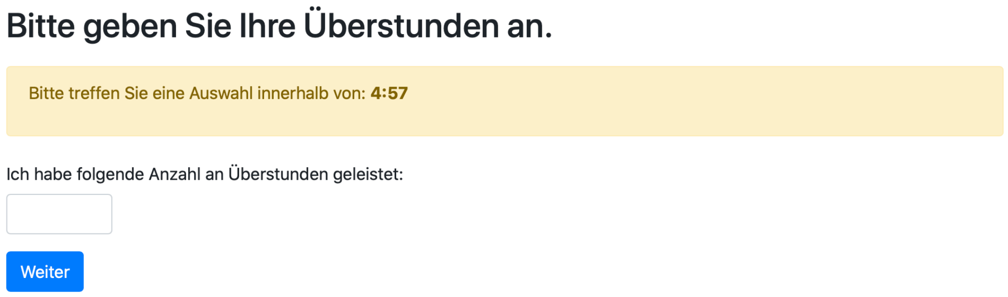


## Chat

Bitte beenden Sie den Chat in: **6:57**

Sie können nun 7 Minuten lang mit dem anderen Mitarbeiter Ihrer Zweiergruppe chatten.
Drücken Sie "Senden", damit der Andere Ihre Nachricht erhält.
Nachdem Sie auf "Chat beenden" gedrückt haben, können Sie Ihre Überstunden angeben.

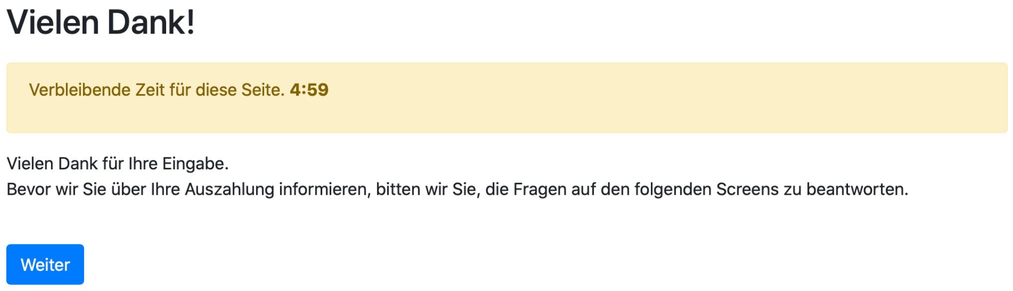
Haben Sie den anderen Mitarbeiter Ihrer Zweiergruppe informiert, dass Sie den Chat verlassen? Wenn ja, dann clicken Sie auf:



## Bitte geben Sie Ihre Überstunden an.

Bitte treffen Sie eine Auswahl innerhalb von: **4:57**

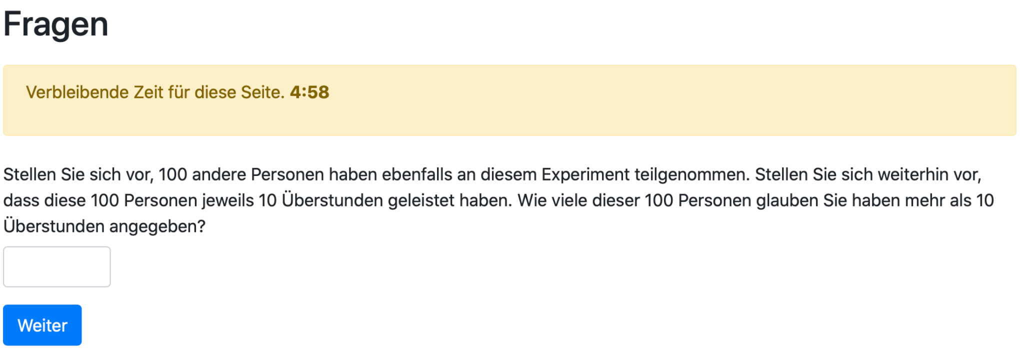
Ich habe folgende Anzahl an Überstunden geleistet:



## Vielen Dank!

Verbleibende Zeit für diese Seite. **4:59**

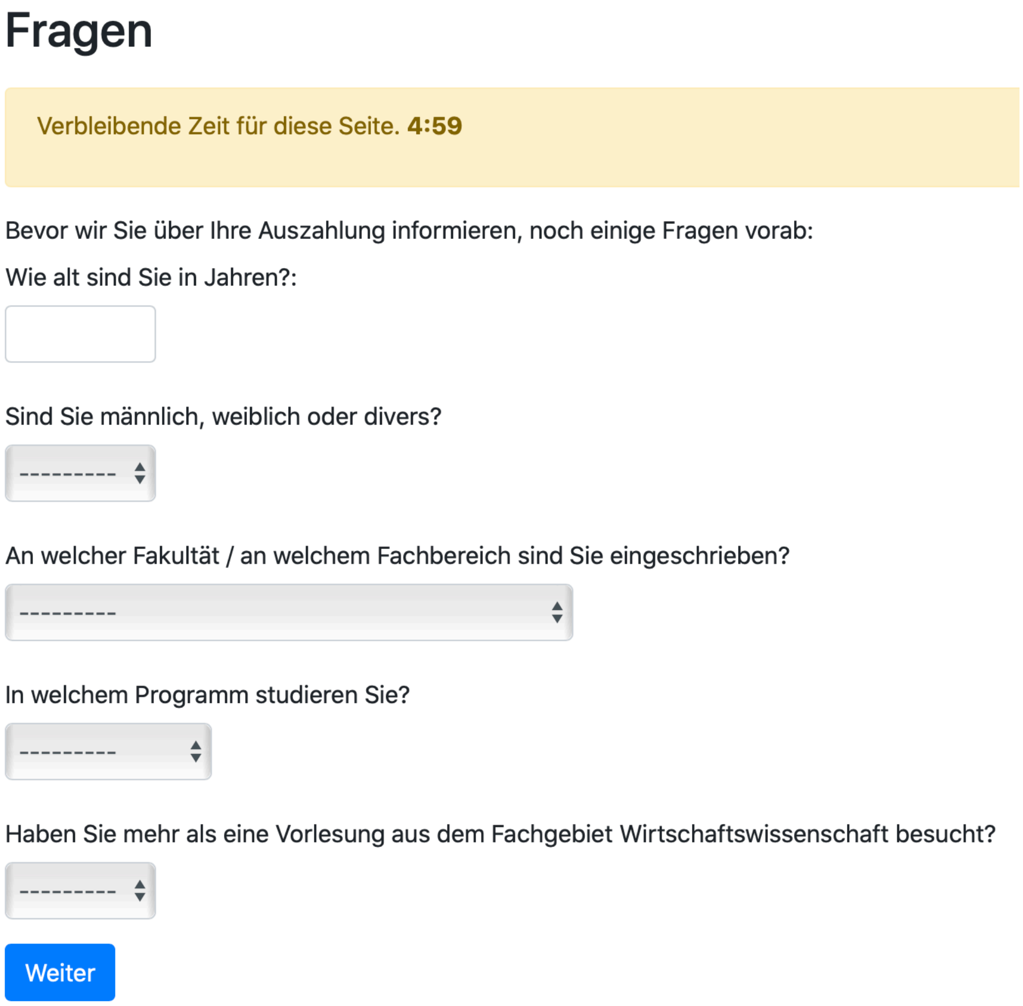
Vielen Dank für Ihre Eingabe.
Bevor wir Sie über Ihre Auszahlung informieren, bitten wir Sie, die Fragen auf den folgenden Screens zu beantworten.



Fragen

Verbleibende Zeit für diese Seite. **4:58**

Stellen Sie sich vor, 100 andere Personen haben ebenfalls an diesem Experiment teilgenommen. Stellen Sie sich weiterhin vor, dass diese 100 Personen jeweils 10 Überstunden geleistet haben. Wie viele dieser 100 Personen glauben Sie haben mehr als 10 Überstunden angegeben?



Fragen

Verbleibende Zeit für diese Seite. **4:59**

Bevor wir Sie über Ihre Auszahlung informieren, noch einige Fragen vorab:

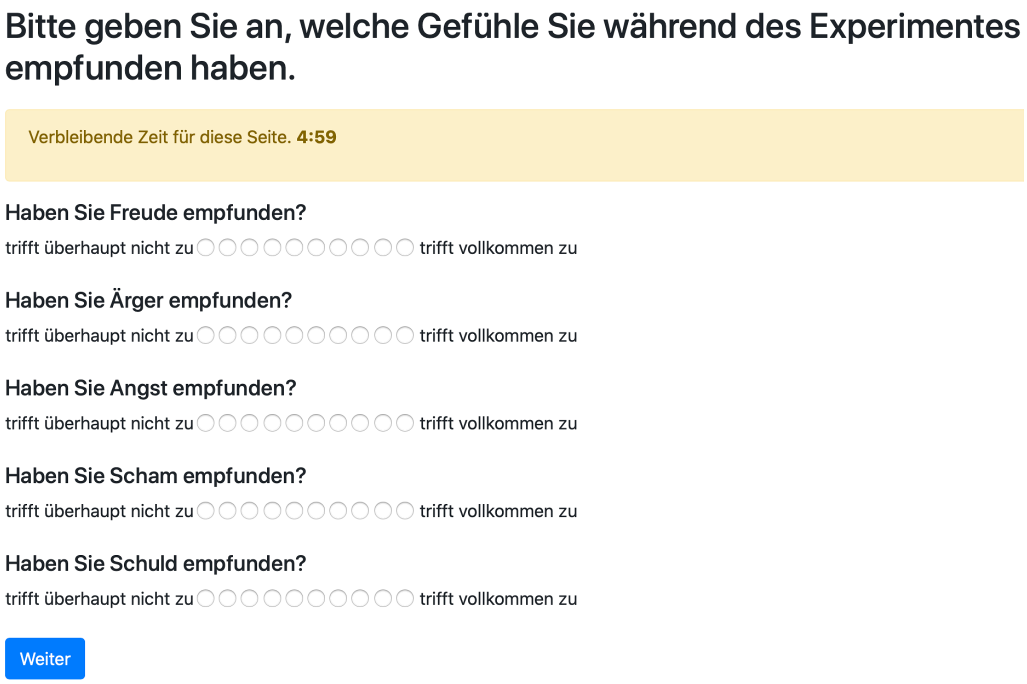
Wie alt sind Sie in Jahren?:

Sind Sie männlich, weiblich oder divers?

An welcher Fakultät / an welchem Fachbereich sind Sie eingeschrieben?

In welchem Programm studieren Sie?

Haben Sie mehr als eine Vorlesung aus dem Fachgebiet Wirtschaftswissenschaft besucht?



## Bitte geben Sie an, welche Gefühle Sie während des Experimentes empfunden haben.

Verbleibende Zeit für diese Seite. **4:59**

**Haben Sie Freude empfunden?**

trifft überhaupt nicht zu ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ trifft vollkommen zu

**Haben Sie Ärger empfunden?**

trifft überhaupt nicht zu ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ trifft vollkommen zu

**Haben Sie Angst empfunden?**

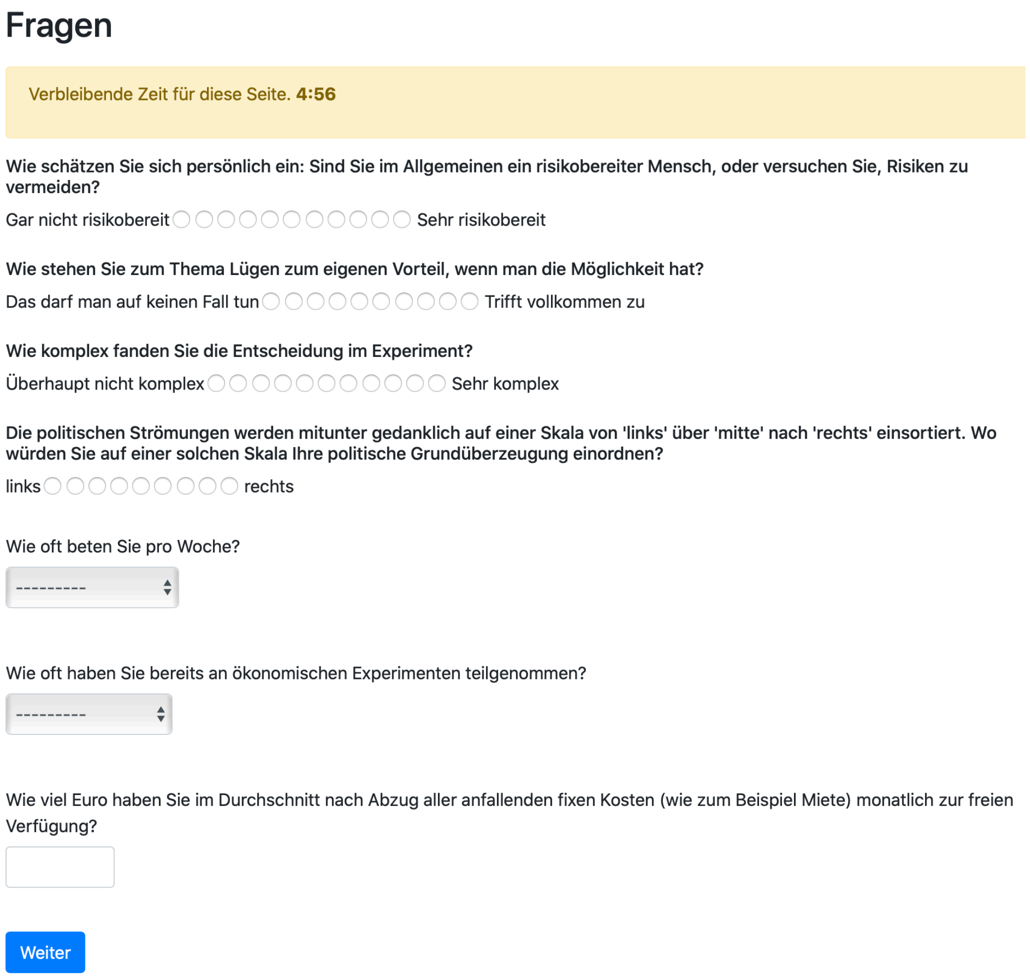
trifft überhaupt nicht zu ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ trifft vollkommen zu

**Haben Sie Scham empfunden?**

trifft überhaupt nicht zu ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ trifft vollkommen zu

**Haben Sie Schuld empfunden?**

trifft überhaupt nicht zu ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ trifft vollkommen zu



# Fragen

Verbleibende Zeit für diese Seite. **4:56**

**Wie schätzen Sie sich persönlich ein: Sind Sie im Allgemeinen ein risikobereiter Mensch, oder versuchen Sie, Risiken zu vermeiden?**

Gar nicht risikobereit ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ Sehr risikobereit

**Wie stehen Sie zum Thema Lügen zum eigenen Vorteil, wenn man die Möglichkeit hat?**

Das darf man auf keinen Fall tun ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ Trifft vollkommen zu

**Wie komplex fanden Sie die Entscheidung im Experiment?**

Überhaupt nicht komplex ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ Sehr komplex

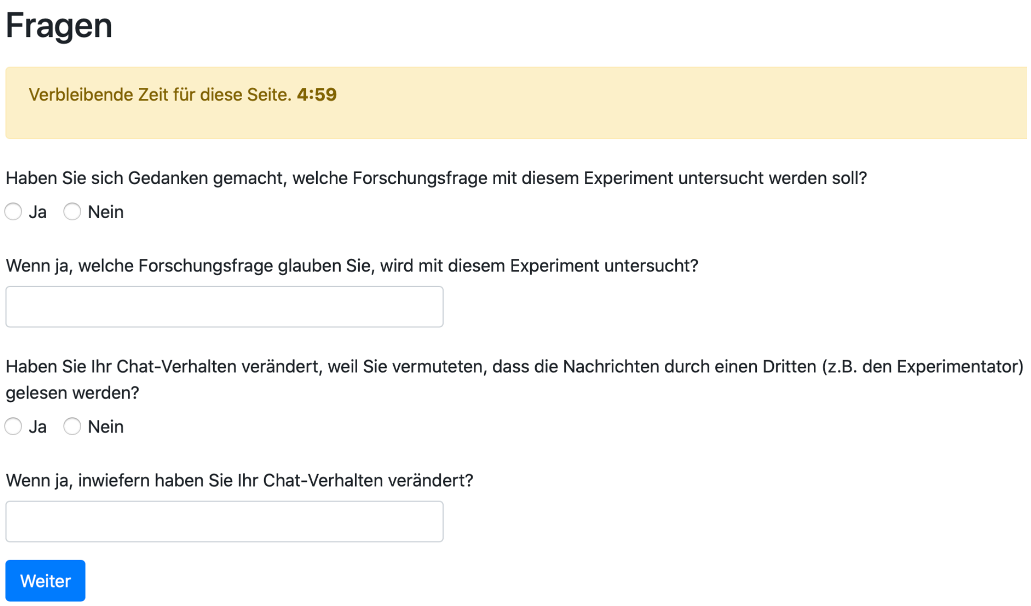
**Die politischen Strömungen werden mitunter gedanklich auf einer Skala von 'links' über 'mitte' nach 'rechts' einsortiert. Wo würden Sie auf einer solchen Skala Ihre politische Grundüberzeugung einordnen?**

links ○ ○ ○ ○ ○ ○ ○ ○ ○ rechts

Wie oft beten Sie pro Woche?

Wie oft haben Sie bereits an ökonomischen Experimenten teilgenommen?

Wie viel Euro haben Sie im Durchschnitt nach Abzug aller anfallenden fixen Kosten (wie zum Beispiel Miete) monatlich zur freien Verfügung?



## Fragen

Verbleibende Zeit für diese Seite. **4:59**

Haben Sie sich Gedanken gemacht, welche Forschungsfrage mit diesem Experiment untersucht werden soll?

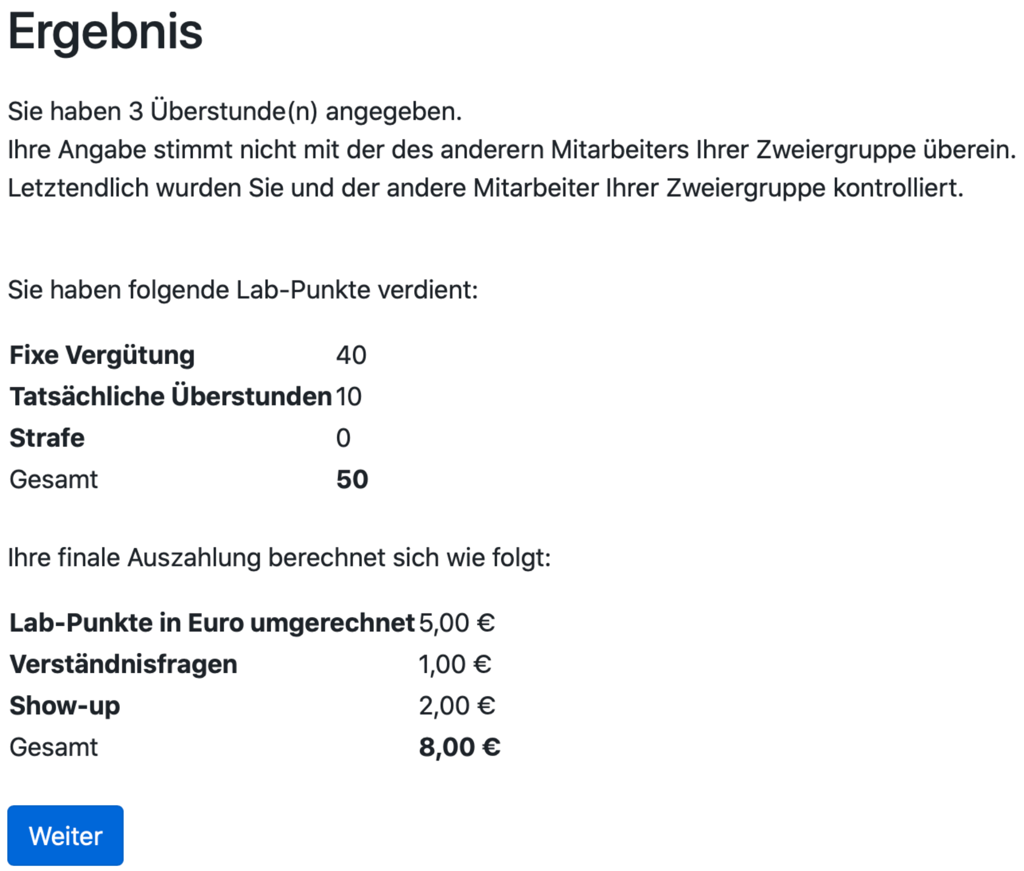
○ Ja ○ Nein

Wenn ja, welche Forschungsfrage glauben Sie, wird mit diesem Experiment untersucht?

Haben Sie Ihr Chat-Verhalten verändert, weil Sie vermuteten, dass die Nachrichten durch einen Dritten (z.B. den Experimentator) gelesen werden?

○ Ja ○ Nein

Wenn ja, inwiefern haben Sie Ihr Chat-Verhalten verändert?



# Ergebnis

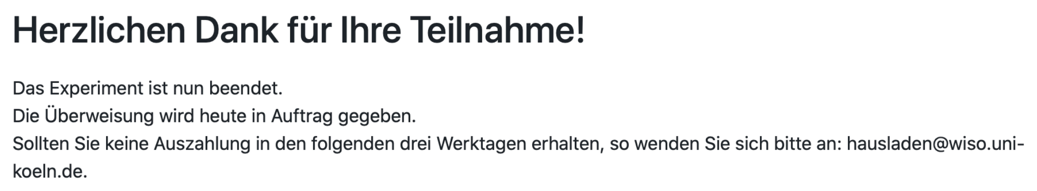
Sie haben 3 Überstunde(n) angegeben.
Ihre Angabe stimmt nicht mit der des anderern Mitarbeiters Ihrer Zweiergruppe überein.
Letztendlich wurden Sie und der andere Mitarbeiter Ihrer Zweiergruppe kontrolliert.

Sie haben folgende Lab-Punkte verdient:

| | |
|---|---|
| **Fixe Vergütung** | 40 |
| **Tatsächliche Überstunden** | 10 |
| **Strafe** | 0 |
| Gesamt | **50** |

Ihre finale Auszahlung berechnet sich wie folgt:

| | |
|---|---|
| **Lab-Punkte in Euro umgerechnet** | 5,00 € |
| **Verständnisfragen** | 1,00 € |
| **Show-up** | 2,00 € |
| Gesamt | **8,00 €** |



# Herzlichen Dank für Ihre Teilnahme!

Das Experiment ist nun beendet.
Die Überweisung wird heute in Auftrag gegeben.
Sollten Sie keine Auszahlung in den folgenden drei Werktagen erhalten, so wenden Sie sich bitte an: hausladen@wiso.uni-koeln.de.

# C
# Appendix, Chapter 3

## C.1 Internal Cluster Validation Indices for Simulated Data

In Figure C.1 all indices are normalized to the unit interval. Indices to be minimized are recoded and reported as inverse.

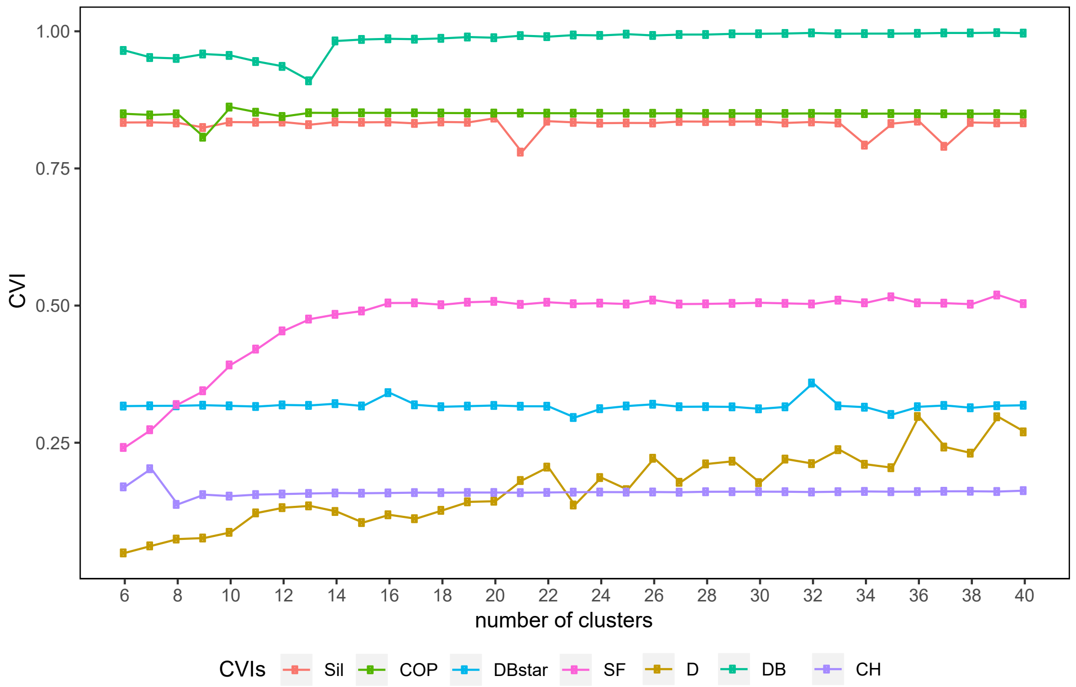


**Figure C.1:** Simulated Data: Internal Cluster Validation Indices

# D Appendix, Chapter 4

## D.1 Replication

**Table D.1:** Overview of all Tables and Figures in Landes and Posner (2011) concerning Circuit Courts

| | |
|---|---|
| Figure 3 | Total votes by year appointed to the court of appeals |
| Table 11 | Court of appeals votes by subject matter and ideology for 538 court of appeals judges only: 1925–2002 |
| Table 12 | Fraction of mixed, conservative and liberal votes for 538 court of appeals judges by president at time of appointment: 1925–2002 |
| Table 13 | Regression analysis of court of appeals votes: 1925–2002 (t-statistics in parentheses) |
| Table 14 | Regression analysis of court of appeals votes: 1960–2002 (t-statistics in parentheses) |
| Table 15 | Circuit effects on ideology of judges' votes |
| Table 16 | Regression analysis of appellate court votes: current judges (t-statistics in parentheses) |

## D.2 Data Preprocessing

We applied preprocessing tailored to our data. As we use data from Lexis, each opinion had a specific structure. We extracted the text and split it into parts when encountering more than a single newline character. Special characters such as "newline"-characters and Roman numbers were removed.

If a potential heading was found within the text, we excluded it, because it would potentially include biasing information such as judge names. It is especially important to exclude judge names, as the model could focus on judge names as a proxy for the directionality as most cases were decided without dissent. This is an issue in our empirical context because we would like to use the predicted data to analyze judge characteristics. Including the judges in the prediction would induce mechanical correlation.

In a second step, we applied regular expressions to capture the part of the opinion in which judges might dissent from the majority. Including a dissenting part that goes against the directionality of the majority in the input would not only add noise but may also lead the classifier to average over the different directions, leading to overall worse performance. If we found a dissent, we split off the relevant paragraph and saved it as an extra entry in the database, marking it as "dissent". We excluded those entries and did not use them as input.

## D.3 All Classifier Input Combinations

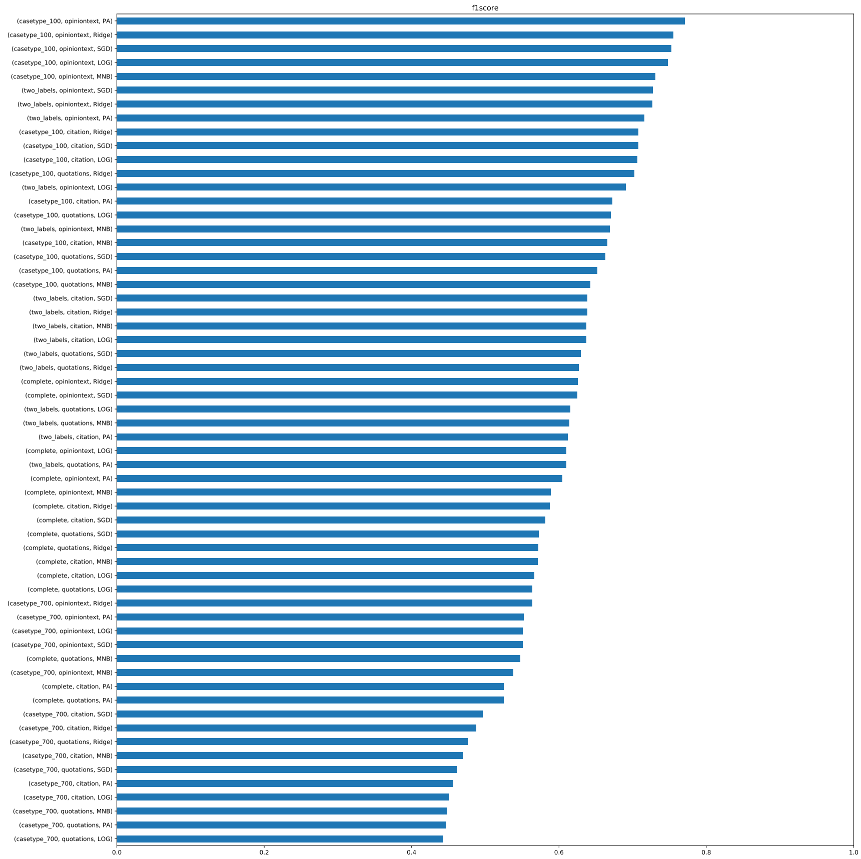


**Figure D.1:** F1 score

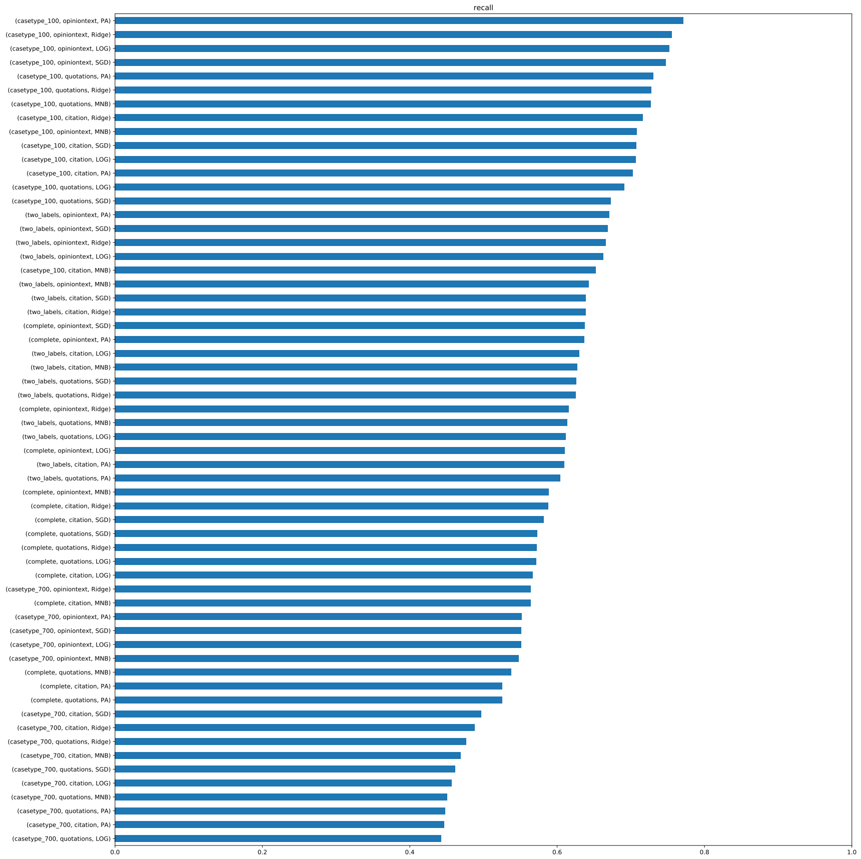


**Figure D.2:** Recall Score

## D.4 Judges

Tables D.2, D.3, D.4 and D.5 present yet another way how to assess the performance of the best classifier. We predict the directionality of an opinion and use it to calculate the fraction of conservative or liberal votes by a judge. We split the judges' population by the party of the appointing president, resulting in four different specifications. Overall, actual and predicted fractions of votes by the ten highest ranked judges by specification are pretty similar and reassure that our classifier performs sufficiently well for our analysis.

**Table D.2:** 10 Judges with the Highest Fraction of Conservative Votes, Appointed by Conservative Presidents

| Frac con | sum | name |
|---:|---:|:---|
| 0.89 | 48 | Barksdale, Rhesa H. |
| 0.85 | 69 | Loken, James B. |
| 0.84 | 65 | Hansen, David R. |
| 0.83 | 110 | Easterbrook, Frank H. |
| 0.82 | 28 | O'Scannlain, Diaruid F. |
| 0.82 | 61 | Luttig, J. Michael |
| 0.80 | 93 | Edmondson, James L. |
| 0.80 | 72 | Magill, Frank J. |
| 0.80 | 104 | Boudin, Michael |
| 0.80 | 45 | DeMoss, Harold R., Jr. |
| *Note:* | | hand-labelled data |

| Frac con | sum | name |
|---:|---:|:---|
| 0.87 | 48 | Barksdale, Rhesa H. |
| 0.87 | 69 | Loken, James B. |
| 0.83 | 66 | Arnold, Morris S. |
| 0.82 | 109 | Easterbrook, Frank H. |
| 0.80 | 15 | Lewis, Robert E. |
| 0.80 | 65 | Hansen, David R. |
| 0.80 | 44 | DeMoss, Harold R., Jr. |
| 0.79 | 61 | Jones, Edith H. |
| 0.79 | 103 | Boudin, Michael |
| 0.78 | 97 | Higginbotham, Patrick E. |
| *Note:* | | predicted data |

**Table D.3:** 10 Judges with the Highest Fraction of Liberal Votes, Appointed by Conservative Presidents

| Frac lib | sum | name |
|---:|---:|:---|
| 0.71 | 11 | Thomas, Clarence |
| 0.63 | 44 | Hitz, William |
| 0.59 | 137 | Gibbons, John J. |
| 0.58 | 39 | Waddill, Edmund, Jr. |
| 0.58 | 46 | Miller, William Ernest |
| 0.58 | 73 | Mansmann, Carol Los |
| 0.58 | 43 | Pratt, George C. |
| 0.56 | 56 | Roth, Jane R. |
| 0.56 | 142 | Northcutt, Elliott |
| 0.56 | 107 | Lively, Frederick P. |
| *Note:* | | hand-labelled data |

| Frac lib | sum | name |
|---:|---:|:---|
| 0.63 | 44 | Hitz, William |
| 0.62 | 11 | Thomas, Clarence |
| 0.57 | 119 | Wilbur, Curtis D. |
| 0.56 | 116 | Van Orsdel, Josiah A. |
| 0.56 | 70 | Thompson, Joseph W. |
| 0.56 | 46 | Miller, William Ernest |
| 0.56 | 55 | Roth, Jane R. |
| 0.56 | 142 | Northcutt, Elliott |
| 0.56 | 108 | Lively, Frederick P. |
| 0.55 | 43 | Pratt, George C. |
| *Note:* | | predicted data |

## D.5 Robustness Checks

Additionally to the histograms that Figure 4.8 provides, we go on to analyze the EBA's statistics on civil cases, displayed by Table D.6a.

For civil cases, we estimated 510 regression models. Figure 4.8a provides information about the share of regression coefficients that are statistically significant as well as lower (column 1) or greater (column 2) than zero. There was no coefficient significant for which the size of at least 50% of estimated coefficients lies below zero. By contrast, there were three coefficients found to be significant while having values larger than zero in at least 50% of the estimated models. These were the fraction of republican senators at the point of election (92%), the fraction of miscellaneous votes (64%) as well as circuit 1 (100%).

Consequently, Leamer (1985)'s EBA (column 3), defines circuit 1 as the only robust vari-

**Table D.4:** 10 Judges with the Highest Fraction of Conservative Votes, Appointed by Liberal Presidents

| Frac con | sum | name |
|---|---|---|
| 0.89 | 45 | Evans, Terence Thomas |
| 0.84 | 38 | Parker, Robert Manley |
| 0.78 | 69 | Williams, Jerre S. |
| 0.76 | 83 | Garza, Reynaldo |
| 0.75 | 60 | Anderson, Robert P. |
| 0.74 | 27 | King, Carolyn Dineen |
| 0.74 | 78 | Mehaffy, Pat |
| 0.73 | 131 | Miller, Wilbur K., Jr. |
| 0.73 | 37 | Murphy, Michael R. |
| 0.73 | 11 | Kravitch, Phyllis A. |
| *Note:* | | hand-labelled data |

| Frac con | sum | name |
|---|---|---|
| 0.82 | 44 | Evans, Terence Thomas |
| 0.81 | 37 | Parker, Robert Manley |
| 0.80 | 20 | Rutledge, Wiley Blount |
| 0.78 | 27 | King, Carolyn Dineen |
| 0.76 | 82 | Garza, Reynaldo |
| 0.75 | 134 | Breyer, Stephen G. |
| 0.74 | 163 | McMillian, Theodore |
| 0.74 | 19 | Cole, Ransey Guy, Jr. |
| 0.74 | 68 | Williams, Jerre S. |
| 0.73 | 30 | Stewart, Carl Edmond |
| *Note:* | | predicted data |

**Table D.5:** 10 Judges with the Highest Fraction of Liberal Votes, Appointed by Liberal Presidents

| Frac lib | sum | name |
|---|---|---|
| 0.71 | 11 | Faris, Charles |
| 0.71 | 11 | Thomas, Sidney Runyan |
| 0.67 | 16 | Hough, Charles M. |
| 0.66 | 24 | Russell, Robert L. |
| 0.66 | 51 | Haney, Bert E. |
| 0.65 | 29 | Ferguson, Warren J. |
| 0.63 | 99 | Higginbotham, Aloyisus Leon |
| 0.63 | 14 | Sarokin, Haddon Lee |
| 0.63 | 22 | Strum, Louie |
| 0.62 | 24 | Clark, William |
| *Note:* | | hand-labelled data |

| Frac lib | sum | name |
|---|---|---|
| 0.66 | 24 | Russell, Robert L. |
| 0.63 | 14 | Sarokin, Haddon Lee |
| 0.63 | 22 | Strum, Louie |
| 0.62 | 27 | O'Connell, John J. |
| 0.62 | 24 | Clark, William |
| 0.61 | 16 | Hough, Charles M. |
| 0.61 | 98 | Higginbotham, Aloyisus Leon |
| 0.60 | 150 | Robinson, Spottswood W., III |
| 0.60 | 51 | Haney, Bert E. |
| 0.57 | 31 | Lucero, Carlos |
| *Note:* | | predicted data |

able.

Furthermore, Table D.6a includes results from Sala-i-Martin (1997)'s EBA (columns 4 and 5). Figure 4.8a suggests that a normal distribution does not sufficiently well approximate the regression coefficients' distribution. For this reason, we focus on Sala-i-Martin (1997) EBA results from a model that does make assumptions about the coefficients' distributions. As a rule of thumb, those variables for which more than 90% of the regression coefficients' cumulative distribution is located either above or below zero, can be interpreted as being robustly connected with the dependent variable (Hlavac 2016). For the variables of being black (96%), the years of having served as a district court judge (93%), as well as for the fraction of economic votes (93%), more than 90% of the cumulative distributions lie below zero. By contrast, for the variables of being appointed by a conservative president (99%), the fraction of miscellaneous votes (98%) as well as for circuit 1 (100%), more than 90% of the cumulative distributions lie above zero.

EBA statistics for criminal cases, displayed in D.6b, are interpreted below. Overall, 127

**Table D.6:** Extreme Bounds Analysis I

**(a)** Civil Cases

|  | $\beta$ sign & $< 0$ | $\beta$ sign & $> 0$ | leamer robust | cdf $\beta <= 0$ generic | cdf $\beta > 0$ generic |
|---|---|---|---|---|---|
| (Intercept) | 0.25 | 0.50 | FALSE | 0.47 | 0.53 |
| Pres | 0.00 | 0.92 | FALSE | 0.01 | 0.99 |
| SenRep | 0.00 | 0.00 | FALSE | 0.30 | 0.70 |
| YrAppt | 0.00 | 0.50 | FALSE | 0.11 | 0.89 |
| Gender | 0.00 | 0.00 | FALSE | 0.33 | 0.67 |
| Black | 0.47 | 0.00 | FALSE | 0.96 | 0.04 |
| District Court | 0.01 | 0.00 | FALSE | 0.93 | 0.07 |
| FracEcon | 0.50 | 0.00 | FALSE | 0.95 | 0.05 |
| FracMisc | 0.00 | 0.64 | FALSE | 0.02 | 0.98 |
| Circuit 1 | 0.00 | 1.00 | TRUE | 0.00 | 1.00 |
| Circuit 2 | 0.00 | 0.00 | FALSE | 0.52 | 0.48 |
| Circuit 3 | 0.00 | 0.00 | FALSE | 0.91 | 0.09 |
| Circuit 4 | 0.00 | 0.00 | FALSE | 0.62 | 0.38 |
| Circuit 5 | 0.00 | 0.00 | FALSE | 0.29 | 0.71 |
| Circuit 6 | 0.00 | 0.00 | FALSE | 0.42 | 0.58 |
| Circuit 7 | 0.00 | 0.00 | FALSE | 0.08 | 0.92 |
| Circuit 8 | 0.00 | 0.00 | FALSE | 0.21 | 0.79 |
| Circuit 9 | 0.00 | 0.00 | FALSE | 0.83 | 0.17 |
| Circuit 10 | 0.00 | 0.00 | FALSE | 0.71 | 0.29 |
| Circuit 11 | 0.00 | 0.00 | FALSE | 0.43 | 0.57 |

regression models were estimated.

Columns 1 and 2 of Table D.6b show the fraction of the respective regression coefficients that are statistically significant and lower or greater than zero at the same time. Only for the dummy variable *Black*, more than 88% of the values estimated were significant and smaller than zero.

By contrast, there were three coefficients, *Pres* (100%), *circuit 8* (100%) and *circuit 10* (100%) found to be significant and showing more than 50% of its values larger than zero. Table D.6b summarizes results from Leamer (1985)'s EBA (column 3). This test concludes that three variables are found to be robustly connected with the dependent variable, which are *Pres* as well as *circuits 8* and *10*.

Furthermore, Table D.6b includes results from Sala-i-Martin (1997)'s EBA (columns 4 and 5). As was the case with civil cases, Figure 4.8b suggests that a normal distribution does not fit the coefficients' distribution very well. For this reason, we focus on EBA results from

**Table D.6:** Extreme Bounds Analysis II

**(b)** Criminal Cases

| | $\beta$ sign & $< 0$ | $\beta$ sign & $> 0$ | leamer robust | cdf $\beta <= 0$ generic | cdf $\beta > 0$ generic |
|---|---|---|---|---|---|
| (Intercept) | 0.00 | 0.50 | FALSE | 0.14 | 0.86 |
| Pres | 0.00 | 1.00 | TRUE | 0.00 | 1.00 |
| SenRep | 0.00 | 0.00 | FALSE | 0.70 | 0.30 |
| YrAppt | 0.00 | 0.00 | FALSE | 0.44 | 0.56 |
| Gender | 0.00 | 0.00 | FALSE | 0.61 | 0.39 |
| Black | 0.88 | 0.00 | FALSE | 0.99 | 0.01 |
| District Court | 0.00 | 0.00 | FALSE | 0.78 | 0.22 |
| Circuit 1 | 0.00 | 0.00 | FALSE | 0.06 | 0.94 |
| Circuit 2 | 0.00 | 0.00 | FALSE | 0.60 | 0.40 |
| Circuit 3 | 0.00 | 0.00 | FALSE | 0.71 | 0.29 |
| Circuit 4 | 0.00 | 0.00 | FALSE | 0.46 | 0.54 |
| Circuit 5 | 0.00 | 0.00 | FALSE | 0.25 | 0.75 |
| Circuit 6 | 0.00 | 0.00 | FALSE | 0.49 | 0.51 |
| Circuit 7 | 0.00 | 0.00 | FALSE | 0.07 | 0.93 |
| Circuit 8 | 0.00 | 1.00 | TRUE | 0.01 | 0.99 |
| Circuit 9 | 0.00 | 0.00 | FALSE | 0.33 | 0.67 |
| Circuit 10 | 0.00 | 1.00 | TRUE | 0.01 | 0.99 |
| Circuit 11 | 0.00 | 0.00 | FALSE | 0.14 | 0.86 |

a parameter-free model. For *Black* (99%), more than 90% of the cumulative distributions lie below zero. By contrast, for the variables of being appointed by a conservative president (*Pres*) (100%), for circuit 1 (94%), circuit 7 (93%), circuit 8 (99%) and circuit 10 (99%) more than 90% of the cumulative distributions lie above zero.

# References

Abadie, Alberto (2020). "Statistical Nonsignificance in Empirical Economics". In: *American Economic Review: Insights* 2.2, pp. 193–208. DOI: 10.1257/aeri.20190252.

Aletras, Nikolaos et al. (2016). "Predicting Judicial Decisions of the European Court of Human Rights: A Natural Language Processing Perspective". In: *PeerJ Computer Science* 2016.10, e93. DOI: 10.7717/peerj-cs.93.

Alonso-Diaz, Santiago, Jessica F. Cantlon, and Steven T. Piantadosi (2018). "A Threshold-Free Model of Numerosity Comparisons". In: *PLoS ONE* 13.4. DOI: 10.1371/journal.pone.0195188.

Amin, Engi, Mohamed Abouelela, and Amal Soliman (2018). "The Role of Heterogeneity and the Dynamics of Voluntary Contributions to Public Goods: An Experimental and Agent-Based Simulation Analysis". In: *Journal of Artificial Societies and Social Simulation* 21.1.

Andreoni, James (1995). "Cooperation in Public-Goods Experiments: Kindness or Confusion?" In: *American Economic Review*, pp. 891–904.

Andres, Maximilian, Lisa Bruttel, and Jana Friedrichsen (2019). "The Effect of Leniency Rule on Cartel Formation and Stability: Experiments with Open Communication".

Arad, Ayala and Stefan Penczynski (2018). "Multi-Dimensional Reasoning in Competitive Resource Allocation Games : Evidence from Intra-Team Communication".

Arbelaitz, Olatz et al. (2013). "An Extensive Comparative Study of Cluster Validity Indices". In: *Pattern Recognition* 46.1, pp. 243–256.

Arifovic, Jasmina and John Ledyard (2012). "Individual Evolutionary Learning, Other-Regarding Preferences, and the Voluntary Contributions Mechanism". In: *Journal of Public Economics* 96.9-10, pp. 808–823.

Ash, Elliott and Daniel L. Chen (2019). "Mapping the Geometry of Law Using Document Embeddings". In: *SSRN Electronic Journal*. DOI: 10.2139/ssrn.3305761.

Ash, Elliott, Daniel L. Chen, and Wei Lu (2018). "Motivated Reasoning in the Field : Polarization in Precedent, Prose, Vote, and Retirement in U.S. Circuit Courts". In: *SSRN Electronic Journal* 614708, pp. 1–33.

Badir Alnidawy, Abdul Azez (2015). "The Effect of Emotional Intelligence on Job Satisfaction: Applied Study in the Jordanian Telecommunication Sector". In: *International Journal of Business Administration* 6.3, pp. 1–34. DOI: 10.5430/ijba.v6n3p63.

Banea, Carmen et al. (2015). *SimCompass: Using Deep Learning Word Embeddings to Assess Cross-level Similarity*. DOI: 10.3115/v1/s14-2098.

Bapna, Ravi et al. (2004). "User Heterogeneity and its Impact on Electronic Auction Market Design: An Empirical Exploration". In: *MIS Quarterly*, pp. 21–43.

Begum, Nurjahan et al. (2015). "Accelerating Dynamic Time Warping Clustering with a Novel Admissible Pruning Strategy". In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 49–58.

Bereby-Meyer, Yoella et al. (2018). "Honesty Speaks a Second Language". In: *Topics in Cognitive Science*, pp. 1–12. DOI: 10.1111/tops.12360.

Berndt, Donald J. and James Clifford (1994). "Using Dynamic Time Warping to Find Patterns in Time Series." In: *KDD workshop*. Vol. 10. 16. Seattle, WA, USA: pp. 359–370.

Bezdek, James C (2013). *Pattern Recognition With Fuzzy Objective Function Algorithms*. Springer Science & Business Media.

BMF (2020). *BMF-Monatsbericht November 2019 - Verfolgung von Steuerstraftaten und Steuerordnungswidrigkeiten im Jahr 2018*. URL: https://www.bundesfinanzministerium.de/Monatsberichte/2019/11/Inhalte/Kapitel-3-Analysen/3-4-steuerstrafsachen.html (visited on 07/15/2020).

Boella, Guido et al. (2012). "Using Legal Ontology to Improve Classification in the Eunomos Legal Document and Knowledge Management System". In: *Semantic Processing of Legal Texts (SPLeT-2012) Workshop Programme*, p. 13.

Brehmer, Yvonne, Helena Westerberg, and Lars Bäckman (2012). "Working-Memory Training in Younger and Older Adults: Training Gains, Transfer, and Maintenance". In: *Frontiers in Human Neuroscience* 6, p. 63. DOI: 10.3389/fnhum.2012.00063.

Breiman, Leo (1996). "Bagging Predictors". In: *Machine Learning* 24.2, pp. 123–140. DOI: 10.1007/bf00058655.

– (1998). "Arcing Classifiers". In: *Annals of Statistics* 26.3, pp. 801–849. DOI: 10.1214/aos/1024691079.

– (2001). "Random Forests". In: *Machine Learning* 45.1, pp. 5–32. DOI: 10.1023/A:1010933404324.

Burchardi, Konrad B. and Stefan P. Penczynski (2014). "Out of Your Mind: Eliciting Individual Reasoning in One Shot Games". In: *Games and Economic Behavior* 84, pp. 39–57. DOI: 10.1016/j.geb.2013.12.005.

Cao, Yu, Elliott Ash, and Daniel L. Chen (2018). "Automated Fact-Value Distinction in Court Opinions". In: *SSRN Electronic Journal*. DOI: 10.2139/ssrn.3202438.

Capra, C. Mónica (2019). "Understanding Decision Processes in Guessing Games: A Protocol Analysis Approach". In: *Journal of the Economic Science Association* 5.1, pp. 123–135. DOI: 10.1007/s40881-019-00074-0.

Capraro, Valerio (2017). "Does the Truth Come Naturally? Time Pressure Increases Honesty in One-Shot Deception Games". In: *Economics Letters* 158.September, pp. 54–57. DOI: 10.1016/j.econlet.2017.06.015.

Casper, Gerhard and Richard A. Posner (1974). "A Study of the Supreme Court's Caseload". In: *The Journal of Legal Studies* 3.2, pp. 339–375. DOI: 10.1086/467517.

Chandler, Seth (2007). "The Network Structure of Supreme Court Jurisprudence". In: *The Mathematica Journal* 10.3, p. 1. DOI: 10.3888/tmj.10.3-5.

Chaudhuri, Ananish (2011). "Sustaining Cooperation in Laboratory Public Goods Experiments: a Selective Survey of the Literature". In: *Experimental Economics* 14.1, pp. 47–83.

Chen, Daniel L., Martin Schonger, and Chris Wickens (2016). "oTree – An Open-Source Platform for Laboratory, Online, and Field Experiments". In: *Journal of Behavioral and Experimental Finance* 9, pp. 88–97. DOI: 10.1016/j.jbef.2015.12.001.

Crammer, Koby et al. (2006). "Online Passive-Aggressive Algorithms". In: *Journal of Machine Learning Research* 7, pp. 551–585. DOI: 10.1.1.9.3429.

Cuturi, Marco (2011). "Fast Global Alignment kernels". In: *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pp. 929–936.

Dainow, Joseph (1966). "The Civil Law and the Common Law: Some Points of Comparison". In: *The American Journal of Comparative Law* 15.3, p. 419. DOI: 10.2307/838275.

Dana, Jason, Roberto A. Weber, and Jason Xi Kuang (2007). "Exploiting Moral Wiggle Room: Experiments Demonstrating an Illusory Preference for Fairness". In: *Economic Theory* 33.1, pp. 67–80. DOI: 10.1007/s00199-006-0153-z.

DePaulo, Bella M. et al. (2003). "Cues to Deception". In: *Psychological Bulletin* 129.1, pp. 74–112. DOI: 10.1037//0033-2909.129.1.74.

Diederich, Johannes, Timo Goeschl, and Israel Waichman (2016). "Group Size and the (In)Efficiency of Pure Public Good Provision". In: *European Economic Review* 85, pp. 272–287.

Dotan, Dror, Florent Meyniel, and Stanislas Dehaene (2018). "Online Confidence Monitoring During Decision-Making". In: *Cognition* 171, pp. 112–121. DOI: 10.1016/j.cognition.2017.11.001.

Elten, Jonas van and Stefan P. Penczynski (2020). "Coordination Games with Asymmetric Payoffs: An Experimental Study with Intra-Group Communication". In: *Journal of Economic Behavior and Organization* 169, pp. 158–188. DOI: 10.1016/j.jebo.2019.11.006.

Engel, Christoph (2019). "Estimating Heterogeneous Reactions to Experimental Treatments". In: *SSRN Electronic Journal*, pp. 1–30. DOI: 10.2139/ssrn.3322322.

– (2020). "Estimating Heterogeneous Reactions to Experimental Treatments". Working paper.

Engel, Christoph, Sebastian Kube, and Michael Kurschilgen (2020). "Managing Expectations: How Selective Information Affects Cooperation". Working paper.

Engel, Christoph and Michael Kurschilgen (2013). "The Coevolution of Behavior and Normative Expectations: An Experiment". In: *American Law and Economics Review* 15.2, pp. 578–609.

– (2014). "The Jurisdiction of the Man Within". Working paper.

Engel, Christoph and Michael Kurschilgen (2020). “The Fragility of a Nudge: the Power of Self-Set Norms to Contain a Social Dilemma”. In: *Journal of Economic Psychology*, forthcoming.

Engel, Christoph and Bettina Rockenbach (2020). “What Makes Cooperation Precarious?” Working paper.

Epstein, Lee et al. (2012). “Ideology and the Study of Judicial Behavior”. In: *Ideology, Psychology, and Law* 705. DOI: 10.1093/acprof:oso/9780199737512.003.0027.

Erb, Christopher D. et al. (2016). “Reach Tracking Reveals Dissociable Processes Underlying Cognitive Control”. In: *Cognition* 152, pp. 114–126. DOI: 10.1016/j.cognition.2016.03.015.

Fabian Pedregosa et al. (2011). “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12.Oct, pp. 2825–2830.

Fehr, Ernst and Bettina Rockenbach (2003). “Detrimental Effects of Sanctions on Human Altruism”. In: *Nature* 422.6928, pp. 137–140. DOI: 10.1038/nature01474.

Fehr, Ernst and Klaus Schmidt (2002). “Theories of Fairness and Reciprocity. Evidence and Economic Applications”. In: *Advances in Economics and Econometrics. 8th World Congress*. Ed. by Mathias Dewatripont and Stephen J. Turnovsky. Cambridge: Cambridge University Press, pp. 208–257.

Ficici, Sevan G., David C. Parkes, and Avi Pfeffer (2012). “Learning and Solving Many-Player Games Through a Cluster-Based Representation”. In: *arXiv preprint arXiv:1206.3253*.

Fischbacher, Urs and Franziska Föllmi-Heusi (2013). “Lies in Disguise–An Experimental Study on Cheating”. In: *Journal of the European Economic Association* 11.3, pp. 525–547. DOI: 10.1111/jeea.12014.

Fischbacher, Urs and Simon Gachter (2010). “Social Preferences, Beliefs, and the Dynamics of Free Riding in Public Goods Experiments”. In: *American Economic Review* 100.1, pp. 541–56.

Fischbacher, Urs, Simon Gachter, and Ernst Fehr (2001). “Are People Conditionally Cooperative? Evidence from a Public Goods Experiment”. In: *Economics Letters* 71.3, pp. 397–404.

Fochmann, Martin et al. (2019). “Dishonesty and Risk-Taking: Compliance Decisions of Individuals and Groups”.

Foerster, Anna et al. (2013). “Honesty Saves Time (and Justifications)”. In: *Frontiers in Psychology* 4.July, p. 473. DOI: 10.3389/fpsyg.2013.00473.

Fong, Christian and Matthew Tyler (2020). “Machine Learning Predictions as Regression Covariates”.

Frederick, Shane (2005). “Cognitive Reflection and Decision-Making”. In: *Journal of Economic Perspectives* 19.4, pp. 25–42. DOI: 10.1257/089533005775196732.

Frey, Bruno S. and Felix Oberholzer-Gee (1997). “The Cost of Price Incentives: An Empirical Analysis of Motivation Crowding-Out”. In: *American Economic Review* 87.4, pp. 746–755. DOI: 10.2307/2951373.

Friedman, Jason, Scott Brown, and Matthew Finkbeiner (2013). “Linking Cognitive and Reaching Trajectories via Intermittent Movement Control”. In: *Journal of Mathematical Psychology* 57.3-4, pp. 140–151. DOI: 10.1016/j.jmp.2013.06.005.

Ganis, Giorgi et al. (2003). “Neural Correlates of Different Types of Deception: An fMRI Investigation”. In: *Cerebral Cortex* 13.8, pp. 830–836. DOI: 10.1093/cercor/13.8.830.

Georgalos, Konstantinos and John Hey (2019). “Testing for the Emergence of Spontaneous Order”. In: *Experimental Economics*, pp. 1–21. DOI: 10.1007/s10683-019-09637-8.

Giles, Micheal W., Virginia A. Hettinger, and Todd Peppers (2001). “Picking Federal Judges: A Note on Policy and Partisan Selection Agendas”. In: *Political Research Quarterly* 54.3, pp. 623–641. DOI: 10.1177/106591290105400307.

Ginn, Martha Humphries, Kathleen Searles, and Amanda Jones (2015). “Vouching for the Court? How High Stakes Affect Knowledge and Support of the Supreme Court”. In: *Justice System Journal* 36.2, pp. 163–179. DOI: 10.1080/0098261X.2014.965854.

Gneezy, Uri (2005). “Deception: The Role of Consequences”. In: *American Economic Review* 95.1, pp. 384–394. DOI: 10.1257/0002828053828662.

Gneezy, Uri and Aldo Rustichini (2000). “Pay Enough or Don’t Pay at all”. In: *Quarterly Journal of Economics* 115.3, pp. 791–810. DOI: 10.1162/003355300554917.

Gunia, Brian C. et al. (2012). “Contemplation and Conversation: Subtle Influences on Moral Decision-Making”. In: *Academy of Management Journal* 55.1, pp. 13–33. DOI: 10.5465/amj.2009.0873.

Halko, Nathan, Per-Gunnar Martinsson, and Joel A. Tropp (2011). “Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions”. In: *SIAM Review* 53.2, pp. 217–288. DOI: 10.1137/090771806.

Harris, Zellig S. (1954). “Distributional Structure”. In: *WORD* 10.2-3, pp. 146–162. DOI: 10.1080/00437956.1954.11659520.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media.

Hlavac, Marek (2016). “ExtremeBounds: Extreme Bounds Analysis in R”. In: *Journal of Statistical Software* 72.9. DOI: 10.18637/jss.v072.i09.

Honnibal, Mattew and Ines Montani (2017). “spaCy2: Natural Language Understanding with Bloom Embeddings, Convolutional Neural Networks and Incremental Parsing”. In: *Features* 7.1. DOI: 10.5281/zenodo.1212304.

Houser, Daniel, Stefan Vetter, and Joachim Winter (2012). “Fairness and Cheating”. In: *European Economic Review* 56.8, pp. 1645–1655. DOI: 10.1016/j.euroecorev.2012.08.001.

Hu, Baotian, Zhengdong Lu, et al. (2014). “Convolutional Neural Network Architectures for Matching Natural Language Sentences”. In: *Advances in Neural Information Processing Systems*. Vol. 3. January, pp. 2042–2050.

Hu, Xiaoqing, J. Peter Rosenfeld, and Galen V. Bodenhausen (2012). “Combating Automatic Autobiographical Associations: The Effect of Instruction and Training in Strategically Concealing Information in the Autobiographical Implicit Association Test". In: *Psychological Science* 23.10, pp. 1079–1085. DOI: 10.1177/0956797612443834.

Isaac, R. Mark and James M. Walker (1988). "Group Size Effects in Public Goods Provision: The Voluntary Contributions Mechanism". In: *Quarterly Journal of Economics* 103.1, pp. 179–199.

Janke, S. and A. Glöckner-Rist (2014). *Deutsche Version des Positive and Negative Affect Schedule (PANAS).* DOI: 10.6102/zis146.

Jeffreys, Harold (1998). *The Theory of Probability*. OUP Oxford.

Joachims, Thorsten (1998). "Text Categorization with Support Vector Machines: Learning with Many Relevant Features". In: *Lecture Notes in Computer Science*. Vol. 1398. Springer, pp. 137–142. DOI: 10.1007/s13928716.

Johnson, Ray, Heather Henkell, et al. (2008). "The Self in Conflict: The Role of Executive Processes During Truthful and Deceptive Responses about Attitudes". In: *NeuroImage* 39.1, pp. 469–482. DOI: 10.1016/j.neuroimage.2007.08.032.

Johnson, Susan W., Donald R. Songer, and Nadia A. Jilani (2011). "Judge Gender, Critical Mass, and Decision-Making in the Appellate Courts of Canada". In: *Journal of Women, Politics and Policy* 32.3, pp. 237–260. DOI: 10.1080/1554477X.2011.589293.

Joulin, Armand et al. (2017). "Bag of Tricks for Efficient Text Classification". In: *arXiv preprint arXiv:1607.01759* 2, pp. 427–431. DOI: 10.18653/v1/e17-2068.

Kassow, Benjamin, Donald R. Songer, and Michael P. Fix (2012). "The Influence of Precedent on State Supreme Courts". In: *Political Research Quarterly* 65.2, pp. 372–384. DOI: 10.1177/1065912910391477.

Kenter, Tom and Maarten De Rijke (2015). "Short Text Similarity with Word Embeddings". In: *Proceedings of the 24th ACM international on conference on information and knowledge management*. Vol. October, pp. 1411–1420. DOI: 10.1145/2806416.2806475.

Kim, Yoon (2014). "Convolutional Neural Networks for Sentence Classification". In: *arXiv preprint arXiv:1408.5882*. DOI: 10.3115/v1/d14-1181.

Kocher, Martin G., Simeon Schudy, and Lisa Spantig (2018). "I lie? We lie! Why? Experimental Evidence on a Dishonesty Shift in Groups". In: *Management Science* 64.9, pp. 3995–4008. DOI: 10.1287/mnsc.2017.2800.

Kosfeld, Michael, Akira Okada, and Arno Riedl (2009). "Institution Formation in Public Goods Games". In: *American Economic Review* 99.4, pp. 1335–55.

Landes, William M. and Richard A. Posner (2011). "Rational Judicial Behavior: A Statistical Study". In: *Journal of Legal Analysis* 1.2, pp. 775–831. DOI: 10.2139/ssrn.1463483.

Lauderdale, Benjamin E. and Tom S. Clark (2014). "Scaling Politically Meaningful Dimensions Using Texts and Votes". In: *American Journal of Political Science* 58.3, pp. 754–771. DOI: 10.1111/ajps.12085.

Lauderdale, Benjamin E. and Alexander Herzog (2016). "Measuring Political Positions from Legislative Speech". In: *Political Analysis* 24.3, pp. 374–394. DOI: 10.1093/pan/mpw017.

Laver, Michael, Kenneth Benoit, and John Garry (2003). “Extracting Policy Positions from Political Texts Using Words as Data”. In: *American Political Science Review* 97.2, pp. 311–331. DOI: 10.1017/S0003055403000698.

Le, Quoc and Tomas Mikolov (2014). “Distributed Representations of Sentences and Documents”. In: *31st International Conference on Machine Learning*. Vol. 4, pp. 2931–2939.

Leamer, Edward E. (1985). “Sensitivity Analyses Would Help”. In: *The American Economic Review* 75.3, pp. 308–313.

Ledyard, John O (1995). “Public Goods: A Survey of Experimental Research”. In: *Handbook of Experimental Economics*. Ed. by John Kagel and Al Roth. Princeton: Princeton University Press, pp. 111–194.

Levenshtein, Vladimir (1966). “Binary Codes Capable of Correcting Deletions, Insertions, and Reversals”. In: *Soviet Physics Doklady*. Vol. 10. 8, pp. 707–710.

Liao, T. Warren (2005). “Clustering of Time Series Data – a Survey”. In: *Pattern Recognition* 38.11, pp. 1857–1874.

Lipton, Zachary C., Charles Elkan, and Balakrishnan Naryanaswamy (2014). “Optimal Thresholding of Classifiers to Maximize F1 Measure”. In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, pp. 225–239.

Lohse, Tim, Sven A. Simon, and Kai A. Konrad (2018). “Deception under Time Pressure: Conscious Decision or a Problem of Awareness?” In: *Journal of Economic Behavior and Organization*. Vol. 146, pp. 31–42. DOI: 10.1016/j.jebo.2017.11.026.

López-Pérez, Raúl and Eli Spiegelman (2019). “Do Economists lie More?” In: *Dishonesty in Behavioral Economics*. Academic Press, pp. 143–162. DOI: 10.1016/B978-0-12-815857-9.00003-0.

Lu, Yixin et al. (2016). “Exploring Bidder Heterogeneity in Multichannel Sequential B2B Auctions”. In: *MIS Quarterly* 40.3, pp. 645–662.

Lucas, Pablo, Angela de Oliveira, and Sheheryar Banuri (2012). “The Effects of Group Composition and Social Preference Heterogeneity in a Public Goods Game: An Agent-Based Simulation”. In: *Journal of Artificial Societies and Social Simulation* 17.3, pp. 148–174.

MacLeod, Colin M. and Kevin Dunbar (1988). “Training and Stroop-Like Interference: Evidence for a Continuum of Automaticity”. In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 14.1, pp. 126–135. DOI: 10.1037/0278-7393.14.1.126.

Martin, Andrew D. and Kevin M. Quinn (2001). “The Dimensions of Supreme Court Decision Making: Again Revisiting the Judicial Mind”. In: *Meeting of the Midwest Political Science Association*.

– (2002). “Dynamic Ideal Point Estimation via Markov Chain Monte Carlo for the U.S. Supreme Court, 1953–1999”. In: *Political Analysis* 10.2, pp. 134–153. DOI: 10.1093/pan/10.2.134.

Martin, Andrew D., Kevin Quinn, and Lee Epstein (2005). “The Median Justice on the U.S. Supreme Court”. In: *North Carolina Law Review* 83.5, pp. 1275–1322.

Masood, Ali S. and Donald R. Songer (2013). "Reevaluating the Implications of Decision-Making Models". In: *Journal of Law and Courts* 1.2, pp. 363–389. DOI: 10.1086/670745.

Mcfadden, Daniel (1974). "Conditional Logit Analysis of Qualitative Choice Behavior". In: *Frontiers in Econometrics*, pp. 105–42.

McLachlan, Geoffrey J. and David Peel (2004). *Finite Mixture Models*. John Wiley & Sons.

Mikolov, Tomas et al. (2013). "Distributed Representations of Words and Phrases and their Compositionality". In: *Advances in Neural Information Processing Systems*, pp. 3111–3119.

Milham, M. P. et al. (2003). "Practice-Related Effects Demonstrate Complementary Roles of Anterior Cingulate and Prefrontal Cortices in Attentional Control". In: *NeuroImage* 18.2, pp. 483–493. DOI: 10.1016/S1053-8119(02)00050-2.

Nikiforakis, Nikos and Hans-Theo Normann (2008). "A Comparative Statics Analysis of Punishment in Public-Good Experiments". In: *Experimental Economics* 11.4, pp. 358–369.

Norvig, Peter (2007). *How to Write a Spelling Corrector*.

Olesen, Pernille J., Helena Westerberg, and Torkel Klingberg (2004). "Increased Prefrontal and Parietal Activity after Training of Working Memory". In: *Nature Neuroscience* 7.1, pp. 75–79. DOI: 10.1038/nn1165.

Penczynski, Stefan P. (2019). "Using Machine Learning for Communication Classification". In: *Experimental Economics* 22.4, pp. 1002–1029. DOI: 10.1007/s10683-018-09600-z.

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning (2014). "GloVe: Global Vectors for Word Representation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543. DOI: 10.3115/v1/d14-1162.

Pinheiro-Chagas, Pedro et al. (2017). "Finger Tracking Reveals the Covert Stages of Mental Arithmetic". In: *Open Mind* 1.1, pp. 30–41. DOI: 10.1162/OPMI-a-00003.

Pirolli, Peter L. and John R. Anderson (1985). "The Role of Practice in Fact Retrieval". In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 11.1, pp. 136–153. DOI: 10.1037/0278-7393.11.1.136.

Platt, John C, Nello Cristianini, and John Shawe-Taylor (2000). "Large Margin DAGs for Multiclass Classification". In: *Advances in neural information processing systems*, pp. 547–553.

Randazzo, Kirk A., Richard W. Waterman, and Michael P. Fix (2011). "State Supreme Courts and the Effects of Statutory Constraint: A Test of the Model of Contingent Discretion". In: *Political Research Quarterly* 64.4, pp. 779–789. DOI: 10.1177/1065912910379229.

Ratanamahatana, Chotirat Ann and Eamonn Keogh (2004). "Everything You Know about Dynamic Time Warping is Wrong". In: *Third Workshop on Mining Temporal and Sequential Data*. Vol. 32. Citeseer.

Rauh, Christian (2018). “Validating a Sentiment Dictionary for German Political Language - a Workbench Note”. In: *Journal of Information Technology and Politics* 15.4, pp. 319–343. DOI: 10.1080/19331681.2018.1485608.

Rehurek, Radim and Petr Sojka (2010). “Software Framework for Topic Modelling with Large Corpora”. In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pp. 45–50.

Reid, Rebecca and Kirk A. Randazzo (2016). “Statutory Language and the Separation of Powers”. In: *Justice System Journal* 37.3, pp. 246–258. DOI: 10.1080/0098261X.2015.1024569.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin (2016). “‘Why Should I Trust You?” Explaining the Predictions of Any Classifier”. In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, pp. 1135–1144. DOI: 10.1145/2939672.2939778.

Rifkin, Ryan, Gene Yeo, and Tomaso Poggio (2003). “Regularized Least-Squares Classification”. In: *Nato Science Series Sub Series III Computer and Systems Sciences* 190, pp. 131–154.

Rusic, Milos, Malte Pietsch, and Timo Möller (2020). *Deepset*. URL: https://deepset.ai (visited on 07/15/2020).

Sala-i-Martin, Xavier X. (1997). *I Just Ran Four Million Regressions*. Tech. rep.

Sardá-Espinosa, Alexis (2017). “Comparing Time-Series Clustering Algorithms in R using the dtwclust Package”. In: *R Package Vignette* 12, p. 41.

Schmidt, Mark, Nicolas Le Roux, and Francis Bach (2017). “Minimizing Finite Sums with the Stochastic Average Gradient”. In: *Mathematical Programming* 162.1-2, pp. 83–112. DOI: 10.1007/s10107-016-1030-6.

Segal, Jeffrey A. et al. (1995). “Ideological Values and the Votes of U.S. Supreme Court Justices Revisited”. In: *The Journal of Politics* 57.3, pp. 812–823. DOI: 10.2307/2960194.

Shalvi, Shaul, Ori Eldar, and Yoella Bereby-Meyer (2012). “Honesty Requires Time (and Lack of Justifications)”. In: *Psychological Science* 23.10, pp. 1264–1270. DOI: 10.1177/0956797612443835.

Shrestha, Prasha et al. (2017). “Convolutional Neural Networks for Authorship Attribution of Short Texts”. In: *15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017 - Proceedings of Conference*. Vol. 2, pp. 669–674. DOI: 10.18653/v1/e17-2106.

Sidorov, Grigori et al. (2014). “Syntactic N-Grams as Machine Learning Features for Natural Language Processing”. In: *Expert Systems with Applications* 41.3, pp. 853–860. DOI: 10.1016/j.eswa.2013.08.015.

Slapin, Jonathan B. and Sven Oliver Proksch (2008). “A Scaling Model for Estimating Time-Series Party Positions from Texts”. In: *American Journal of Political Science* 52.3, pp. 705–722. DOI: 10.1111/j.1540-5907.2008.00338.x.

Socher, Richard et al. (2013). “Reasoning with Neural Tensor Networks for Knowledge Base Completion”. In: *Advances in Neural Information Processing Systems*, pp. 926–934.

Songer, Donald R. (1993). *The United States Court of Appeals Database – Documentation for Phase I.*

Straube, Sirko and Mario M. Krell (2014). “How to Evaluate an Agent’s Behavior to Infrequent Events? - Reliable Performance Estimation Insensitive to Class Distribution”. In: *Frontiers in Computational Neuroscience* 8, p. 43.

Sturm, H. P. and C. Herman Pritchett (1949). “The Roosevelt Court, a Study in Judicial Politics and Values, 1937–1947”. In: *The Western Political Quarterly* 2.3, p. 465. DOI: 10.2307/442080.

Suen, Ching Y. (1979). “N-Gram Statistics for Natural Language Understanding and Text Processing”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-1.2, pp. 164–172. DOI: 10.1109/TPAMI.1979.4766902.

Sulea, Octavia-Maria et al. (2017). “Exploring the Use of Text Classification in the Legal Domain”. In: *arXiv preprint arXiv:1710.09306*.

Sun, Shiliang and Qiaona Chen (2011). “Hierarchical Distance Metric Learning for Large Margin Nearest Neighbor Classification”. In: *International Journal of Pattern Recognition and Artificial Intelligence* 25.7, pp. 1073–1087. DOI: 10.1142/S021800141100897X.

Tabatabaeian, Maryam, Rick Dale, and Nicholas D. Duran (2015). “Self-Serving Dishonest Decisions can show Facilitated Cognitive Dynamics”. In: *Cognitive Processing* 16.3, pp. 291–300. DOI: 10.1007/s10339-015-0660-6.

Tang, Grace (2012). “White Lies”. In: *Nature* 487.7407, p. 400. DOI: 10.1038/487400a.

Undavia, Samir, Adam Meyers, and John E. Ortega (2018). “A Comparative Study of Classifying Legal Documents with Neural Networks”. In: *Federated Conference on Computer Science and Information Systems (FedCSIS)*, pp. 515–522. DOI: 10.15439/2018F227.

Van Bockstaele, Bram et al. (2012). “Learning to Lie: Effects of Practice on the Cognitive Cost of Lying”. In: *Frontiers in Psychology* 3, p. 526. DOI: 10.3389/fpsyg.2012.00526.

Van’t Veer, Anna E., Mariëlle Stel, and Ilja van Beest (2014). “Limited Capacity to Lie: Cognitive Load Interferes with being Dishonest”. In: *Judgment and Decision Making* 9.3, pp. 199–206. DOI: 10.2139/ssrn.2351377.

Vaswani, Ashish et al. (2017). “Attention is All You Need”. In: *Advances in Neural Information Processing Systems*. Vol. 2017-Decem, pp. 5999–6009.

Vorobeychik, Yevgeniy, Michael P. Wellman, and Satinder Singh (2007). “Learning Payoff Functions in Infinite Games”. In: *Machine Learning* 67.1-2, pp. 145–168.

Walczyk, Jeffrey J. et al. (2009). “Cognitive Lie Detection: Response Time and Consistency of Answers as Cues to Deception”. In: *Journal of Business and Psychology* 24.1, pp. 33–49. DOI: 10.1007/s10869-009-9090-8.

Wolpert, David H. (1992). “Stacked Generalization”. In: *Neural Networks* 5.2, pp. 241–259. DOI: 10.1016/S0893-6080(05)80023-1.

Zelmer, Jennifer (2003). "Linear Public Goods Experiments: A Meta-Analysis". In: *Experimental Economics* 6.3, pp. 299–310.

Zhang, Harry (2004). "The Optimality of Naive Bayes". In: *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference, FLAIRS 2004* 2, pp. 562–567.

# Carina Ines Hausladen

## Personal Data

| | |
|---|---|
| Name: | Carina Ines Hausladen |
| Nationality: | German |
| DoB: | 22. July 1993 |
| Phone: | +49 175 7481665 |
| email: | hausladen@wiso.uni-koeln.de |

## Academic Background

2017 – . Ph.D. Candidate, Department of Economics, University of Cologne, Cologne Germany / Max Planck Institute for Research on Collective Goods, Bonn, Germany.

2020 Visiting Research Fellow, Department of Experimental and Applied Psychology, Vrije Universiteit Amsterdam, Amsterdam, Netherlands, 3 months.

2019 Visiting Research Fellow, Department of Finance, Accounting, Controlling, and Taxation, Freie Universität Berlin, Berlin, Germany, 3 months.

2019 Visiting Research Fellow, Center for Law, Economics, and Data Science, Swiss Federal Institute of Technology (ETH), Zurich, Switzerland, 2 months.

2016-2017 Visiting Student Researcher, Psychophysiology Laboratory, Leland Stanford Junior University, Palo Alto, California, 6 months.

2015-2015 Research Assistant, Department of Economic Policy, University of Passau, Passau, Germany, 1 year. 2014 Student Research Intern, International Economics, Ifo Institute, Munich, Germany, 2 months.

## Education

2014-2017 M.Sc. Economics, University of Passau, Passau, Germany.

2011-2014 B.Sc. Business, University of Augsburg, Augsburg, Germany.

## Publications

Hausladen, C. I., Schubert, M. H., & Ash, E. (2020). Text Classification of ideological direction in judicial opinions. International Review of Law and Economics, 62, 105903.

doi: `10.1016/j.irle.2020.105903`

## Honors and Awards

| | |
|---|---|
| 2019 | C-SEB Startup Grant, University of Cologne, Cologne, Germany, 3000€. |
| 2019 | IPAK Travel Grant, DAAD & University of Cologne, Cologne, Germany, 1500€. |
| 2019 | Travel Grant, Empirical Legal Studies Replication Conference, Claremont, California, $500. |
| 2018 | C-SEB Gender Research Grant, University of Cologne, Cologne, Germany, 4000€. |
| 2016 | Heinz Sauermann Award, Gesellschaft für experimentelle Wirtschaftsforschung, Magdeburg, Germany, 1000€. |
| 2015 | Scholarship, German National Academic Foundation, Bonn, Germany, $\sim$ 15000€. |
| 2014 | Scholarship, Solidaris gGmbH, Munich, Germany, 500€. |
| 2013 | Scholarship, University of Augsburg, Augsburg, Germany, 3000€. |

## Conference Presentations and Invited Talks

Feb. 2020 Engel, C., Hausladen, C. I., Schubert, M. H., Identifying Theories about the Composition of the Type Space through Cluster Analysis of Linear Public Good Experiments, Amsterdam Cooperation Lab, Vrije Universiteit Amsterdam, Amsterdam, Netherlands.

May 2019 Hausladen, C. I., Nikolaychuk, O., Color me honest! Mousetracking, time-pressure, and (dis-)honest behavior. Sixth International Meeting on Experimental and Behavioral Social Sciences (IMEBESS), Utrecht University, Utrecht, Netherlands.

Apr. 2019 Hausladen, C. I., Schubert, M. H., & Ash, E. (2020). Text Classification of ideological direction in judicial opinions. PELS Replication Conference, Claremont McKenna College, Claremont, California.

Feb. 2019 Fochmann, M., Hausladen, C. I., Mohr, P., Classifying (dis-)honest decision making based on experimentally collected chat data.12th IMPRS Uncertainty Thesis Workshop, Wittenberg, Germany.

Sep. 2018 Hausladen, C. I., Nikolaychuk, O., Color me honest! Mousetracking, time-pressure, and (dis-)honest behavior. 6th Swiss Young Researchers Workshop in Behavioral Economics and Experimental Research, University of Neuchâtel, Neuchâtel, Switzerland.

## Teaching

2018-2019 Advanced Analytics and Applications, Lecture, Co-organizer and regular speaker, University of Cologne, Chair of sustainable energy and economics, Cologne, Germany.

2015-2016 Economics of Institutions, Tutorial, University of Passau, Chair of macroeconomics, Passau, Germany.

## Technical Skills

Programming Languages: R, Python, and LATEX.

Additional Software Skills: oTree, Gensim, NLTK, pandas, PostgreSQL, sklearn, and spaCy.

## Supervisors

Prof. Dr. Dr. h.c. Christoph Engel
Director of the Max Planck Institute for Research on Collective Goods
Kurt-Schumacher-Straße 10, 53113 Bonn
engel[at]coll.mpg.de

Prof. Dr. Martin Fochmann
Department of Finance, Accounting, Controlling and Taxationn
Thielallee 73, 14195 Berlin
martin.fochmann[at]fu-berlin.de

Pemfling, den 05. August 2020

Carina Ines Hausladen

“Hiermit versichere ich an Eides Statt, dass ich die vorgelegte Dissertation selbstständig und ohne die Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus anderen Quellen direkt oder indirekt übernommenen Aussagen, Daten und Konzepte sind unter Angabe der Quelle gekennzeichnet. Bei der Auswahl und Auswertung folgenden Materials haben mir die nachstehend aufgeführten Personen in der jeweils beschriebenen Weise entgeltlich/unentgeltlich (zutreffendes bitte unterstreichen) geholfen:

Weitere Personen neben den in der Einleitung der Dissertation aufgeführten Koautorinnen und Koautoren waren an der inhaltlich-materiellen Erstellung der vorliegenden Dissertation nicht beteiligt. Insbesondere habe ich hierfür nicht die entgeltliche Hilfe von Vermittlungs- bzw. Beratungsdiensten in Anspruch genommen. Niemand hat von mir unmittelbar oder mittelbar geldwerte Leistungen für Arbeiten erhalten, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen. Die Dissertation wurde bisher weder im In- noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt. Ich versichere, dass ich nach bestem Wissen die reine Wahrheit gesagt und nichts verschwiegen habe.”

Pemfling, den 05. August 2020

Carina Ines Hausladen