Essays in Behavioral Economics and Microeconomic Theory

D I S S E R T A T I O N

zur Erlangung des akademischen Grades
doctor rerum politicarum
(Doktor der Wirtschaftswissenschaft)

eingereicht an der

Wirtschaftswissenschaftlichen Fakultät
der Humboldt-Universität zu Berlin

von

Pauline Lisa Vorjohann, M.Sc.

Präsident der Humboldt-Universität zu Berlin:
Prof. Peter A. Frensch, PhD (komm.)

Dekan der Wirtschaftswissenschaftlichen Fakultät:
Prof. Dr. Daniel Klapper

Gutachter:
1. Prof. Georg Weizsäcker, Ph.D.
2. Prof. Dr. Steffen Huck

Tag des Kolloquiums: 13.7.2022

Zusammenfassung der Dissertationsergebnisse

Kapitel 1: Reference-dependent choice bracketing Im Rahmen des Erwartungsnutzenmodells leite ich ein theoretisches Modell von *choice bracketing* aus zwei verhaltensökonomischen Axiomen ab. Das erste Axiom etabliert einen direkten Zusammenhang zwischen *narrow bracketing* und *correlation neglect*. Das zweite Axiom identifiziert den Referenzpunkt als den Ort, an dem *broad* und *narrow* Präferenzen miteinander verbunden sind. In meinem Modell ist der *narrow bracketer* durch die Unfähigkeit, Veränderungen vom Referenzpunkt in unterschiedlichen Dimensionen gleichzeitig zu verarbeiten, charakterisiert. Mit einem Experiment demonstriere ich die empirische Testbarkeit meines Modells und präsentiere erste Beweise für seine Validität.

Kapitel 2: Welfare-based altruism Warum geben Menschen, wenn man sie fragt, präferieren aber, nicht gefragt zu werden, und nehmen sogar, wenn sich die Gelegenheit ergibt? Wir zeigen, dass Axiome wie Separabilität, *narrow bracketing*, und *scaling invariance* diese scheinbar widersprüchlichen Beobachtungen vorhersagen. Insbesondere implizieren diese Axiome, dass die Interdependenz von Präferenzen ("Altruismus") ein Ergebnis des Interesses für das Wohlbefinden anderer im Gegensatz zu ihren bloßen Auszahlungen ist. Hierbei wird das Wohlbefinden durch die referenzabhängige Wertfunktion aus der Prospekttheorie erfasst. Unser Modell erlaubt es uns, konsistente Vorhersagen von Entscheidungen aus einflussreichen Experimenten über eine Vielzahl von Verteilungssituationen hinweg zu treffen.

Kapitel 3: Fake news and information transmission Wir untersuchen, wie sich *fake news* auf den Informationsfluss zwischen Nachrichtenportalen und ökonomischen Agenten auswirkt. Wir erweitern das klassische *cheap-talk*-Modell um Unsicherheit über die Präferenzen des *sender* (Nachrichtenportal). Es gibt zwei Typen von Nachrichtenportalen. Ein *fake-news*-Portal möchte im Agenten unabhängig vom wahren Zustand eine maximale Erwartung wecken. Ein legitimes Nachrichtenportal möchte die Wahrheit offenbaren. Wir zeigen, dass jedes informative perfekte Bayesianische Gleichgewicht durch einen Schwellenwert charakterisiert ist. Während der Agent alle Zustände unter dem Schwellenwert unterscheiden kann, ist es ihm unmöglich, Zustände über dem Schwellenwert zu unterscheiden.

Summary of dissertation results

Chapter 1: Reference-dependent choice bracketing I derive a theoretical model of choice bracketing from two behavioral axioms in an expected utility framework. The first behavioral axiom establishes a direct link between narrow bracketing and correlation neglect. The second behavioral axiom identifies the reference point as the place where broad and narrow preferences are connected. In my model, the narrow bracketer is characterized by an inability to process changes from the reference point in different dimensions simultaneously. I present an experiment which demonstrates the empirical testability of my model and provides preliminary evidence in support of its validity.

Chapter 2: Welfare-based altruism Why do people give when asked, but prefer not to be asked, and even take when possible? We show that standard behavioral axioms including separability, narrow bracketing, and scaling invariance predict these seemingly inconsistent observations. Specifically, these axioms imply that interdependence of preferences ("altruism") results from concerns for the welfare of others, as opposed to their mere payoffs, where individual welfares are captured by the reference-dependent value functions known from prospect theory. The resulting preferences are non-convex, which captures giving, sorting, and taking directly. This allows us to consistently predict choices across seminal experiments covering distributive decisions in many contexts.

Chapter 3: Fake news and information transmission We present a theoretical model to investigate how the presence of fake news affects information transmission from media outlets to economic agents. In a standard cheap talk framework we introduce uncertainty about the sender's (media outlet's) preferences. There are two types of media outlets. A fake news outlet wants to push the agent's belief to the maximum irrespective of the state of the world. A legitimate outlet wants to reveal the true state to the agent. We show that any informative perfect Bayesian equilibrium of our game is characterized by a threshold value. While the agent can perfectly separate amongst states below the threshold value, there is no separation amongst states above the threshold value. We determine the unique most informative threshold value for a general class of equilibria.

Acknowledgements

I am grateful to my advisers Georg Weizsäcker and Steffen Huck for their invaluable support at every step of the way towards this dissertation. Their enthusiasm, trust, and advice made all the difference and I consider myself very lucky to have been able to learn from and work with them for the past few years.

Yves Breitmoser deserves a special thank you as well. For initiating and nurturing my excitement about economics from the very beginning of my studies. His skill, patience, and good nature make him a perfect coauthor and mentor to me.

I want to thank my colleagues, in particular Philipp Albert, Kai Barron, Johannes Leutgeb, Thibaud Pierrot, and Roel van Veldhuizen who always took the time for inspiring conversations about research and beyond during countless coffee breaks. Their openness, interest, and advice kept me going whenever I was stuck.

Furthermore, I would like to express my gratitude to the institutions and programs that supported my doctoral studies. Many thanks go to my two academic homes for the past years, the Microecnomics Research Group at Humboldt Universität zu Berlin and the Wissenschaftszentrum Berlin für Sozialforschung (WZB). These insitutions and in particular the people that comprise them provided the perfect environment and support system for me to grow both as a researcher and as a person. I am also very grateful to have been a member of the Berlin Behavioral Economics Group, the Collaborative Research Center Transregio "Rationality and Competition" (CRC TRR 190), and the Berlin Doctoral Program in Economics and Management Science (now Berlin School of Economics).

Finally, I am indebted to my family and friends. Especially my partner Jan Henner, my parents Ute John and Walter Vorjohann, and my close friend Laetitia Lenel. Their love and support is what made this dissertation and everything else possible for me.

Introduction

In this dissertation, I study how human behavior is affected by its economic environment. Oftentimes, slight changes in the rules that govern the outcomes of our decisions or their mere framing can have a large impact on the decisions we make. Whether we manage to avoid mistakes in complex investment choices depends on how easy it is to keep an overview of the vast amount of available financial products and their properties. Whether we regard a distribution of money between ourselves and another person as fair is largely determined by how that distribution came about. Whether we trust the news we read is influenced by what we know about the strategic incentives of the news outlets that produce them.

Each of these three examples alludes to a different channel through which decision-making is influenced by the economic environment. First, humans make mistakes. Living in an overwhelmingly complex world, our capacity and willingness to deliberate our decisions down to the smallest detail is limited and shaped by how they are framed. Second, humans are social beings. We do not only care about our own monetary payoffs but derive satisfaction from achieving distributions that we and others regard as fair with our concept of fairness being inherently context-dependent. Third, strategic considerations change as the motivations of the people we interact with change. The information that we can infer from another person's behavior depends on what we know about her incentives. In each of the three chapters of this dissertation, I use a theoretical approach to analyze one of these three channels. In the first two chapters, I additionally use experimental data to test the assumptions and predictions of my models.

In the first chapter, I focus on a specific systematic mistake in human decision making. I investigate the human tendency of isolating individual decisions from one another to simplify an overall interdependent decision problem. This tendency is referred to as narrow bracketing. Empirical evidence has shown that narrow bracketing adversely affects behavior in a large variety of important economic settings including, for example, labor supply decisions, investment decisions, and consumption decisions. I present a theoretical model of narrow bracketing. The model is derived from two basic behavioral axioms in the context of expected utility. Additionally, I present an experiment that demonstrates the empirical testability of the model and

provides first evidence in support of my behavioral axioms.

A distinctive feature of my model is that a narrow bracketer behaves as if her preferences were context-dependent. In particular, her decisions over complex multi-dimensional objects are influenced by what I call her reference point. In the spirit of Kahneman and Tversky (1979), the reference point can be thought of as the status quo or an outcome that the narrow bracketer is used to. While the narrow bracketer is comfortably able to keep an overview of a complex multi-dimensional outcome as long as it coincides with her reference point, she is only able to keep track of how another outcome departs from her reference point in one dimension at a time. As a result, her decisions are influenced by her reference point which may change as the context or presentation of her decision environment changes.

In the second chapter, which is based on joint work with Yves Breitmoser, we look at altruistic preferences. Evidence from standard dictator games suggests that people have stable preferences for giving. Still, existing models of social preferences have been unable to account for the drastic changes in giving behavior observed in slight variations of that game including, for example, taking and sorting games. Meanwhile, the large literature on real-world charitable giving converged to a very similar set of puzzles. We propose a novel axiomatic approach to analyzing social preferences. Starting from general behavioral principles, we derive a preference representation that reconciles the seemingly contradictory evidence on general distribution games and charitable giving. Distinctively, our model characterizes altruism as a concern for the welfare of others as opposed to their mere payoffs. We complement our theoretical analysis by re-analyzing existing experimental data showing that our model reliably predicts giving behavior across different contexts.

Although in this chapter we study an inherently different behavioral phenomenon, there is a very close connection to the first chapter. As in my model of narrow bracketing, the preferences of what we call a welfare-based altruist are reference-dependent. Indeed, the central axiom that distinguishes our model of welfare-based altruism from standard payoff-based altruism states that the decision-maker engages in a form of narrow bracketing. The resulting reference-dependence of the welfare-based altruist's preferences directly implies that she may judge the same monetary distribution differently depending on how it was generated. As the rules of a distribution game change, the reference point with respect to which outcomes are evaluated changes which in

turn affects the associated welfares of the involved parties.

In the third chapter, which is based on joint work with Steffen Huck, we leave the realm of individual decision-making and consider how strategic behavior is affected by changes in the economic environment. In particular, we investigate how the information flow from media outlets to economic agents is impaired by the presence of fake news. In recent years, many societies have experienced a fundamental shift in the prevalence of fake news accompanied by a growing concern about the effects of this shift on the general public's trust in the media. We analyze a theoretical model of the information transmission between a potentially biased media outlet and an economic agent. Our analysis reveals that the presence of fake news can have a substantial negative impact on the possibilities for information transmission from legitimate news outlets to economic agents.

Other than the first two chapters, the third chapter demonstrates that strong responses in behavior to slight changes in the economic environment do not only occur when the decision-maker is prone to make systematic mistakes or when she has non-standard preferences. We show that the mere chance of encountering a fake news outlet even if very small has the potential to erode the trust that rational economic agents place in the media to an alarming extent.

Contents

List of Figures

List of Tables

1 Reference-dependent choice bracketing

1.1 Abstract

I derive a theoretical model of choice bracketing from two behavioral axioms in an expected utility framework. The first behavioral axiom establishes a direct link between narrow bracketing and correlation neglect. The second behavioral axiom identifies the reference point as the place where broad and narrow preferences are connected. In my model, the narrow bracketer is characterized by an inability to process changes from the reference point in different dimensions simultaneously. As a result, her tradeoffs between dimensions are distorted. While she disregards interactions between actual outcomes, she appreciates these interactions mistakenly with respect to the reference point. In addition to the theoretical contribution, I present an experiment which demonstrates the empirical testability of my model and provides preliminary evidence in support of its validity.

1.2 Introduction

The amount of decisions that we face and the interdependencies between all of these decisions force us to apply a simplified view of the world. We isolate decisions from one another to be able to make them at all. Following Read et al. (1999b) this mental procedure is referred to as choice bracketing. A decision maker who assesses all of her decisions jointly to find the optimal combination is referred to as a broad bracketer. A narrow bracketer takes some or all of her decisions in isolation, disregarding their interdependencies. As a result, the combination of decisions that a narrow bracketer makes is rarely optimal.

I present a theoretical model of choice bracketing. The model is derived from a choice-theoretic foundation in the context of expected utility. My model is applicable to a large variety of economic settings. In particular, it is the first theoretical model of choice bracketing that allows for multidimensional outcomes. Consequently, my model opens up the possibility to study the effects of narrow bracketing in many important economic settings ranging from basic consumption basket choice to complex multiattribute negotiations. Furthermore, I resolve the general incompatibility of narrow bracketing and

budget balance. Finally, my model enables me to derive meaningful predictions for the behavior of a narrow bracketer who is not loss-averse at the same time, isolating the two behavioral biases from one antoher. In addition to the theoretical contribution, I present an experiment which demonstrates the testability of my model and provides preliminary evidence in support of its validity.

Empirical and experimental evidence suggests that narrow bracketing affects behavior in many important economic settings including, for example, labor supply decisions (Fallucchi and Kaufmann, 2021; Camerer et al., 1997), investment decisions (Kumar and Lim, 2008; Thaler et al., 1997; Gneezy and Potters, 1997), trade between agents (Kahneman et al., 1990), retirement savings decisions (Choi et al., 2009; Brown et al., 2008), consumption decisions (Abeler and Marklein, 2017; Read and Loewenstein, 1995), decisions under risk (Rabin and Weizsäcker, 2009), and intertemporal decisions (Koch and Nafziger, 2020; Andreoni et al., 2018; Read et al., 1999a). The prevalence of narrow bracketing is demonstrated by Ellis and Freeman (2020). Across three different contexts they find that only $0-15\%$ of subjects in their experiment are consistent with broad bracketing while $40-44\%$ of subjects are consistent with narrow bracketing. Furthermore, Mu et al. (2020) show that under the assumption of broad bracketing the principle of stochastic dominance is incompatible with the common observation that decision makers are risk-averse over small gambles, providing a theoretical argument for the importance of accounting for narrow bracketing when modeling decision making under risk.

Despite the ample evidence of both prevalence and relevance of narrow bracketing, we still lack a generally applicable theoretical model of this important behavioral bias. Providing such a model is the main contribution of my paper.

A decision-maker (DM) faces a series of intermediate decisions. Together, these intermediate decisions comprise the prospect she receives. A prospect is a probability distribution on a multidimensional outcome set. Each prospect is decomposed into several subprospects representing the intermediate decisions. There is one subprospect for each dimension of the outcome set. The subprospect corresponding to a given dimension of the outcome set is the marginal distribution on that dimension induced by the prospect it comprises.

DM is characterized by two preference relations on prospects. Her *broad preference relation* captures her true preferences. If DM brackets broadly,

she makes choices in line with her broad preference relation. If DM brackets narrowly, her choices are governed by her *narrow preference relation* instead. DM's narrow preference relation is characterized by a *system of brackets*. The system of brackets partitions the subprospects comprising an overall prospect into distinct groups (brackets). I take the system of brackets as given.[1] It determines the degree to which DM brackets narrowly. While a fully narrow DM puts each subprospect into a distinct bracket, a fully broad DM has only one bracket including all subprospects that comprise the overall prospect.

I derive a representation for DM's narrow preference relation from her broad preference relation and two behavioral axioms. I do so in the framework of expected utility. My first behavioral axiom specifies the mistake that a narrow bracketer makes. It identifies *correlation neglect*[2] as the central flaw of narrow decision making. A narrow DM considers the subprospects inside a given bracket in isolation, disregarding all subprospects outside of that bracket. Of course, if these other subprospects are entirely independent of the considered subprospects, there is no harm done in disregarding them. If, however, these other subprospects are correlated with the considered subprospects or there are important interdependencies between the subprospect outcomes, disregarding them becomes a problem.

My second behavioral axiom ties the narrow preference relation to its broad couterpart. The broad and narrow preference relations belong to one and the same DM. While the one captures DM's true preferences, the other captures the choices she makes. Therefore, the narrow preference relation may depart from the broad preference relation only if that departure can be rationalized by DM's bracketing behavior. In principle a narrow bracketer disregards all interdependencies between subprospects across brackets. However, I assert that the narrow bracketer is not entirely ignorant with respect to these across-bracket interdependencies. I assume that there exists a specific outcome, the *reference point*[3], at which she retains her ability to process all brackets simultaneously. Therefore, the narrow preference relation agrees with the broad preference relation for any two prospects that differ from each

[1] For models of endogeneous bracket formation in the context of intertemporal decision making see, e.g., Galperti (2019); Hsiaw (2018); Koch and Nafziger (2016). Relatedly, Kőszegi and Matějka (2020) present a model of how people form mental budgets.

[2] For related papers on correlation neglect see, e.g., Enke and Zimmermann (2018); Ellis and Piccione (2017); Eyster and Weizsäcker (2016).

[3] The concept of a reference point was introduced by Kahneman and Tversky (1979) in the context of prospect theory.

other and the reference point in at most one bracket. Intuitively, the reference point captures an outcome that DM is used to and therefore comfortably able to keep the overview of.

The derived expected utility representation of the narrow preference relation is additively separable across brackets. The narrow bracketer's expected utility from a given prospect can be decomposed into a sum of expected utilities from its bracketwise subprospects. Additive separability is implied by my correlation neglect axiom. To establish it I apply a theorem derived by Fishburn (1967) in the framework of multiattribute utility theory to my setting. The axiom that ties the narrow preference relation to its broad counterpart via the reference point imposes further structure on the narrow bracketer's bracketwise expected utilities. For a given bracket the expected utility function of the narrow bracketer is equivalent to the broad bracketer's expected utility function with all outside-bracket outcomes fixed at the reference point.

My representation theorem reveals that when evaluating a prospect, the narrow bracketer can be modeled as using the same expected utility function as the broad bracketer. However, she applies that expected utility function separately to each bracket in her system of brackets. For each bracket, she evaluates the broad expected utility function at the subprospects inside that bracket while keeping all other subprospects fixed at the reference point. Finally, she takes the sum of all of these bracketwise expected utilities. As a result, the narrow bracketer disregards any interactions between subprospects across brackets. However, she appreciates these interactions mistakenly with respect to her reference point.

My model of choice bracketing is simple in the sense that the derived representation of the narrow preference relation can be treated in exactly the same way as any broad expected utility representation. We can thus use the standard economics toolbox and the large body of existing results from microeconomic theory to study the choices of a narrow bracketer.

In particular, the model can be used in standard constrained (expected) utility maximization problems. One of the main obstacles towards formalizing the intuition of choice bracketing is that narrow bracketing is not readily compatible with the principle of budget balance. While narrow bracketing is associated with a decision maker's inability to think multidimensionally, budget balance requires her to make tradeoffs between dimensions. My model resolves this incompatiblity of narrow bracketing and budget balance by in-

troducing the reference point. At the reference point, the narrow bracketer retains her ability to think multidimensionally. However, since she is unable to process changes from the reference point in different dimensions simultaneously, her tradeoffs between dimensions are distorted.

Existing experiments on choice bracketing circumvent dealing with the general incompatibility of narrow bracketing and budget balance by design (see, e.g., Ellis and Freeman, 2020; Rabin and Weizsäcker, 2009; Tversky and Kahneman, 1981). They restrict attention to settings where the intermediate decisions that comprise an overall decision are not connected via a budget constraint. Then, a subject's choice in one intermediate decision has no influence on the choices available to her in any other intermediate decision.

Barberis and Huang (2009) and Barberis et al. (2001) present theoretical models of choice bracketing that remedy the incompatibility of narrow bracketing and budget balance by assuming that the narrow bracketer evaluates her utility function separately for each decision she takes and then maximizes the sum over all these individually evaluated utilities. Their model has been used to study choice bracketing in economic applications including portfolio choice (Barberis and Huang, 2009; Barberis et al., 2006; Benartzi and Thaler, 1995), asset pricing (Barberis and Huang, 2001; Barberis et al., 2001), and self-control problems (Koch and Nafziger, 2016; Hsiaw, 2018). My model of choice bracketing contributes to this literature in three respects. First, it provides a choice theoretic foundation for the additive formulation of Barberis and Huang (2009). Second, it extends the set of possible applications considerably by allowing for multidimensional outcomes. Third, by explicitly modeling the system of brackets, it allows for more subtle forms of partial narrow bracketing.

Barberis and Huang (2009) and Barberis et al. (2006) capture partial narrow bracketing through a global-plus-local utility function. This means that they model the partial narrow bracketer as evaluating the weighted sum of broad and fully narrow utility such that the weight attached to the fully narrow utility measures the extent of narrow bracketing. This formulation obviously has the advantage of being more simple than my approach. However, this simplicity comes at a cost. It blurs the very basic intution of choice *bracketing* and does not allow for investigations of the effects that a change in the system of brackets has on the behavior of a narrow bracketer. Furthermore, experimental results of Ellis and Freeman (2020) suggest that the costs of simplicity

as imposed by the global-plus-local formulation may outweigh its benefits.

My model of choice bracketing reveals a tight relation between narrow bracketing and budgeting, which besides narrow bracketing is another important aspect of mental accounting as outlined by Thaler (1999). It is intuitively appealing to think of a consumer who chooses a complex consumption bundle as following a two-stage procedure (Gilboa et al., 2010). In the first stage, the budgeting stage, she optimally distributes her budget across general categories of goods like clothing, food, and entertainment. In the second stage, she decides separately for each good category how to allocate her category budget from the first stage across the individual goods belonging to that category. Such a budgeting procedure is generally admissibile if and only if the utility function is additively separable across good categories (Gorman, 1959; Strotz, 1957, 1959). Thus, akin to Blow and Crawford (2018)'s definition of boundedly rational mental accounting, additive separability of the narrow preference representation implies that a narrow bracketer can be interpreted as using the described budgeting procedure although her broad preferences do not allow it.

To demonstrate the effects that narrow bracketing has on behavior in basic economic settings, I apply my model to the economics 101 consumer's constrained utility maximization problem with two goods. Additive separability of the narrow preference representation implies that any interactions between the two goods in her bundle are disregarded by the narrow bracketer. This disregard is nicely illustrated by the shape of the narrow indifference curves in comparison to their broad counterparts. If the goods have negative interactions akin to substitutabilities, the narrow indifference curves are more convex than their broad counterparts. If the goods have positive interactions akin to complementarities, the narrow indifference curves are less convex than their broad counterparts. Intuitively, the more convex an indifference curve, the more complementary are the two goods. Thus, a narrow bracketer regards two substitutable goods as more complementary than they actually are and vice versa for two complementary goods.

However, while disregarding interactions for the consumption bundle she receives, the narrow bracketer is not fully ignorant of their existence. She mistakenly appreciates the interactions separately for each good dimension of her bundle with respect to her reference point. The narrow bracketer does not consider changes from the reference point in the two good dimensions

simultaneously. Thus, when thinking about an alteration of her bundle away from the reference point in one good dimension, she keeps the respective other good dimension fixed at its reference point level. As a result, the tradeoffs she makes between the two good dimensions are distorted.

For illustration, suppose the two goods have positive interactions and the reference point is unbalanced towards the first good dimension. The higher reference point in the first good dimension implies that increases in the second good dimension are percieved by the narrow bracketer as more attractive than they actually are. At the same time, the lower reference point in the second good dimension makes increases in the first good dimension seem less attractive than they actually are. The narrow bracketer's mistaken attribution of interactions to the respective reference point levels instead of the amounts in her actual bundle move her optimum away from the reference point. In contrast, if the two goods have negative interactions, the narrow bracketer's chosen bundle is closer to the reference point than her optimal bundle.

I also study the implications of choice bracketing in an Edgeworth-box exchange economy assuming status-quo reference points. I find that, starting from any initial endowment structure, in the case of positive interactions the volume of trade is higher if the trading parties bracket narrowly. In contrast, in the case of negative interactions narrow bracketing results in a lower volume of trade. This result has important implications for how the procedures of negotiations affect their outcomes. Especially, it calls into question the general practice of splitting up multidimensional negotiations, negotiating every aspect of a deal separately, since this might induce the involved parties to bracket narrowly.

A recent related literature shows how a consumer's limited attention to price or preference shocks provokes behavior akin to the narrow consumer's behavior in my model. For different definitions of limited attention, papers by Kőszegi and Matějka (2020), Lian (2020), and Gabaix (2014) show that in reaction to such a shock in one good dimension, the inattentive consumer behaves as if she (partially) disregards interactions of that good with the other goods in her bundle. I model narrow bracketing directly. Therefore, in contrast to models based on limited attention my model has bite also in settings with perfect information on prices and preferences. Indeed, experimental evidence suggests that narrow bracketing readily occurs even in such deterministic settings (see, e.g., Ellis and Freeman, 2020; Rabin and Weizsäcker, 2009).

Finally, I present the results of an online laboratory experiment. The main goal of the experiment is to demonstrate the empirical testability of my model. On a secondary note, my experimental results provide preliminary evidence for the validity of my model. I show how to construct an experimental design that can test both the validity of my behavioral axioms and my model's predictions on the role of the reference point in narrow decision making. I compare behavior within subject in equivalent two-dimensional decision problems across two treatments. In the *braod treatment* subjects can access information on both dimensions of the decision problem jointly. In the *narrow treatment* subjects can access information on the two dimensions of the decision problem only separately. Furthermore, I impose a waiting time in-between accessing the information on each of the two dimensions of the decision problem that makes switching between the information on the dimensions costly.

A decision problem in the experiment is a multiple choice list between a portfolio and an increasing certain payment. I use the multiple choice list to elicit subjects' willingness to pay (WTP) for the portfolio. Each portfolio consists of two assets, a blue asset and an orange asset. The assets yield blue and orange point earnings respectively depending on the toss of a coin. Payments are determined by the combination of blue and orange point earnings. The payment rule induces interactions between blue and orange points to make the problem interesting in the context of my model.

To gain control over the reference point that subjects use in my experiment I introduce a base-portfolio. The base-portfolio is deterministic and kept constant over the course of the experiment. Every decision in the experiment is implemented with probability 0.5. If a decision is not implemented, the subject receives the base-portfolio instead. This approach of influencing the reference point that subjects use is inspired by Abeler et al. (2011).

The portfolios for which I elicit WTP are chosen such that my behavioral axioms and model predictions can be tested by comparing the WTP differences for pairs of portfolios across the two treatments. Despite observing a relatively small treatment effect, I find support for my model of choice bracketing. The experimental evidence is partially in line with my correlation neglect axiom. Furthermore, my second behavioral axiom on the connection of broad and narrow preferences via the reference point is fully supported. However, I do not find support for my model prediction on the role of the reference point. Overall, the results of my experiment serve as preliminary evidence for

the validity of my model. More generally, the experiment demonstrates that my model of choice bracketing is empirically testable and provides a guideline for future experimental investigations, possibly amplifying the suggested treatment variation to induce a larger treatment effect.

The main respect in which my experiment departs from the experimental literature studying narrow bracketing is the treatment design. Existing apporaches to experimentally separate broad from narrow bracketing can be roughly categorized into two groups. First, broad and narrow treatments differ in whether subjects make decisions simultaneously or sequentially (see, e.g., Rabin and Weizsäcker, 2009; Read et al., 2001, 1999a). Second, broad and narrow treatments differ in whether subjects' rewards from their decisions are aggregated or separated (see, e.g., Koch and Nafziger, 2020; Stracke et al., 2017; Gneezy and Potters, 1997).[4] I employ a treatment variation that allows for a direct test of my behavioral axioms. Instead of fully isolating the two dimensions of the decision problem in the narrow treatment, I preserve its multidimensional nature across treatments. I only vary the ease at which subjects can jointly access information on the two dimensions of the decision problem. As a side effect, my treatment variation is less convoluted with other factors such as, for example, reduced complexity, time preferences, and presentation effects.

1.3 The model

1.3.1 Theoretical framework

The outcome set X is a Cartesian product $\prod_{i\in I} X_i$. I is a finite set $\{1,2,...,n\}$ indexing the dimensions of an outcome $x \in X$. Let $\mathcal{P}$ denote the set of all finite discrete probability distributions on the set of all subsets of X. A *prospect* $P \in \mathcal{P}$ is a probability distribution over the multidimensional outcomes assigning to each outcome $x \in X$ its probability $P(x)$. If $P \in \mathcal{P}$, then $0 \leq P(x) \leq 1$ for all $x \in X$ and $\sum_{x\in X} P(x) = 1$.

The domain of preference is the set of all prospects. A decision maker (DM) is characterized by two preference relations on the set of prospects. Her

[4] Penczynski et al. (2020) present an interesting combination of the two approaches to understand the effects of decomposing simple games into components, each highlighting a different motive of the underlying game. While their design does not aim at directly separating broad from narrow bracketing, their treatment variation is quite closely related to the one implemented in my experiment.

broad preference relation $\succcurlyeq_b$ and her *narrow preference relation* $\succcurlyeq_n$. Consider prospects $P, Q \in \mathcal{P}$. $[P \succcurlyeq_b Q]$ indicates that P is weakly preferred to Q according to $\succcurlyeq_b$. As usual, $[P \succ_b Q]$ indicates $[P \succcurlyeq_b Q$ and not $P \succcurlyeq_b Q]$ while $[P \sim_b Q]$ indicates $[P \succcurlyeq_b Q$ and $P \succcurlyeq_b Q]$. The indications apply analoguously to $\succcurlyeq_n$.

I interpret DM's broad preference relation as capturing her true preferences in the sense that if she brackets broadly, her choices are in line with $\succcurlyeq_b$. If DM brackets narrowly, her choices may not be in line with her true preferences. I interpret $\succcurlyeq_n$ as the preference relation that governs the narrow DM's choices.

Assumption 1 (Richness)**.** Every probability distribution over outcomes that takes only finitely many values is available in the preference domains of $\succcurlyeq_b$ and $\succcurlyeq_n$.

So far, my theoretical framework closely follows the literature on multiattribute utility theory (see e.g. Keeney and Raiffa, 1993; Fishburn, 1965, 1967). To accomodate the idea of choice bracketing I now carry the multiattribute nature of outcomes over to the prospect that generates them.

Let $\mathcal{P}_i$ be the set of all finite probability distributions on X_i. For every prospect $P \in \mathcal{P}$ there exists an element $P_i \in \mathcal{P}_i$ which is the marginal distribution on X_i induced by P. Refer to P_i as *subprospect* i of prospect P. Any prospect $P \in \mathcal{P}$ is thus associated with a collection of subprospects corresponding to its outcome dimensions, $(P_1, P_2, ..., P_n)$.

The decomposition of prospects into subprospects captures that a DM's overall decision for a specific prospect is the result of several intermediate decisions. In each intermediate decision DM chooses a subprospect. Taken together these subprospects then generate the overall prospect. In the multidimensional outcome arising from this prospect each dimension represents the outcome of one subprospect.

As long as DM brackets broadly, i.e. makes choices in line with $\succcurlyeq_b$, the above decomposition of prospects is redundant. A broad bracketer chooses the same prospect independent of whether this choice is the result of just one or several intermediate choices. A narrow bracketer, however, does not keep track of the interdependencies between all intermediate decisions. Therefore, a narrow bracketer's overall decision for a specific prospect depends on whether it is decomposed into subprospects or not.

In its most extreme form, narrow bracketing means that DM decides about each subprospect in isolation disregarding its interdependencies with any other subprospect she chooses. I allow for less extreme forms of narrow bracketing in which DM retains her ability to process subsets of her intermediate decisions jointly. Therefore, I define a *system of brackets* characterizing the narrow preference relation. The system of brackets partitions the collection of subprospects that generate the overall prospect into distinct groups (brackets).

The *system of brackets* B characterizing $\succcurlyeq_n$ is a set $\{B_1, B_2, ..., B_m\}$ of nonempty subsets of the outcome dimension index set I with $\bigcup_{j=1}^{m} B_j = I$. We refer to B_j as *bracket* j of the system of brackets B. Let $\mathcal{P}^j$ be the set of all finite discrete probability distributions on the set of all subsets of the outcome set in bracket B_j, $X^j := \prod_{i \in B_j} X_i$. For every prospect $P \in \mathcal{P}$ there exists an element $P^j \in \mathcal{P}^j$ which is the marginal distribution on X^j induced by P. We refer to P^j as the *jth bracket prospect* of P. Given a system of brackets B, each prospect P induces a collection of bracketwise prospects, $(P^1, P^2, ..., P^m)$, and each outcome x can be written as a collection of bracketwise outcomes, $x = (x^1, x^2, ..., x^m)$ where $x^j = (x_i)_{i \in B_j}$ for $j = 1, 2, ..., m$.

When a prospect $P \in \mathcal{P}$ is deterministic, i.e. $P(x) = 1$ for some $x \in X$, I refer to that prospect directly by its outcome x. Similarly, I refer to a deterministic subprospect $P_i \in \mathcal{P}_i$ by its outcome $x_i \in X_i$ and to a deterministic bracketwise prospect $P^j \in \mathcal{P}^j$ by its bracketwise outcome $x^j \in X^j$.

Given two prospects $P, Q \in \mathcal{P}$, denote by $(P^j, Q^{-j}) \in \mathcal{P}$ the prospect generated by combining the jth bracket prospect P^j of P with all but the jth bracket prospects of Q. Given two outcomes $x, y \in X$, denote by $(x^j, y^{-j}) \in X$ the outcome that combines the jth bracket outcome, x^j, in x with all but the jth bracket outcomes in y.

In gerneral, you can think of the multidimensional nature of outcomes in my framework in two ways. First, in line with what is normally thought of in the multiattribute utility literature, the outcomes of different subprospects may as such be qualitatively different from one another, naturally giving rise to a multiattribute formulation. For example, the overall outcome could be a consumption basket which is comprised of many individual goods, the different outcome dimensions, each of which was individually put into the basket by DM on her way through the supermarket.

Second, capturing the possibility of narrow bracketing in cases where outcomes do not have a multiattribute nature as such, I allow for a distinction

between outcome dimensions that are qualitatively the same but are the result of distinct intermediate decisions. For example, the overall outcome could be total money earnings from a portfolio comprised of the earnings from a collection of assets, the outcome dimensions, each of which was puchased individually by DM.

1.3.2 Axiomatic foundation

In the following I derive a utility representation for the narrow preference relation $\succcurlyeq_n$ from the broad preference relation $\succcurlyeq_b$. I do so in the framework of expected utility (EU), implicitly assuming that the axioms underlying the EU representation are fulfilled for each of the two preference relations $\succcurlyeq_b$ and $\succcurlyeq_n$.[5]

Assumption 2 (EU)**.**

(1) There exists a function $u : X \to \mathbb{R}$, the *broad utility function*, such that for all prospects $Q, R \in \mathcal{P}$, $Q \succcurlyeq_b R \Leftrightarrow EU(Q) \geq EU(R)$ with $EU(P) := \sum_{x \in X} P(x)u(x)$. u is unique up to positive affine transformation.

(2) There exists a function $\tilde{u} : X \to \mathbb{R}$, the *narrow utility function*, such that for all prospects $Q, R \in \mathcal{P}$, $Q \succcurlyeq_n R \Leftrightarrow \widetilde{EU}(Q) \geq \widetilde{EU}(R)$ with $\widetilde{EU}(P) := \sum_{x \in X} P(x)\tilde{u}(x)$. $\tilde{u}$ is unique up to positive affine transformation.

My approach for finding a utility representation of the narrow preference relation proceeds as follows. I ask myself two basic questions about the behavior of a narrow bracketer. The answers to these questions are captured in my two behavioral axioms. Together with Assumption 2 (EU) these two behavioral axioms determine the shape of the narrow bracketer's preference representation.

What is the narrow bracketer's mistake? First, I restrict my attention to the narrow preference relation. The following behavioral axiom clarifies what exactly it is that the narrow bracketer misses when choosing between two prospects.

Axiom 1 (correlation neglect)**.** For any two prospects $P, Q \in \mathcal{P}$, if all bracket-wise prospects induced by P and Q on the system of brackets B are the same, i.e. $P^j = Q^j$ for all $j \in \{1, 2, ..., m\}$, then $P \sim_n Q$.

[5] For axiomatizations of EU see, for example, Fishburn (1970) and Wakker (2010).

Axiom 1 states that the narrow bracketer is ignorant with respect to the correlation between the bracketwise prospects that comprise an overall prospect. When making a choice between two prospects, she only considers the individual bracketwise subprospects without keeping track of the overall prospects they comprise. Therefore, any two prospects that are comprised of the same subprospects, i.e. that induce the same marginal distributions on all bracketwise outcome sets, look exactly the same to her. This holds irrespective of whether the overall prospects, i.e. the joint distributions on the overall outcome set, are the same as well.

Of course, Axiom 1 only has bite in the sense that it harms the narrow bracketer, if there are meaningful interactions between the subprospects or their outcomes across brackets. Only then does the correlation structure of a prospect matter for the broad preference relation and only then does the correlation neglect axiom imply that the narrow preference relation deviates from its broad counterpart.

Axiom 1 is closely related to the concept of independence used in the multiattribute utility theory literature. In particular, Fishburn (1967) introduced an assumption equivalent to Axiom 1. I make heavy use of the results from that paper in the proof of my representation theorem. His assumption is a weaker version of mutual independence between the attributes of an outcome as defined in Fishburn (1965) allowing mutual independence to hold only between subsets of the attributes of an outcome.

Where are broad and narrow the same? Axiom 1 pins down the narrow bracketer's mistake. I now identify the instances in which the narrow bracketer's choice should not deviate from her true preferences. The following axiom considers the connection between the narrow preference relation and its broad counterpart.

Axiom 2 (Referene Point)**.** There exists an outcome $r \in X$, the *reference point*, such that for any two prospects $P, Q \in \mathcal{P}$, if the bracketwise prospects induced by P and Q differ from each other and r in at most one bracket, i.e. $P^j = Q^j = r^j$ for all but at most one $B_j \in \{B_1, B_2, ..., B_m\}$, then $P \succcurlyeq_b Q \Leftrightarrow P \succcurlyeq_n Q$.

Axiom 2 states that there exists an outcome, the reference point, which ties together broad and narrow preference relation. At the reference point the narrow bracketer is perfectly able to consider all brackets jointly. She

can properly process changes from the reference point as long as they only occur inside one bracket at a time. In a way, Axiom 2 tames the narrow preference relation. It allows for departures from the broad preference relation only if prospects differ from each other and the reference point in more than one bracket. The narrow bracketer is never fully ignorant of the existence of interactions between the subprospects across brackets since at and around the reference point she makes choices in line with her true preferences.

1.3.3 Representation theorem

I am now ready to state my representation theorem for the narrow preference relation.

Theorem 1 (Narrow Preference Representation)**.** *Under Assumptions 1 (Richness) and 2 (EU), Axioms 1 (Correlation neglect) and 2 (Reference point) hold if and only if for all prospects $P \in \mathcal{P}$ and corresponding bracketwise prospects $P^j \in \mathcal{P}^j$*

$$\widetilde{EU}(P) = \sum_{j=1}^{m} \widetilde{EU}_j(P^j) \quad \textit{with} \quad \widetilde{EU}_j(P^j) := \sum_{x^j \in X^j} P^j(x^j)\tilde{u}_j(x^j)$$

where $\tilde{u}_j : X^j \to \mathbb{R}$ for brackets $B_j \in \{B_1, B_2, ..., B_m\}$ are bracketwise utility functions with

$$\tilde{u}_j(x^j) := u(x^j, r^{-j}) \quad \forall x^j \in X^j \tag{1}$$

where $u(\cdot, r^{-j})$ denotes the broad utility function evaluated at the reference point for all brackets except bracket j, r^{-j}, which is treated as a fixed parameter of $\tilde{u}_j$.

Proof.

Step 1: The narrow utility function is additively separable across brackets. This result follows from Axiom 1 (Correlation neglect) using the results of Fishburn (1967). I restate his Theorem 1 translated to my framework:

Theorem (Fishburn, 1967)**.** *Under Assumptions 1 (Richness) and 2 (EU), Axiom 1 (Correlation neglect) holds if and only if there exist bracketwise utility*

functions $\tilde{u}_j : X^j \to \mathbb{R}$ *for all brackets* $B_j \in \{B_1, B_2, ..., B_m\}$ *such that*

$$\widetilde{EU}(P) = \sum_{j=1}^{m} \widetilde{EU}_j(P^j) \quad \textit{with} \quad \widetilde{EU}_j(P^j) := \sum_{x^j \in X^j} P^j(x^j)\tilde{u}_j(x^j)$$

for all prospects $P \in \mathcal{P}$ *and corresponding bracketwise prospects* P^j. $\widetilde{EU}$ *is unique up to positive affine transformation.*

Step 2: The jth bracket utility function corresponds to the broad utility function evaluated at the reference point outside of bracket j. This result follows from Axiom 2 (Reference point). Consider any two prospects $P, Q \in \mathcal{P}$ with correspoinding bracketwise prospects P^j, Q^j such that $P^j = Q^j = r^j$ for all but at most one $B_j \in \{B_1, B_2, ..., B_m\}$. Without loss of generality take $B_j = B_1$ as the bracket for which P^j, Q^j and r^j may differ. By Assumption 2 (EU) for the broad preference relation, $P \succcurlyeq_b Q$ if and only if $EU(P) \geq EU(Q)$. We can rewrite P and Q as (P^1, r^{-1}) and (Q^1, r^{-1}), obtaining $EU(P^1, r^{-1}) \geq EU(Q^1, r^{-1})$. We thus have

$$P \succcurlyeq_b Q \quad \Leftrightarrow \quad \sum_{x^1 \in X^1} P^1(x^1)u(x^1, r^{-1}) \geq \sum_{x^1 \in X^1} Q^1(x^1)u(x^1, r^{-1}). \tag{2}$$

Similarly, by Assumption 2 (EU) for the narrow preference relation, $P \succcurlyeq_n Q$ if and only if $\widetilde{EU}(P) \geq \widetilde{EU}(Q)$. Rewriting P and Q as above, we obtain

$$P \succcurlyeq_n Q \quad \Leftrightarrow \quad \sum_{x^1 \in X^1} P^1(x^1)\tilde{u}(x^1, r^{-1}) \geq \sum_{x^1 \in X^1} Q^1(x^1)\tilde{u}(x^1, r^{-1}).$$

Now, by Step 1 we can rewrite the above expression as

$$P \succcurlyeq_n Q \quad \Leftrightarrow \sum_{x^1 \in X^1} P^1(x^1)\tilde{u}_1(x^1) + \sum_{j=2}^{m} \tilde{u}_j(r^j) \geq \sum_{x^1 \in X^1} Q^1(x^1)\tilde{u}_1(x^1) + \sum_{j=2}^{m} \tilde{u}_j(r^j)$$

and simplify it to

$$P \succcurlyeq_n Q \quad \Leftrightarrow \sum_{x^1 \in X^1} P^1(x^1)\tilde{u}_1(x^1) \geq \sum_{x^1 \in X^1} Q^1(x^1)\tilde{u}_1(x^1). \tag{3}$$

Now, by Axiom 2 (Reference point) $P \succcurlyeq_b Q \Leftrightarrow P \succcurlyeq_n Q$. Combining expres-

sions 2 and 3 we therefore have

$$\sum_{x^1 \in X^1} P^1(x^1) u(x^1, r^{-1}) \geq \sum_{x^1 \in X^1} Q^1(x^1) u(x^1, r^{-1})$$
$$\Leftrightarrow \quad \sum_{x^1 \in X^1} P^1(x^1) \tilde{u}_1(x^1) \geq \sum_{x^1 \in X^1} Q^1(x^1) \tilde{u}_1(x^1).$$

The above statement requires $\tilde{u}_1$ to be a positive affine transformation of u evaluated at r^{-1}. Now, by Axiom 2 (Reference point) this requirement holds for all bracketwise utility functions in the sequence $\tilde{u}_1, \tilde{u}_2, ..., \tilde{u}_m$. Furthermore, by Theorem 2 of Fishburn (1967) a transformation of a bracketwise utility function $\tilde{u}_j$ cannot be performed individually, i.e. without appropriately transforming all other bracketwise utility functions in accordance with the admissible transformations of $\widetilde{EU}$.[6] □

The first part of Theorem 1 is essentially a restatement of Fishburn (1967)'s Theorem 1. Applied to my setting, his finding implies that under Assumptions 1 (Richness) and 2 (EU) Axiom 1 (Correlation neglect) holds if and only if the narrow utility function $\tilde{u}$ is additively separable across brackets. For each bracket B_j in the system of brackets characterizing the narrow preference relation, there exists a bracketwise utility function $\tilde{u}_j$, mapping the jth bracket outcome to the real numbers. The narrow utility function can be written as the sum of all bracketwise utility functions. This means that we can write the narrow expected utility of a prospect $P \in \mathcal{P}$ as a sum of bracketwise expect utilities from all bracketwise prospects P^j induced by P.

The important new insight of Theorem 1 is that the jth bracket utility function, $\tilde{u}_j$ is equivalent to the broad utility function keeping all outcomes except the jth bracket outcome fixed at the reference point. This means that we can interpret the narrow bracketer as actually using the same utility function she would use if she bracketed broadly. However, she applies that utility function separately to each bracket in her system of brackets. For a given bracket she evaluates her broad utility function at the outcomes inside that bracket while keeping all outside-bracket outcomes fixed at their reference point levels. Finally, her overall utility from a specific outcome is determined by the sum of all of these bracketwise evaluated utilities.

To illustrate the content of Theorem 1, consider the special case of $n = 2$

[6] For a detailed discussion of the admissible transformations on the sequence of functions $\widetilde{EU}_1, \widetilde{EU}_2, ..., \widetilde{EU}_m$ see Fishburn (1967).

such that every prospect consists of two subprospects and suppose the system of brackets characterizing $\succcurlyeq_n$ separates these two subprospects into distinct brackets. Consider any prospect $P \in \mathcal{P}$. The expected utility of the broad bracketer is given by

$$EU(P) = \sum_{x \in X} P(x)u(x). \tag{4}$$

Theorem 1 implies that the expected utility of the narrow bracketer can be expressed as

$$\begin{aligned} \widetilde{EU}(P) &= \sum_{x_1 \in X_1} P_1(x_1)u(x_1, r_2) + \sum_{x_2 \in X_2} P_2(x_2)u(r_1, x_2) \\ &= \sum_{x \in X} P(x)[u(x_1, r_2) + u(r_1, x_2)] \end{aligned} \tag{5}$$

with u equivalent across the two expected utility formulas.

The narrow bracketer's expected utility representation is an additively separable version of its broad counterpart. Consider the first formulation of $\widetilde{EU}(P)$ in (5) and compare it to the broad expected utility formula in (4). $\widetilde{EU}(P)$ is additively separable across brackets. It consists of the sum of two separate expected utility formulas, one evaluating the first subprospect P_1 and one evaluating the second subprospect P_2. This additive separability reflects the fact that any correlation between the two subprospects are disregarded by the narrow bracketer. By evaluating their expected utilities separately, she treats them as if they were entirely independent.

Furthermore, the narrow bracketer disregards any interactions between the outcomes of the two subprospects. This is nicely illustrated by the second formulation of $\widetilde{EU}(P)$ in (5). The utility that a narrow bracketer derives from an outcome x of the overall prospect P is, again, additively separable across brackets. Instead of evaluating the broad utility function at the overall outcome x as in (4), she evaluates the broad utility function separately for each bracket, once at the outcome of the first subprospect x_1 and once at the outcome of the second subprospect x_2. Since she never evaluates the broad utility at x_1 and x_2 jointly, she does not keep track of possible complementarities or substitutabilities between the two subprospect outcomes.

However, since the narrow bracketer uses the same utility function in her evaluation as the broad bracketer, she is never fully ignorant of the existence of interactions between the two outcome dimensions. She simply appreciates

these interactions mistakenly with respect to the reference point. When the narrow bracketer evaluates the outcome of the first subprospect x_1, she keeps the outcome of the second subprospect fixed at r_2 and vice versa. Thus, while she considers the interdependencies between x_1 and r_2 as well as the interdependencies between r_1 and x_2, she fails to keep track of the interdependencies between x_1 and x_2. As a result, her tradeoffs between the outcome dimensions are distorted.

1.3.4 Discussion

Budget balance A major obstacle towards modeling narrow bracketing is that there exists a tension between the behavioral bias and the economic principle of budget balance. Intuitively, narrow bracketing is associated with "...making each choice in isolation" (Read et al., 1999b). Adhering to this basic intution, one might be drawn to model the narrow bracketer as sequentially making each decision in a set of concurrent decisions as if it were the only decision she faces overall. Such a modeling approach works nicely when applied to the specific environments studied in large parts of the experimental literature on choice bracketing. These experiments are designed such that the specific option a decision maker chooses in one decision does not influence the set of options that are available to her in any other decision (see, e.g., Ellis and Freeman, 2020; Rabin and Weizsäcker, 2009; Tversky and Kahneman, 1981). However, the approach of modeling narrow bracketing as fully isolated decision making runs into serious problems when applied to economically more relevant settings in which decision makers face resource constraints which tie together the option sets of concurrent decisions.

For illustration consider the constrained utility maximization problem of a consumer who has a fixed budget to spend on food and clothing. Suppose the consumer narrowly brackets these two good categories. As long as her budget is tight enough, full isolation of her decisions in these two categories implies that the consumer spends her whole budget on either one of the two categories leaving nothing for the respective other category. Once she enters a, say, clothing store she fully ignores that she might also want to get dinner later on and therefore spends her whole budget on a new outfit. Only later, when she passes by her favourite restaurant she realizes how hugry she is. Of course, the irrationality displayed by the consumer's behavior in this example

is not what we observe in reality and goes far beyond what we actually think of when we talk about narrow bracketing.

The example demonstrates that a reasonable model of narrow bracketing needs to balance the isolated nature of narrow decision making with the integrated thinking required for making meaningful tradeoffs across brackets to satisfy budget balance. By defining the narrow preference relation on the same fully multidimensional prospects as the broad preference relation, I implicitly model the narrow bracketer's decision making as simultaneous. Therefore, my framework allows me to in principle cover the whole spectrum of isolation and integration in the narrow bracketer's decision making. Axiom 1 (Correlation neglect) imposes a limit on the ability of the narrow bracketer to integrate subprospects across brackets. This limit is balanced by Axiom 2 (Reference point) which retains the narrow bracketer's ability to integrate subprospects across brackets at and around the reference point. It is the combination of these two axioms that enables me to derive a representation of the narrow preference relation which captures the narrow bracketer's tendency to isolate intermediate decisions from one another and at the same time resolves the general incompatibility of this behavior with the principle of budget balance.

Mental accounting and budgeting Thaler (1999) defines mental accounting as "...the set of cognitive operations used by individuals and households to organize, evaluate, and keep track of financial activities". Choice bracketing is one component of such mental accounting. Another important component of mental accounting is *budgeting*. In the context of consumption choice budgeting describes the assignment of goods into categories with a fixed budget for each category. An important implication of budgeting is the violation of monetary fungiblity across categories.

Already long before behavioral economics was introduced into the scientific debate, economists contemplated how a general but sufficiently tractalbe utility function capturing consumer behavior should look like. Strotz (1957) argues that it is intuitively appealing to think of the consumer as follwing a two-stage maximization procedure akin to budgeting. In the first stage, the consumer allocates her overall budget across general good categories like, for example, food, clothing, and travel. Then, in the second stage she considers each category in isolation and allocates the previously determined category

budget across the individual goods inside that category.

Gorman (1959) investigates the characteristics a utility function needs to have in order for the solution to a full constrained utility maximization problem to be equivalent to the solution obtained in the described two-stage-procedure. A necessary and sufficient condition for budgeting to be rational is that the consumer's utility is either additively separable across budget categories or separable with budgetwise utilities entering through an intermediate function that is homogeneous of degree one.

This reveals how in my model narrow bracketing implies a boundedly rational form of budgeting as discussed by Blow and Crawford (2018). The narrow bracketer's expected utility representation is additively separable across brackets. Thus, she behaves as if she employed the described two-stage budgeting procedure with budgeting categories equivalent to the brackets in her system of brackets. However, her broad expected utility representation is not generally additively separable across brackets. Therefore, such budgeting behavior is not generally admissible according to the narrow bracketer's true preferences.

1.4 Model predictions

1.4.1 Constrained utility maximization

Consider economics 101 consumption bundle choice. DM faces the problem of allocating a given budget or wealth w across two goods. She chooses a consumption bundle $x \in \mathbb{R}^2_+$. We can write $x = (x_1, x_2)$ where x_1 denotes the amount of good 1 and x_2 denotes the amount of good 2. The per-unit prices of the two goods are p_1 and p_2 respectively.

As benchmark consider the maximization problem solved by a broad bracketer:

$$\max_{x_1,x_2} u(x_1,x_2) \quad \text{subject to} \quad p_1x_1 + p_2x_2 \leq w. \tag{6}$$

Denote by $x^* = (x_1^*, x_2^*)$ the broad optimum, i.e. the argument that maximizes (6). I am interested in how a narrow DM's choice deviates from her broad optimum. Suppose DM brackets each good in her consumption bundle separately, i.e. $B = \{\{x_1\}, \{x_2\}\}$. She solves

$$\max_{x_1,x_2} u(x_1,r_2) + u(r_1,x_2) \quad \text{subject to} \quad p_1x_1 + p_2x_2 \leq w. \tag{7}$$

Denote by $\tilde{x} = (\tilde{x}_1, \tilde{x}_2)$ the narrow optimum, i.e. the argument that maximizes (7).

The following assumption assures that the broad consumer's optimization problem is well-behaved. By the subsequent lemma, this assumption also implies well-behavedness of the narrow consumer's optimization problem.

Assumption 3 (Quasi-concavity of u)**.** For all $x, y \in \mathbb{R}^2_+$ and all $\lambda \in [0,1]$, the broad utility function $u : \mathbb{R}^2_+ \to \mathbb{R}$ satisfies $u(\lambda x + (1-\lambda)y) \geq \min\{u(x), u(y)\}$.

Lemma 1 (Quasi-concavity of $\tilde{u}$)**.** *Assumption 3 implies that the narrow utility function $\tilde{u} : \mathbb{R}^2_+ \to \mathbb{R}$ is quasi-concave.*

Proof. For all $x, y, r \in \mathbb{R}^2_+$ and all $\lambda \in [0,1]$, quasi-concavity of u implies,

$$u(\lambda x_1 + (1-\lambda)y_1, r_2) \geq \min\{u(x_1, r_2), u(y_1, r_2)\} \text{ and}$$

$$u(r_1, \lambda x_2 + (1-\lambda)y_2) \geq \min\{u(r_1, x_2), u(r_1, y_2)\}.$$

Since $\tilde{u}(\lambda x + (1-\lambda)y) = u(\lambda x_1 + (1-\lambda)y_1, r_2) + u(r_1, \lambda x_2 + (1-\lambda)y_2)$, $\tilde{u}(x) = u(x_1, r_2) + u(r_1, x_2)$, and $\tilde{u}(y) = u(y_1, r_2) + u(r_1, y_2)$ it follows, that $\tilde{u}(\lambda x + (1-\lambda)y) \geq \min\{\tilde{u}(x), \tilde{u}(y)\}$. □

The direction in which the narrow optimum departs from its broad counterpart depends crucially on the type of interdependencies between the two goods captured by the sign of the broad utility function's cross-derivative.

Definition 1. Goods 1 and 2 have negative interactions if $\frac{\partial^2 u}{\partial x_1 \partial x_2} < 0$ for all $x \in \mathbb{R}^2_+$. They have positive interactions if $\frac{\partial^2 u}{\partial x_1 \partial x_2} > 0$ for all $x \in \mathbb{R}^2_+$. The two goods have no interactions if $\frac{\partial^2 u}{\partial x_1 \partial x_2} = 0$ for all $x \in \mathbb{R}^2_+$.

Roughly, negative interactions are associated with substitutabilities between the two goods while postive interactions are associated with complementarities between the two goods.[7]

In Section 1.3.3 (Representation theorem) I alluded to the fact that the additive separability of the narrow utility function implies that the narrow bracketer disregards interactions. In the context of consumption bundle choice the following proposition illustrates this fact by comparing the indifference

[7] See Chambers and Echenique (2009) and Topkis (1998) for a detailed discussion on when a positive cross-derivative of the utility function implies complementarity.

curves of the narrow bracketer to their broad counterparts. Like all further proofs, the proof of the proposition is relegated to the appendix.

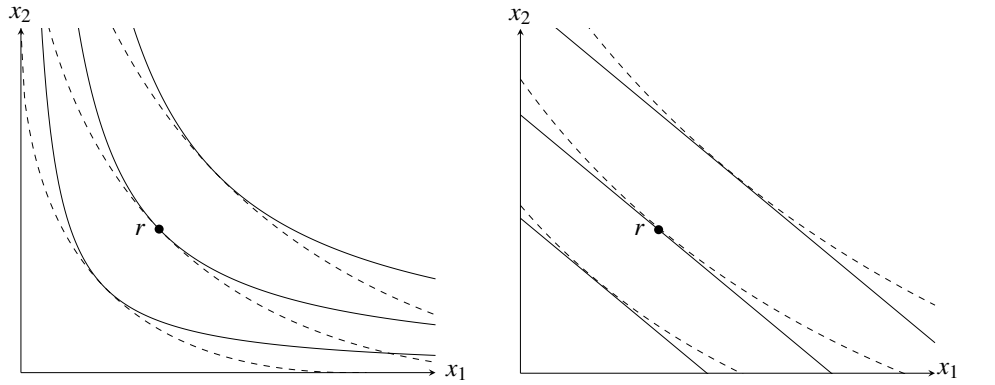
Denote by $MRS(x)$ and $\widetilde{MRS}(x)$ the marginal rates of substitution between good 1 and good 2 at bundle $x = (x_1, x_2)$ for the broad and the narrow bracketer respectively, i.e. $MRS(x) = \frac{\partial u}{\partial x_1} / \frac{\partial u}{\partial x_2}$ and $\widetilde{MRS}(x) = \frac{\partial \tilde{u}}{\partial x_1} / \frac{\partial \tilde{u}}{\partial x_2}$.

Proposition 1 (Indifference curves)**.** *Assume goods 1 and 2 have either positive, negative, or no interactions. For any amount of good 1, x_1, there exists a corresponding amount of good 2, $f(x_1)$, such that $MRS(x_1, f(x_1)) = \widetilde{MRS}(x_1, f(x_1))$ where $f(r_1) = r_2$, $f(x_1) < r_2$ for $x_1 < r_1$ and $f(x_1) > r_2$ for $x_1 > r_1$. Furthermore,*

- *Positive interactions $\Rightarrow MRS(x) > \widetilde{MRS}(x)$ for all $x \in \mathbb{R}^2_+$ with $x_2 > f(x_1)$ and $MRS(x) < \widetilde{MRS}(x)$ for all $x \in \mathbb{R}^2_+$ with $x_2 < f(x_1)$*
- *Negative interactions $\Rightarrow MRS(x) < \widetilde{MRS}(x)$ for all $x \in \mathbb{R}^2_+$ with $x_2 > f(x_1)$ and $MRS(x) > \widetilde{MRS}(x)$ for all $x \in \mathbb{R}^2_+$ with $x_2 < f(x_1)$*
- *No interactions $\Rightarrow MRS(x) = \widetilde{MRS}(x)$ for all $x \in \mathbb{R}^2_+$.*

Proposition 1 states that at the reference point the slopes of broad and narrow indifference curves are the same. Furthermore, for every amount of good 1, there exists a corresponding amount of good 2 such that the slopes of broad and narrow indifference curves are the same at that bundle. If there are positive interactions between the two goods, the narrow indifference curve is flatter than the broad indifference curve to the left of that bundle and steeper than the broad indifference curve to the right of that bundle. Therefore, narrow indifference curves are less convex than their broad counterparts if the two goods have positive interactions. Conversely, narrow indifference curves are more convex than their broad counterparts if the two goods have negative interactions. Intuitively, the more convex the indifference curves, the more complementary are the two goods. Therefore, in the case of positive interactions, the narrow bracketer can be interpreted as behaving as if the two goods were less complementary than they actually are and vice versa for the case of negative interactions.

Figure 1.1 illustrates the content of Proposition 1 for two specific broad utility functions given the reference point r. Consider first Figure 1.1a. The figure shows the indifference curve maps of broad (solid) and narrow (dashed)



(a) Positive interactions: $u(x_1,x_2) = \sqrt{x_1 x_2}$ (b) Negative interactions: $u(x_1,x_2) = \sqrt{x_1 + x_2}$

Figure 1.1: Comparison of broad (solid) and narrow (dashed) indifference curves with reference point r.

bracketer for a broad utility function belonging to the Cobb-Douglas family. The utility function is characterized by complementarities which is reflected by the convex shape of the broad indifference curves. The corresponding narrow indifference curves are less convex than their broad counterparts, reflecting the fact that the narrow bracketer disregards the positive interactions between the two goods. However, at the reference point and at any bundle with a distribution of amounts between the two goods proportional to the reference point distribution, broad and narrow indifference curves have the same slope. This illustrates how the narrow bracketer's tradeoffs between the two goods remain undistorted at the reference point and proportional bundles.

In contrast, Figure 1.1b depicts the indifference curve maps of broad and narrow bracketer for a perfect substitutes broad utility function with negative interactions between the two goods[8]. Perfect substitutability between the two goods implies that the broad indifference curves are straight lines. The narrow bracketer, however, disregards the negative utility interactions between the two goods. As a result, her indifference curves are convex. She treats the two goods as more complementary than they are. Again, her tradeoffs at the reference point and at bundles proportial to the reference point remain undistorted.

The next proposition investigates how the narrow bracketer's chosen consumption bundle departs from her optimal consumption bundle.

[8]The utility function is widely used in the context of decision making under risk since it has the CRRA property.

Denote by $d(x,y)$ the Euclidean distance between two consumption bundles $x,y \in \mathbb{R}^2_+$, i.e. $d(x,y) := \sqrt{(x_1-y_1)^2+(x_2-y_2)^2}$.

Proposition 2 (Narrow optimum)**.** *Assume $w = p_1r_1 + p_2r_2$ and $r \neq x^*$. The follwing two statements hold at any interior solutions x^* and $\tilde{x}$ to the maximization problems (6) and (7) respectively.*

- *Positive interactions* $\Rightarrow d(r,x^*) < d(r,\tilde{x})$
- *Negative interactions* $\Rightarrow d(r,x^*) > d(r,\tilde{x})$

Proposition 2 states that for budget balanced reference points, unless $r = x^*$, the narrow optimum $\tilde{x}$ is further away (in terms of Euclidean distance) from the reference point than the broad optimum x^* if the two goods have positive interactions. Conversely, the narrow optimum $\tilde{x}$ is closer to the reference point if the the two goods have negative interactions.

Considering Proposition 1 (Indifference curves) in isolation, one might expect that the narrow bracketer's disregard of interactions between the two goods and the resulting shape of her indifference curves imply that the narrow bracketer underdiversifies in the case of positive interactions and overdiversifies in the case of a negative interactions. However, while this intuition is not generally flawed, it does not take into account the role that the reference point plays for the narrow bracketer's decisions. The important role of the reference point is clarified by Proposition 2.

While the narrow bracketer disregards the interdependencies between the goods in her bundle, she is not fully ignorant of their existence. However, she does not consider changes from the respective reference quantities for the two goods simultaneously. Thus, when thinking about an alteration in the amount she might purchase of good 1, from r_1 to $x_1 \neq r_1$, she keeps the amount of good 2 fixed at the reference quantity of good 2, r_2. The reverse holds for alterations in the amount she purchases of good 2. Therefore, the narrow bracketer's appreciation of the interactions between the two goods only occurs separately for the two quantities she purchases and mistakenly with respect to the reference quantity of the respective other good. This implies that the reference point has a profound influence on the narrow bracketer's choice.

For example, if the goods have positive interactions, an unbalanced reference point with $r_1 > r_2$ pushes the narrow bracketer towards increasing her

consumption of good 2 and decreasing her consumption of good 1. This happens because the high reference quantity of good 1, r_1, makes investments in good 2 seem more attractive than investments in good 1, which are in the narrow bracketers mind combined with the relatively low reference quantity of good 2, r_2. Now, if the optimal consumption basket of the broad bracketer x^* prescribes $x_1^* \leq x_2^*$, the fact that the narrow optimum $\tilde{x}$ is pushed further from the reference point r compared to the broad optimum x^* in this constellation always implies that the bundle chosen by the narrow bracketer is less diversified than the bundle chosen by the broad bracketer. If, however, the broad optimum x^* prescribes $x_1^* > x_2^*$, the extra push away from r might induce the narrow bracketer to choose a more diversified consumption bundle than the broad bracketer even though she disregards the positive utility interactions between the chosen quantities x_1 and x_2. Depending on the constellation of reference point and broad optimum, we might therefore observe a narrow bracketer overdiversifying her consumption bundle compared to the broad optimum although the goods have positive interactions. Similarly, we might observe a narrow bracketer underdiversifying her conusmption bundle compared to the broad optimum although the goods have negative interactions.

Interestingly, if the goods have positive interactions the effect of the reference point on the narrow bracketer's chosen bundle goes into the opposite direction of the effect that loss-aversion implies in this setting. The chosen bundle of a loss-averse narrow bracketer is always closer to the reference point than the chosen bundle of a narrow bracketer without loss-aversion. Thus, while narrow bracketing in the case of negative interactions exacerbates the effects of loss-aversion, in the case of positive interactions it actually dampens the effects of loss-aversion. My results reveal that the reference point plays an important role in the decision making of a narrow bracketer independent of whether she is loss-averse or not.

1.4.2 Exchange economy

Consider an exchange economy with two consumers $i = 1,2$ and two goods. Consumer i's consumption bundle is denoted by $x^i = (x_1^i, x_2^i)$. An allocation $x \in \mathbb{R}_+^4$ is an assignment of a consumption bundle to each consumer, i.e. $x = (x^1, x^2) = ((x_1^1, x_2^1), (x_1^2, x_2^2))$. The total endowments of goods 1 and 2 in the

economy are given by $\omega_1 > 0$ and $\omega_2 > 0$ respectively. The initial endowment allocation is denoted $\omega = (\omega^1, \omega^2)$ with $\omega^1 = (\omega_1^1, \omega_2^1)$ denoting consumer 1's endowment such that consumer 2's endowment is given by $\omega^2 = (\omega_1 - \omega_1^1, \omega_2 - \omega_2^1)$. I assume $\omega_1^i, \omega_2^i \geq 0$ for $i = 1,2$. The systems of brackets for the two consumers are given by $B^i = \{\{x_1^i\}, \{x_2^i\}\}$ for $i = 1,2$.

I refer to the *broad economy* as the exchange economy in which both consumers bracket broadly and to the *narrow economy* as the exchange economy in which both consumers bracket narrowly. Furthermore, I refer to the *broad contract curve* as the set of Pareto optimal allocations of the broad economy and to the *broad core* as the set of Pareto optimal allocations that constitute Pareto improvements with respect to the initial endowment allocation in the broad economy. *Narrow contract curve* and *narrow core* are defined analogously. It is a well known fact that any Walrasian equilibrium of an exchange economy is an element of its core (Mas-Colell et al., 1995).

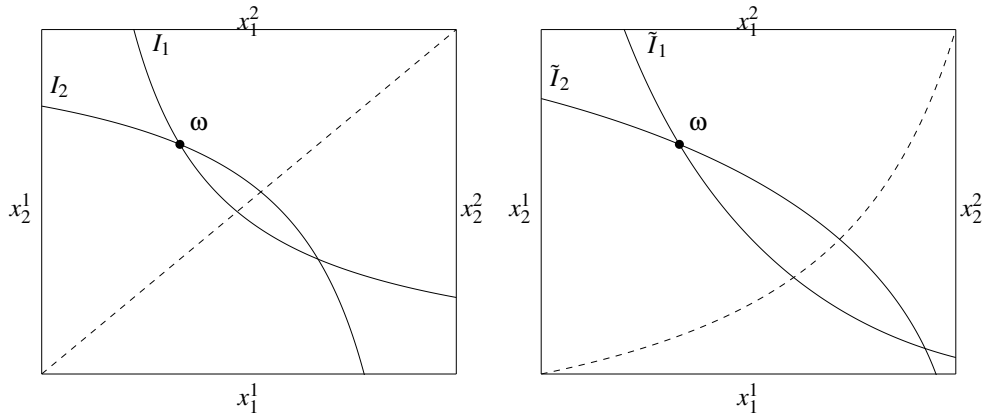
The following proposition shows how choice bracketing systematically affects the volume of trade in the exchange economy.

Proposition 3 (Exchange economy)**.** *Assume that consumer i's reference point is equal to her initial endowment, i.e.* $r^i = \omega^i$ *for* $i = 1,2$*. For any inditial endowment allocation* ω *such that* $MRS^1(\omega^1) \neq MRS^2(\omega^2)$*, if two allocations* x *and* $\tilde{x}$ *are elements of the broad and narrow core respectively and they are not at the corner, then*

- *Positive interactions for both consumers* $\Rightarrow d(\omega, x) < d(\omega, \tilde{x})$.
- *Negative interactions for both consumers* $\Rightarrow d(\omega, x) > d(\omega, \tilde{x})$.

Proposition 3 states that starting from any inital endowment allocation there is more trade in the narrow exchange economy compared to its broad counterpart if the two goods have positive interactions. Conversely, there is less trade in the narrow exchange economy compared to its broad counterpart if the two goods have negative interactions.

Figure 1.2 illustrates the difference between a broad exchange economy and its narrow counterpart when there are positive interactions between the two goods. Consider first Figure 1.2a which shows the broad economy in an Edgeworth-box. At the initial endowment allocation ω consumer 1 holds a bundle that is unbalanced towards good 2 while consumer 2 holds a bundle



(a) Broad economy (b) Narrow economy

Figure 1.2: Edgeworth-box comparison of broad and narrow exchange economy with broad utilities $u^i(x_1^i, x_2^i) = \sqrt{x_1^i, x_2^i}$ for $i = 1,2$ (positive interactions) and reference points $r^i = \omega^i$ for $i = 1,2$. In each Edgeworth-box the lower left corner corresponds to consumer 1's origin and the upper right corner corresponds to consumer 2's origin. I_i and $\tilde{I}_i$ for $i = 1,2$ respectively denote consumer i's broad and narrow indifference curve reached at the initial endowment allocation ω. The dashed graph displays the contract curve of the respective economy. The part of the contract curve that is enclosed by the lense that opens up between the two indifference curves corresponds to the core of the economy.

that is unbalanced towards good 1. The indifference curves that the two consumers reach at this initial endowment allocation intersect. Any allocation inside the lense enclosed by the two indifference curves constitutes a Pareto improvement with respect to ω. In particular, redistributing a small amount of good 1 in exhange for a small amount of good 2 from consumer 2 to consumer 1 resulting in more balanced bundles makes both consumers better off. Performing a series of such small trades allows the consumers to arrive at the broad core which is located on the part of the contract curve that intersects with the lense. At the broad core the consumers have reached a Pareto optimal allocation. Since in this example the broad contract curve is on the 45° line, any such allocation has the property that it equalizes the amounts of good 1 and good 2 allocated to a given consumer. Thus, in the given broad economy we should expect the consumers to perform trades that move them from the initial endowment allocation towards an allocation that fully balances their consumption bundles.

Consider now the corresponding narrow exchange economy displayed in Figure 1.2b. As in the broad economy, the consumer's narrow indifference curves intersect at the initial endowment allocation. Furthermore, moving to an allocation which induces bundles that are more balanced between the two goods for both consumers constitutes a Pareto improvement. However, in the narrow economy the overall set of allocations constituting a Pareto improvement with respect to ω extends much further to the lower right corner of the Edgeworth-box than in the broad economy. This is a direct consequence of the narrow consumers' disregard of the positive interactions between the good dimensions in their bundles. As stated in Proposition 1 (Indifference curves) positive interactions between the two goods imply that the narrow indifference curves are less convex compared to their broad counterparts. The narrow consumers perceive the two good dimensions of their bundles as less complementary than they actually are.

Relatedly, the narrow contract curve is not on the 45° line but bent towards the lower right corner of the Edgeworth-box. As a result, the bundles in the narrow core allocations are not balanced between the two goods. Instead, any allocation in the narrow core has the property that consumer 1's bundle is unbalanced towards good 1 and consumer 2's bundle is unbalanced towards good 2. Interestingly, the imbalance in the consumers' bundles at the narrow core is exactly opposite to the imbalance in the consumers' bundles

at the initial endowment allocation. This property of the narrow core mirrors the logic of Proposition 2 (Narrow optimum). The consumers appreciate the positive interactions between the two good dimensions mistakenly with respect to their reference points. Akin to status-quo based reference points, consumers' reference points are assumed to be equal to their respective bundles in the initial endowment allocation. Consider consumer 1. Her bundle in the initial endowment allocation is unbalanced towards good 2. Due to the complementarity between the two good dimensions, the resulting high reference point in the second good dimension makes increases in the amount of good 1 seem relatively more attractive than they actually are. Similarly, the low reference point in the first good dimension makes increases in the amount of good 2 seem relatively less attractive than they actually are. This constellation implies a push of narrow consumer 1's preference towards bundles that are characterized by an imbalance opposite to the imbalance in her initial endowment, i.e. towards good 1. Similarly, consumer 2's preferences is pushed towards bundles that are imbalanced towards good 2. As a result, the volume of trade predicted for the narrow economy is larger than the volume of trade predicted for the broad economy.

1.5 Experiment

1.5.1 Design

In the experiment I elicit participants' willingnesses to pay (WTP) for portfolios. Each portfolio consists of two assets, a blue and an orange asset. The blue asset yields blue points and the orange asset yields orange points. Point earnings from a portfolio are determined by a coin toss performed by the computer. Importantly, it is the same coin toss that determines a participant's point earnings from the blue and the orange asset in a portfolio. Relating to my theoretical framework (Section 1.3.1) a portfolio in the experiment corresponds to a prospect while the two assets correspond to the subprospects comprising the prospect.

Preferences, or more accurately broad preferences, over portfolios are partly (risk preferences still matter) induced via a payment rule that translates any combination of blue and orange point earnings into payments. The payment rule induces negative interactions between blue and orange points, i.e. the more blue points a participant receives, the less valuable is an increase

	6	12	18	20	30	32	36	40	42	44	50	52	54	74
6					0,24 €	0,47 €	0,90 €	1,29 €	1,47 €	1,64 €	2,09 €	2,22 €	2,34 €	2,99 €
12					0,90 €	1,10 €	1,47 €	1,80 €	1,95 €	2,09 €	2,45 €	2,55 €	2,64 €	
18			0,24 €	0,47 €	1,47 €	1,64 €	1,95 €	2,22 €	2,34 €	2,45 €	2,72 €	2,79 €	2,85 €	
20			0,47 €	0,69 €	1,64 €	1,80 €	2,09 €	2,34 €	2,45 €	2,55 €	2,79 €	2,85 €	2,90 €	
30	0,24 €	0,90 €	1,47 €	1,64 €	2,34 €	2,45 €	2,64 €	2,79 €	2,85 €	2,90 €	2,99 €	3,00 €		
32	0,47 €	1,10 €	1,64 €	1,80 €	2,45 €	2,55 €	2,72 €	2,85 €	2,90 €	2,94 €	3,00 €			
36	0,90 €	1,47 €	1,95 €	2,09 €	2,64 €	2,72 €	2,85 €	2,94 €	2,97 €	2,99 €				
40	1,29 €	1,80 €	2,22 €	2,34 €	2,79 €	2,85 €	2,94 €	2,99 €	3,00 €					
42	1,47 €	1,95 €	2,34 €	2,45 €	2,85 €	2,90 €	2,97 €	3,00 €						
44	1,64 €	2,09 €	2,45 €	2,55 €	2,90 €	2,94 €	2,99 €							
50	2,09 €	2,45 €	2,72 €	2,79 €	2,99 €	3,00 €								
52	2,22 €	2,55 €	2,79 €	2,85 €	3,00 €									
54	2,34 €	2,64 €	2,85 €	2,90 €										
74	2,99 €													

Figure 1.3: The payment table shown to participants in the experiment. The payment associated with a specific combination of blue and orange points can be found by choosing the row according to the amount of blue points and the column according to the amount of orange points.

in orange points and vice versa. Throughout the experiment participants have access to a table stating the respective payments associated with all different combinations of blue and orange point earnings. The payment table is displayed in Figure 1.3.

The experiment has 20 rounds. In each round the participant is provisionally allocated a simple portfolio, the base-portfolio. The base-portfolio is deterministic, i.e. it yields the same point earnings irrespective of the result of the coin toss. It remains constant over the course of the experiment. I use the base-portfolio to induce participants' reference points. The base-portfolio is displayed in the first row of Table 1.1.

For every round of the experiment a random draw determines whether that round is a trade-round or a base-round. Both round types are equally likely. In base-rounds the participant keeps her base-portfolio. In trade-rounds she is offered another portfolio (trade-portfolio). Her WTP for the trade-portfolio is elicited via a multiple choice list. In each row of the choice list the participant has to make a decision between the offered trade-portfolio and an increasing certain payment. For each submitted choice list one row is randomly chosen and the participant's decision in the respective row is implemented. Since participants do not know whether a given round is a trade-round or a base-

round, they fill out a multiple choice list in every round.

In each round the participant is displayed a decision-screen. At the top of the decision screen she sees the base-portfolio. Below, she can click a button to view the trade-portfolio of that round. The participant can switch back and forth between viewing the trade- and the base-portfolio anytime. Below the respective portfolio the multiple choice list for the offered trade-portfolio is displayed. I enforce a single switchpoint.

Overall, I elicit WTP for 10 different trade-portfolios in two treatments. The portfolios used in the experiment are displayed in Table 1.1. The experiment has a within-subject design. Therefore, each participant fills out a multiple choice list for each of the 10 trade-portfolios twice, once in the *broad treatment* and once in the *narrow treatment* (hence the 20 rounds). As suggested by their names, the treatments are designed to induce subjects to bracket broadly in the broad treatment and narrowly in the narrow treatment. The order in which participants see the different trade-portfolios in the two treatments is randomized.

The treatments differ in how the participant can access information about the contents of the trade-portfolio. In the broad treatment the participant views the blue and orange asset comprising the trade-portfolio jointly. In the narrow treatment the particpant can view the blue and orange asset comprising the trade-portfolio only separately. The view of the other asset is kept fixed at the respective asset in the base-portfolio. The participant can change between viewing the blue asset in the trade-portfolio and viewing the orange asset in the trade-portfolio anytime by clicking on a button. After clicking the button, the participant sees a waiting screen for 5 seconds and is then redirected to the respective view of the trade-portfolio. Importantly, the information available to the participant in the narrow treatment is the same as the information available to her in the broad treatment. The treatments only differ in how easy it is for the participant to jointly consider the two assets that comprise the trade-portfolio.

I conducted 10 online-sessions with roughly 18 subjects each. Overall 171 subjects took part in the experiment. The sessions took place in September 2020 with subjects from the WZB-Technical University laboratory subject pool in Berlin. Subjects were invited to participate in the experiment using ORSEE (Greiner, 2015). In conducting the sessions I closely followed the UCSC LEEPS Lab Protocol for Online Economics Experiments (Zhao et al.,

Portfolio	Blue asset		Orange asset		EV	Variance
	Heads	Tails	Heads	Tails		
Base	30	30	6	6	€0.24	0.00
Trade A1	30	42	32	6	€1.96	0.24
Trade A2	30	42	6	32	€1.57	1.77
Trade A3	52	30	12	30	€2.45	0.01
Trade A4	30	52	12	30	€1.95	1.10
Trade B1	30	74	6	6	€1.62	1.89
Trade B2	44	54	6	6	€1.99	0.12
Trade B3	30	30	6	50	€1.62	1.89
Trade B4	30	30	20	30	€1.99	0.12
Trade C1	52	44	30	32	€2.97	0.00
Trade C2	30	32	52	44	€2.97	0.00

Table 1.1: The portfolios used in the experiment. For each portfolio the table shows the number of blue and orange points the portfolio yields depending on the outcome of the coin toss for that portfolio. Furthermore, it shows expected value (EV) and variance of each portfolio rounded to two decimal places.

2020). The experiment was programmed in oTree (Chen et al., 2016). I pre-registered the experiment on the AEA RCT Registry, including a pre-analysis plan and power calculation (Vorjohann, 2020). The experimental sessions were preceded by two pilot sessions run in July 2020. I do not use the data from these pilot sessions in my analysis.

1.5.2 Hypotheses

Table 1.2 summarizes the correspondence between theory and experiment. Portfolios are probability distributions over combinations of blue and orange point earnings. A combination of blue and orange point earnings is a two-dimensional outcome. Thus, there is a direct correspondence between portfolios in the experiment and prospects as defined in my theoretical framework (Section 1.3.1). Furthermore, the blue asset in a portfolio is in effect the marginal distribution over blue points induced by the portfolio. Similarly, the orange asset in a portfolio is the marginal distribution over orange points induced by the portfolio. The assets in a portfolio therefore correspond to its subprospects. The payment rule translates combinations of blue and orange point earnings to money earnings. Following induced value theory (Smith, 1976) the broad utility associated with a combination of blue and orange point

Theory	Experiment
Prospect	Portfolio
Subprospects	Blue asset and orange asset
Outcome	Combination of blue and orange point earnings
Outcome dimensions	Blue and orange points
Broad utility of an outcome	Payment for a combination of blue and orange point earnings
Reference point	Blue and orange point earnings in the base-portfolio

Table 1.2: Correspondence between theory and experiment.

earnings can be measured by the payment it generates.

The reference point as defined by Axiom 2 (Reference point) in Section 1.3.2 plays a central role in my theory of choice bracketing. My model relies on the existence of a reference point but remains agnostic about which specific outcome constitutes the reference point. However, to design a meaningful test for the validity of my model I require additional knowledge about the reference points of subjects in my experiment. Therefore, I introduce the base-portfolio into my experimental design. The base-portfolio serves the purpose of inducing its deterministic outcome as reference point. This purpose is achieved in two ways, each of which builds on a prominent theory of the nature of reference points. First, by provisionally allocating the base-portfolio to subjects at the beginning of each round, the base-portfolio is established as the status-quo. Second, by implementing the base-portfolio instead of a subject's decision with a probability of 0.5 in each round, the base-portfolio enters the subject's expected outcome from a given decision. This design feature is inspired by Abeler et al. (2011) and builds on the theory of expectation-based reference points (Kőszegi and Rabin, 2006a).

Since my axiomatization builds on the connection between broad and narrow preference relation, both characterizing one and the same decision maker, I employ a within-subject design. The idea is to elicit a subject's WTP for a given trade-portfolio twice, once when she brackets the two assets in the portfolio broadly and once when she brackets them narrowly. I manipulate how subjects bracket the assets in a trade-portfolio by varying the ease at which they can be considered jointly. In the broad treatment, subjects see the two assets on the same screen. This makes it relatively easy to integrate them and

keep track of the overall portfolio they comprise. In contrast, to integrate the assets in the narrow treatment, subjects have to recall them since they are accessible only on separate screens. The waiting time imposed when switching between viewing each of the assets in the trade-portfolio further complicates joint consideration.

The main goal of my experiment is to test the validity of the behavioral axioms underlying my theoretical model. Additionally, I test one of my model's predictions concering the role of the reference point. The trade-portfolios for which I elicit subjects' WTP (Table 1.1) can be classified into three groups. Trade-portfolios A1-A4 are designed to test Axiom 1 (Correlation neglect). Trade-portfolios B1-B4 are designed to test Axiom 2 (Reference point). Trade-portfolios C1 and C2 are designed to test the model prediction.

Consider first trade-portfolios A1 and A2. The two portfolios induce the same marginal distributions over blue and orange points, i.e. a fifty-fifty chance between 30 and 42 blue points and a fifty-fifty chance between 32 and 6 orange points. Thus, if Axiom 1 (Correlation neglect) holds, subjects in the narrow treatment are expected to have the same WTP for the two portfolios. However, overall trade-portfolios A1 and A2 are not the same. They differ in the joint distribution over blue and orange points they induce. Trade-portfolio A1 induces a fifty-fifty chance between the overall outcomes (30 blue points, 32 orange points) and (42 blue points, 6 orange points). This is equivalent to a fifty-fifty chance between receiving €2.45 and €1.47 (see the payment table in Figure 1.3) and implies an expected value of €1.96. In contrast, trade-portfolio A2 induces a fifty-fifty chance between the overall outcomes (30 blue points, 6 orange points) and (42 blue points, 32 orange points). This is equivalent to a fifty-fifty chance between €0.24 and €2.9 and implies an expected value of €1.57. Thus, in the broad treatment a risk neutral subject should have a higher WTP for trade-portfolio A1 compared to trade-portfolio A2. Furthermore, since A2 has a higher variance than A1, the same should hold for a risk averse subject. An equivalent logic applies to the pair of trade-portfolios A3 and A4.

Denote by $WTP_i(P)$ the WTP for trade-portfolio P expressed by subject i. $|WTP_i(P) - WTP_i(Q)|$ denotes the absolute value of the WTP difference between portfolios P and Q expressed by subject i in a given treatment. Hypothesis 1 summarizes the theoretical predictions based on Axiom 1 (Correlation neglect).

Hypothesis 1 (Correlation neglect)**.**

(a) $|WTP_i(A1) - WTP_i(A2)|$ *is higher in the broad treatment than in the narrow treatment.*

(b) $|WTP_i(A3) - WTP_i(A4)|$ *is higher in the broad treatment than in the narrow treatment.*

Next, consider trade-portfolios B1 and B2 in Table 1.1. Both portfolios contain the same orange asset which yields 6 orange points independent of the outcome of the coin toss. Furthermore, the orange asset in the two portfolios is equivalent to the orange asset in the base-portfolio. B1 and B2 differ from each other and the base-portfolio only in the blue asset they contain. Thus, if Axiom 2 (Reference point) holds, we should observe the same ordering between the WTP for the two portfolios across treatments. Since trade-portfolio B2 has both a higher expected value and a lower variance than trade-portfolio B1, risk-neutral and risk-averse subjects should express a lower WTP for B1 than for B2.

Similarly, trade-portfolios B3 and B4 in Table 1.1 differ from each other and the base-portfolio only in the orange asset they contain. Therefore, based on Axiom 2 (Reference point) I expect that if a subject's WTP for B3 is lower than her WTP for B4 in the broad treatment, this also holds for the same subject in the narrow treatment. Again, based on expected value and variance the WTP difference between B3 and B4 should be negative for risk-neutral and risk-averse subjects. The theoretical predictions based on Axiom 2 are summarized in Hypothesis 2.

Hypothesis 2 (Reference point)**.**

(a) The sign of $WTP_i(B1) - WTP_i(B2)$ *is the same across treatments.*

(b) The sign of $WTP_i(B3) - WTP_i(B4)$ *is the same across treatments.*

Finally, consider trade-portfolios C1 and C2 in Table 1.1. The only difference between C1 and C2 is that the labeling of the assets they contain is interchanged. The asset that yields 52 points for heads and 44 points for tails is called the blue asset in C1 and the orange asset in C2. Similarly, the asset that yields 30 points for heads and 32 points for tails is called the orange

asset in C1 and the blue asset in C2. However, since blue and orange points enter the payment rule symmetrically (see the payment table in Figure 1.3), the lottery over payments the two portfolios induce is exactly the same. Both portfolios are equivalent to a fifty-fifty chance between €3 and €2.94. Therefore, a subject in the broad treatment should have the same WTP for C1 as for C2.

Based on my model of narrow bracketing I do not expect the same to hold for subjects in the narrow treatment. A narrow bracketer evaluates the broad expected utility function separately for each asset in the portfolio keeping the respective other asset fixed at the reference point level (see Theorem 1). Suppose a subject is risk neutral such that her broad expected utility from a portfolio is equivalent to the expected value of that portfolio. Consider trade-portfolio C1. Regarding the blue asset in the portfolio, the narrow bracketer can be modeled as calculating the expected value of a fictitious portfolio that combines the blue asset in C1 with the orange asset in the base-portfolio. Such a portfolio would induce a fifty-fifty chance between the outcomes (52 blue points, 6 orange points) and (44 blue points, 6 orange points) associated with an expected value of €2.28. In turn, regarding the orange asset in C1, the narrow bracketer can be modeled as calculating the expected value of another fictitious portfolio that combines the blue asset in the base-portfolio with the orange asset in C1 which is €2.05. Similarly, the expected values resulting from such a separate evaluation of the assets in C2 are €1.62 and €1.69. Since according to my representation theorem the narrow bracketer's preferences over portfolios are representable by the sum of these separately evaluated expected values, I expect that in the narrow treatment a risk-neutral subject's WTP for C1 is higher than her WTP for C2.

The prediction derived for trade-portfolios C1 and C2 is a manifestation of the general intution concerning the role that an unbalanced reference point plays for the behavior of a narrow bracketer discussed in Section 1.4 (Model predictions). The base-portfolio yields considerably more blue points than orange points. It therefore induces a reference point that is unbalanced towards blue points. Since C1 yields more blue than orange points irrespective of the outcome of the coin toss, C1 is characterized by an imbalance towards blue points as well. Conversely, C2 is characterized by an imbalance towards orange points. Now, the negative interactions between blue and orange points induced by the payment rule push the narrow bracketer's preference towards

the otherwise equivalent trade-portfolio which is characterized by an imbalance towards blue points, namely C1. Hypothesis 3 summarizes the experimental prediction on the role of the reference point derived from my model.

Hypothesis 3 (Role of the reference point)**.** $WTP_i(C1) - WTP_i(C2)$ *is lower in the broad treatment than in the narrow treatment.*

1.5.3 Results

A prerequisite for me to be able to use my experimental data to test my model of choice bracketing is that the treatment manipulation worked. I require that subjects bracket the blue and orange asset comprising a trade-portfolio jointly in the broad treatment and separately in in the narrow treatment. Suppose the treatment manipulation did not work. Then, a subject's WTP for a given trade-portfolio should be the same across treatments. For each trade-portfolio used in the experiment, Figure 1.4 shows the distribution of the within subject WTP difference between treatments. While the distributions are centered around zero for all portfolios, the plots also show considerable variation in the WTP for the same portfolio between treatments. Across the different portfolios 41-77% of subjects in my sample express a WTP in the broad treatment that differs by more than €0.1 from the WTP they express for the same portfolio in the narrow treatment. These results indicate that while the treatment manipulation may not have been successful for all subjects in my experiment, there still is a considerable share of subjects that behaves differently across the two treatments.

Consider first Hypothesis 1 (Correlation neglect). For each of the two trade-portfolio pairs A1&A2 and A3&A4 the hypothesis states that the difference in WTP between the two portfolios should be higher in the broad treatment compared to the narrow treatment. Hypothesis 1a concerning A1&A2 is fulfilled for roughly 50% of subjects in my sample. The share of subjects for whom Hypothesis 1b concerning A3&A4 is fulfilled is 42%. Figure 1.5 compares the means of the absolute value of the WTP difference between the portfolios in each of the two portfolio pairs across treatments. In line with Hypothesis 1a, the mean absolute WTP difference between trade-portfolios A1 and A2 is higher in the broad treatment than in the narrow treatment. Furthermore, this observation is confirmed to be statistically significant in a one-sided paired two-sample t-test (p=0.04). However, Hypothesis 1b is not

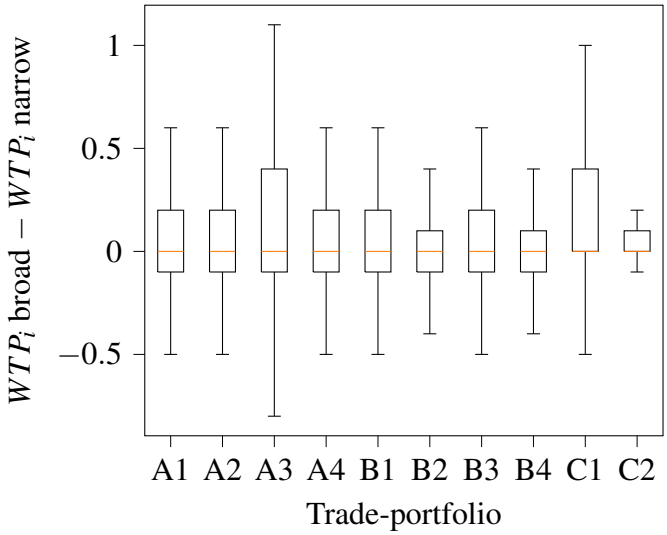


Figure 1.4: Comparison of WTP for trade-portfolios across treatments. Each boxplot visualizes the distribution of the difference between WTP for a trade-portfolio in the broad treatment and WTP for the same trade-portfolio in the narrow treatment. As usual, the box shows the interquartile range with the horizontal line between lower and upper quartile marking the median. The whiskers extend to the extrema of the distribution excluding outliers.

supported in my experimental data. Figure 1.5 already shows that the mean absolute WTP difference between A3 and A4 is virtually the same across the two treatments, a result which is confirmed by the corresponding t-test.

According to Axiom 1 (Correlation neglect), a subject that brackets the two assets in a portfolio narrowly should be indifferent between the portfolios in the two trade-portfolio pairs A1&A2 and A3&A4. Thus, if all subjects bracketed narrowly in the narrow treatment, we should observe a mean WTP difference of zero for the two pairs in that treatment. Even for A1&A2 this is clearly not the case. Only 20% of subjects in my sample show a WTP difference between A1 and A2 of at most €0.1 in the narrow treatment. With a mere 11% the share of subjects that can be accordingly classified as correlation neglecters in the narrow treatment is much lower for A3&A4. However, for both portfolio pairs I do observe a considerable drop in the share of correlation neglecters in the broad treatment compared to the narrow treatment. In the broad treatment only roughly 8% of subjects show a WTP difference of at most €0.1 between the portfolios in the respective pair. I interpret this drop in the share of correlation neglecters when moving from the narrow to

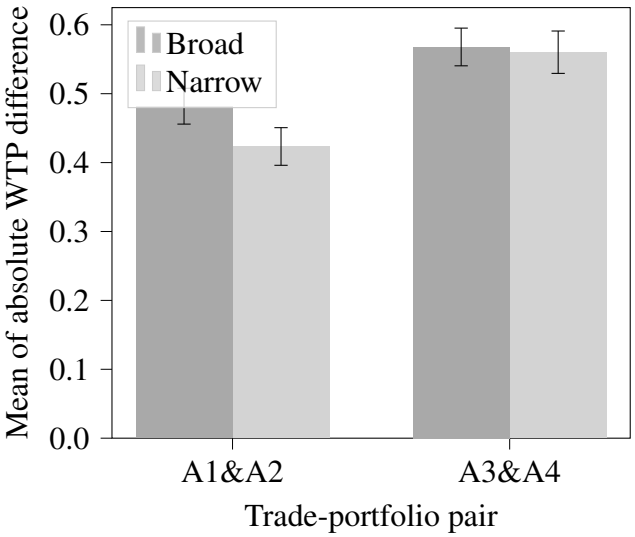


Figure 1.5: Visualization of the results for Hypothesis 1 (Correlation neglect). The graph shows the mean of the absolute value of the WTP difference between the portfolios in each of the two trade-portfolio pairs A1&A2 and A3&A4 separated by treatment. The error bars indicate the standard error of the respective mean measured by the standard deviation divided by the squareroot of the sample size.

the broad treatment as additional suggestive evidence that narrow bracketing is related to correlation neglect as asserted by Axiom 1 (Correlation neglect).

Result 1 (Correlation neglect)**.** *My experimental results concerning Hypothesis 1 provide preliminary evidence for the validity of Axiom 1.*

Consider now Hypothesis 2 (Reference point). For each of the two portfolio pairs B1&B2 and B3&B4, the hypothesis states that the ordering of WTP for the two portfolios in the pair should be the same across treatments. Hypothesis 2 is fulfilled for the majority of subjects in my sample. For 64 and 69% of subjects respectively I observe the same ordering of WTP within portfolio pairs B1&B2 and B3&B4. For each of the two portfolio pairs Table 1.3 shows a contingency table providing the respective frequencies of the feasible combinations of WTP orderings in the broad and narrow treatment. Consider first Table 1.3a. As expected, for the majority of subjects $WTP_i(B1) - WTP_i(B2)$ is negative in both treatments. This means that the majority of subjects prefers the portfolio in the pair that is characterized by a higher expected value and a lower variance in both treatments. In a Pearson's chi-squared test the null hypothesis of independence of the signs of

	Narrow treatment			
Broad treatment	$-$	0	$+$	Total
$-$	96	9	15	120
0	15	3	6	24
$+$	13	3	11	27
Total	124	15	32	171

(a) Sign of $WTP_i(B1) - WTP_i(B2)$

	Narrow treatment			
Broad treatment	$-$	0	$+$	Total
$-$	100	4	24	128
0	6	6	1	13
$+$	15	3	12	30
Total	121	13	37	171

(b) Sign of $WTP_i(B3) - WTP_i(B4)$

Table 1.3: Visualization of the results for Hypothesis 2 (Reference point). Each contingency table displays the bivariate frequency distribution of the sign of the WTP difference between the trade-portfolios in the respective pair observed in the two treatments.

$WTP_i(B1) - WTP_i(B2)$ in the broad and narrow treatment is clearly rejected (p=0.007). This result is in line with Hypothesis 2a. Considering Table 1.3b a similar picture emerges. For 100 out of 171 subjects $WTP_i(B3) - WTP_i(B4)$ is negative in both treatments. Furthermore, a Pearson's chi squared test clearly rejects the null hypothesis of independence (p<0.001).

Overall, the experimental results concerning Hypothesis 2 are in line with Axiom 2 (Reference point). The portfolios in the pairs B1&B2 and B3&B4 differ from each other and the base-portfolio in only one asset. Therefore, the observation of equal WTP orderings between the portfolios within each pair across treatments is consistent with the existence of a reference point, in this case equal to the outcome of the base-portfolio, that ties together the broad and narrow preference relations on portfolios. However, in conjunction with the preceding analysis of my experimental data, this interpretation should be taken with a grain of salt. My analysis so far suggests that my treatment manipulation was only partially successful in inducing subjects to bracket narrowly in the narrow treatment. At the same time, Hypothesis 2

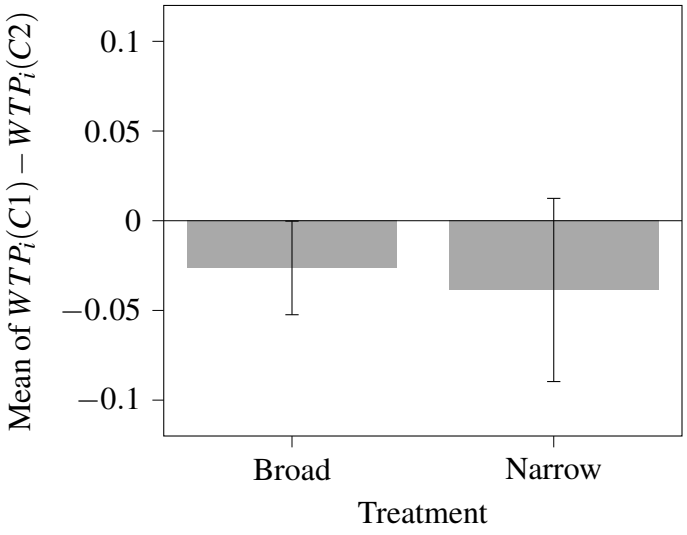


Figure 1.6: Visualization of the result for Hypothesis 3 (Role of the reference point). The graph shows the mean of $WTP_i(C1) - WTP_i(C2)$ in the broad an the narrow treatment. The error bars indicate the standard error of the respective mean measured by the standard deviation divided by the squareroot of the sample size.

should equally hold for a subject who brackets broadly in both treatments. Therefore, in light of my relatively weak treatment effect, the presented test of this hypothesis is not entirely conclusive.

Result 2 (Reference point)**.** *My experimental results concerning Hypothesis 2 support the validity of Axiom 2.*

Finally, consider Hypothesis 3 (Role of the reference point). The hypothesis states that the difference in WTP between trade-portfolios C1 and C2 in the narrow treatment should be lower than the same WTP difference in the broad treatment. Hypothesis 3 is fulfilled for only 29% of subjects in my sample. For the vast majority of subjects (85%) the absolute values of the WTP difference between the two portfolios differ by at most €0.1 across the two treatments. Figure 1.6 compares the means of the WTP difference between C1 and C2 across treatments. In both treatments the mean of this WTP difference is close to zero. A paired two-sample t-test confirms that there is no significant difference between the mean WTP differences across treatments.

For the majority of subjects in my sample (60%) the WTP for portfolios C1 and C2 in the broad treatment differ by at most €0.1. This is as expected since the two portfolios are essentially equivalent. A subject who brackets the

two assets in a portfolio broadly should be indifferent between C1 and C2. In the narrow treatment the share of subjects for whom the WTP for C1 and C2 differ by at most €0.1 drops to 47%. This drop in the share of subjects behaving consistent with broad bracketing in the narrow treatment is reflected by a higher variance in the distribution of the WTP difference between C1 and C2 in that treatment.

Result 3 (Role of the reference point)**.** *My experimental results concerning Hypothesis 3 do not support the validity of my model's prediction on the role of the reference point.*

Overall, I interpret the results of my experiment as providing preliminary evidence for the validity of the behavioral axioms underlying my model of choice bracketing. However, since the treatment effect I observe in the experiment is relatively weak, I am not able to present a conclusive assessment of the validity of my model as a whole. In particular concerning my model's prediction on the role of the reference point, further empirical research will be needed to determine whether the lack of support for the prediction provided by my experiment is an artifact of my relatively weak treatment manipulation or a more general flaw of my modeling approach.

Fortunately, my experimental design can easily be adjusted to make the treatment stronger. For example, one could increase the waiting time that subjects need to endure in the narrow treatment when switching between the information on the two assets in a portfolio. This would make it more costly for subjects to repeatedly view each of the two assets and require a better memory for their joint consideration. Another possibility would be to increase the dimensionality of the decision problem by increasing the number of assets in a portfolio. This would make it harder for subjects to bracket broadly overall and especially so in the narrow treatment.

1.6 Conclusion

Narrow bracketing affects individual decision making. Individual decision making is the very basis of almost all economic activity. Therefore, the potential implications of this behavioral bias go through the whole economy. Indeed, empirical evidence suggests that narrow bracketing adversely affects behavior in a vast variety of important economic settings. In this paper I present a generally applicable theoretical model of choice bracketing.

Previous models of choice bracketing are restricted to one-dimensional outcome spaces. Therefore, these models can accomodate only a small subset of the relevant economic applications. Allowing for multidimensional outcome spaces, my model opens up the possibility to systematically study the effects of narrow bracketing in new economic applications ranging from complex contract negotiations to basic consumption bundle choice. Furthermore, I derive my model from basic behavioral assumptions. In contrast to a model that is designed to generate specific predictions in a given setting my model is therefore more likely to make accurate predictions when applied across a variety of different settings. Finally, my model provides a theoretical framework that can inspire and organize future empirical research on choice bracketing.

An essential component of my model of choice bracketing is the reference point. It ties the narrow preference relation to its broad counterpart. However, my model takes the reference point as given and stays agnostic about where it comes from. In my applications I show that the direction and extent of the deviation of a narrow bracketer's choices from her broad optimum crucially depends on the specific form of the reference point. Future research investigating the nature of reference points in narrow bracketing is therefore essential to further our understanding of this behavioral bias.

Another important component of my model is the system of brackets. It characterizes the degree to which a decision maker brackets narrowly. For a given system of brackets my model fully characterizes the representation of the narrow preference relation. A promising direction for future research would be to identify a way to elicit a decision maker's system of brackets from choice data. My experimental results suggest that the system of brackets characterizing the narrow preference relation is not set in stone. Instead, the extent to which a decision maker brackets narrowly depends on how easy it is for her to access information on the different dimensions of her decision problem simultaneously. In that respect my experimental design can serve as a guideline for finding ways to improve individual decision making.

2 Welfare-based altruism

This chapter is based on joint work with Yves Breitmoser.

2.1 Abstract

Why do people give when asked, but prefer not to be asked, and even take when possible? We show that standard behavioral axioms including separability, narrow bracketing, and scaling invariance predict these seemingly inconsistent observations. Specifically, these axioms imply that interdependence of preferences ("altruism") results from concerns for the welfare of others, as opposed to their mere payoffs, where individual welfares are captured by the reference-dependent value functions known from prospect theory. The resulting preferences are non-convex, which captures giving, sorting, and taking directly. This allows us to consistently predict choices across seminal experiments covering distributive decisions in many contexts.

2.2 Introduction

Altruism is widely defined as a concern for the well-being of others. This definition appears to be self-explanatory and seems to lend itself easily to economic modeling. Yet, any attempt at representing altruistic preferences by means of a utility function seems to prove the opposite. In a seminal paper, Andreoni and Miller (2002) showed that giving in the dictator game is well-captured by simple CES preferences—over the payoff pair of "dictator" and "recipient"—while subsequent research showed that no such utility function is compatible with giving in environments that more realistically capture distributive decisions outside laboratories. For example, if we allow the recipient to have income of her own, giving is crowded out only imperfectly (Bolton and Katok, 1998), suggesting that warm glow may affect giving (Korenok et al., 2013). Furthermore, if we allow the dictator to take from the recipient's endowment (List, 2007; Bardsley, 2008), we observe asymmetries between giving and taking, suggesting that the warm glow of giving is weaker than the cold prickle of taking (Korenok et al., 2014), a result that is incompatible with related evidence from public goods games. Or, if we allow subjects to sort out of playing a dictator game (Dana et al., 2006), around half of them do so,

in particular those who otherwise would give much to the recipient (Lazear et al., 2012), suggesting that we may need to distinguish between altruistic givers and social-pressure givers (DellaVigna et al., 2012). Overall, depending on how we extend the clinical dictator game, a different model of giving seems to be required. In turn, we currently cannot say that we have a reliable model of the simplest economic activity, giving, or a reliable representation of the interdependence of preferences governing distributive decisions. Since any economic interaction essentially boils down to some form of giving, this raises a fundamental question: Does the activity of giving not lend itself to (rigorous) economic modeling, and if so, what does?

This paper presents an axiomatic approach to the representation of preferences that allows us to directly address this potential impossibility. The advantage of an axiomatic approach is that it provides a positive result, characterizing the family of utility functions that represent all forms of preferences over payoff profiles under "plausible" assumptions (i.e., axioms). This clarifies the set of candidate models, many of which may not have been "invented" so far, and avoids the difficulties inherent in constructing a model based on evidence from a selected range of experiments—that the constructed model may be just one of many candidates. We show that in addition to the standard axioms completeness, transitivity, and continuity, three simple assumptions, namely separability, narrow bracketing, and scaling invariance, refine the vast set of candidate models surprisingly concisely to models where decision makers exhibit a concern for the welfare of others. Individuals maximize a weighted mean of "individual welfare functions" equivalent to the reference-dependent value functions known from prospect theory. We use the term "individual welfare function" to refer to an individual's utility in one-person decision problems.

Our representation result (Proposition 4) establishes that this individual welfare is the yardstick by which a person evaluates the consequences of her actions on others. Our approach provides a formal foundation for, and a formal representation of, the intuitive understanding of altruism cited above, and in the one-player case, our representation reduces to prospect-theoretic utilities, implying that it is compatible with the range of evidence on choice under risk. Both of these features stand in contrast to current models of altruism in behavioral analyses, but they relate to early models of altruism reviewed below.

In order to show how our model, which we call *welfare-based altruism*, organizes the seemingly inconsistent behavior observed in distributive decisions, we derive and test a number of theoretical predictions. We explicitly consider a large number of distributive problems including the standard dictator game, distribution games with non-trivial endowments, taking games, and sorting games. We focus on distributive decisions made by single decision makers in order to avoid confounds due to projection of preferences (as suspected in ultimatum games), coordination problems (as in public goods games), or simply irrational expectations in strategic beliefs, for discussions of which we refer to Blanco et al. (2011). Our objective is to improve our understanding of preferences over payoff profiles as such, i.e. the distributive concerns that existing results suggest to be inconsistent across similar problems, but we acknowledge that strategic interactions involve additional phenomena that hopefully will be easier to organize once concerns for distribution can be modeled more consistently.

First, we show that a dictator's optimal transfer at an interior solution decreases in her own reference point while it increases in the recipient's reference point. This explains how a reallocation of initial endowments affects the optimal transfer, by shifting the players' reference points, and predicts imperfect crowding out. Another feature inherent in welfare-based altruism is that the resulting preferences are not convex, as individual welfares are S-shaped. Non-convexity directly explains that allowing the dictator to take from the recipient's initial endowment may result in "preference reversals"; this means that a dictator whose optimal choice in a game without the possibility to take is to transfer a positive amount to the recipient may switch to taking from the recipient once this is allowed (List, 2007; Bardsley, 2008).

Relatedly, losses in relation to the reference point loom larger than gains, akin to loss aversion, explaining the asymmetries between giving and taking (Korenok et al., 2014). Welfare-based altruism also predicts the existence of "reluctant sharers", i.e. persons who transfer a positive amount to the recipient in a standard dictator game but choose a costly option to sort out of the game when given the chance (Dana et al., 2006; Lazear et al., 2012). Since the recipient never learns about the game if the dictator sorts out, her reference point is not adjusted to the dictator game environment in this case and her welfare remains neutral. Once the dictator enters the game, the recipient is informed about the scope of the interaction and forms a reference point reflecting her

expectations, which inflicts a negative externality on a welfare-based altruist. If the dictator believes the recipient would form high expectations once informed, she is best off sorting out and leaving the recipient uninformed. It is worth noting that these predictions are explicit, i.e. the opposite results are ruled out by welfare-based altruism (in a sense made precise in Proposition 6).

After demonstrating how the model theoretically predicts the debated range of stylized facts, we evaluate whether welfare-based altruism indeed captures distributive decisions in these contexts quantitatively—in sample and in particular out of sample, which allows us to assess potential overfitting. To this end, we rely on data from controlled laboratory experiments, which allow us to test models very directly, but as reviewed below, the phenomena observed in the field are very similar. Note that the concerns about overfitting we address here apply equally to all models, of course, in particular to all behavioral models generalizing the so-called standard models, regardless of whether they are models of choice, probability weighting, strategic beliefs, learning, or social preferences. Yet, outside the context of choice under risk (Harless et al., 1994; Wilcox, 2008; Hey et al., 2010), analyses quantitatively testing robustness are comparably rare[9], which has been taken as a suggestion that behavioral models may lack robustness—specifically because allowing for social preferences might allow us to explain everything. We theoretically demonstrate the opposite for our model of welfare-based altruism and econometrically demonstrate its robust fit and predictive adequacy, ruling out overfitting as a concern.

We first estimate the distributions of individual reference points in the four types of distribution games representing the corner stones of the current debate: standard dictator games, distribution games with generalized endowments, taking games, and sorting games. The estimated reference points are surprisingly consistent and we identify three clusters resembling the nongivers, altruistic givers, and social-pressure givers observed by DellaVigna et al. (2012) in charitable fundraising. Implicitly, this clarifies how this diversity of types is captured in a formally uniform manner by welfare-based altruism and that adjustments of reference points do not drive model adequacy

[9]The short list of exceptions that we are aware of comprises analyses of learning (Camerer and Ho, 1999), strategic choice in normal-form games (Camerer et al., 2004), stochastic choice in dictator games (Breitmoser, 2013, 2017), bargaining preferences (De Bruyn and Bolton, 2008), and most recently, social preferences (Bruhin et al., 2018).

across conditions, which is a promising first step.

In our second step, we re-analyze behavior across a set of nine well-known laboratory experiments on distribution games, comprising 83 choice conditions and around 6500 decisions from 981 subjects. Besides improving in-sample fit, we find that compared to the standard CES model of altruism, predictions improve substantially by allowing for welfare-based altruism, consistently across all combinations of in- and out-of-sample conditions. We then examine two alternative approaches of extending the standard CES model that are proposed in the literature, capturing either warm glow and cold prickle, or envy and guilt. They both fail to improve on CES altruism out-of-sample, confirming the general suspicion that achieving out-of-sample robustness is indeed challenging when modeling distributive concerns. To be clear, this was our initial reason to pursue an axiomatic approach based on general behavioral traits not related to unilateral giving, such as narrow bracketing and scaling invariance, but it underlines even stronger that the identified generalization of Prospect-theoretic utilities towards welfare-based altruism indeed provides a promising approach for modeling distributive concerns. Further, by unifying social preferences and risk preferences, it is a promising approach also for future work seeking to capture distributive concerns in strategic interactions.

2.3 Related literature

Similarly to us, Becker (1974) treats altruism as a concern for the utility of others, but his representation yields a linear equation system that can be solved to represent altruism again as a concern for payoffs of others. The resulting differences to standard models in games of complete information are negligible (Kritikos and Bolle, 2005). Even earlier, however, Edgeworth considered general models of altruism that contain ours as a special case, where altruism is a concern for "internal utilities" (Dufwenberg et al., 2011) of others, without the particular Prospect-theoretic formulation that we show the axioms to imply. Implicitly, our results revert us back to this classical idea, based on which, most notably, Dufwenberg et al. (2011) show that the agents behave *as-if-classical* in markets, in the sense that their demand functions depend only on own income and prices. This preserves the applicability of standard techniques and allows them to demonstrate that the Second Welfare Theorem continues to hold.

From a general perspective, our axiomatic analysis establishes a somewhat unsuspected interdependence of concepts as diverse as prospect theory, narrow bracketing, altruism, social appropriateness (discussed below), and reference dependence—besides predicting a range of behavioral puzzles that survived for about 20 years of experimental research. This underlines the adequacy of axiomatic analyses towards understanding social preferences. Further, our results imply that decision makers are utilitarists (for recent discussions, see Fleurbaey and Maniquet, 2011, and Piacquadio, 2017) but in a manner that was predicted by Rawls: rational agents "do not take an interest in one another's interests" (Rawls, 1971, p. 13). That is, agents are concerned with the welfare of others in the way that these others would perceive it in one-person decision problems, but they are not concerned with their altruism or envy, for example. This in turn provides a normative argument for "preference laundering" (Goodin, 1986) in behavioral analyses of social welfare, i.e. for the neglect of emotions such as altruism or envy in welfare analyses.

The work of Rawls also bridges our findings to "social appropriateness" of distributions as discussed by Krupka and Weber (2013). This seems to be important, as their results suggest that behavior may be norm-guided rather than payoff or welfare concerned, casting general doubts on the applicability of models (such as ours) proposed in the existing literature. In Appendix A.2.1, we show that the measure of "social appropriateness" they elicit via coordination games strongly correlates with the Rawlsian notion of social welfare as implied by our out-of-sample predictions of each player's individual welfare (it is an affine transformation of the minimum of these individual welfares). That is, we show that social appropriateness has a simple and intuitive Rawlsian foundation in individual welfare—which we interpret to lend further credibility to both, welfare-based altruism and social appropriateness, as dual approaches towards analyzing behavior.

Finally, by generalizing prospect-theoretic utility, welfare-based altruism addresses a number of practical concerns in the literature, such as providing a unified framework for measuring robustness and heterogeneity of preferences across populations and decision problems (Falk et al., 2018), providing a normatively founded framework for measuring reference points across interactions, thereby facilitating a solution to the long-lasting debate on whether and when reference points reflect a status quo (Kahneman et al., 1991), expectations (Kőszegi and Rabin, 2006b), or others' payoffs (Fehr and Schmidt,

1999), and providing a general framework for structural analyses of charitable giving (DellaVigna et al., 2012; Huck et al., 2015).

2.4 Experimental evidence on giving

We are analyzing a variety of distribution problems under complete information, each of which is more or less closely related to the classic dictator game. In each game, there are two players, the dictator and the recipient. Player 1 (dictator) is endowed with B_1 tokens and player 2 (recipient) is endowed with B_2 tokens. Player 1 can choose $p_1 \in P_1 \subset \mathbb{R}$, inducing a payoff of p_1 for herself and a payoff of $p_2(p_1) = t(B_1 + B_2 - p_1)$ for player 2. We refer to $t > 0$ as transfer rate, to $B = B_1 + B_2$ as budget, and to $B_1 - p_1$ as transfer from the dictator to the recipient.

Definition 2 (Distribution game)**.** A distribution game Γ is defined by the tuple $\langle B_1, B_2, P_1, t \rangle$. The following variants will be distinguished.

- *Standard dictator game:* $B_1 > 0$, $B_2 = 0$, $P_1 \subseteq [0, B_1]$
- *Generalized endowments:* $B_1 \geq 0$, $B_2 > 0$, $P_1 \subseteq [0, B_1]$
- *Taking game:* $B_1 \geq 0$, $B_2 > 0$, $P_1 \subseteq [0, B_1 + B_2]$
- *Sorting game:* $B_1 > 0$, $B_2 = 0$, $P_1 \subseteq \{[0, B_1], \tilde{p}_1\}$ where $\tilde{p}_1$ is an outside option for player 1 inducing payoffs $(\tilde{p}_1, 0)$, with $\tilde{p}_1 \leq B_1$, and implying that 2 is not informed about 1's choice or the rules of the game.

Table 2.1 provides an overview of the behavior observed in these distribution problems. Following the early work of Kahneman et al. (1986) and for example Hoffman et al. (1996), comprehensive analyses of behavior in the standard dictator game are presented in Andreoni and Miller (2002) and Fisman et al. (2007). The authors show that the average share of the budget transferred by dictators varies between 20% and 30%, there is an accumulation of transfers at zero and at the payoff-equalizing option, and there is considerable heterogeneity in transfers between subjects. Furthermore, varying budget sets B and transfer rates t, observed transfers to a large extent satisfy the generalized axiom of revealed preference, implying that dictator behavior is consistent with well-behaved preference orderings. As a candidate for a utility function representing these preferences, Andreoni and Miller (2002)

proposed the CES model of altruism, which, using the formulation of Cox et al. (2007), is given by

$$u(\pi) = \pi_1^\beta/\beta + \alpha\,\pi_2^\beta/\beta \qquad \text{(CES altruism)}$$

with $\alpha, \beta \in \mathbb{R}$. Here, α represents the degree of altruism, $\beta = 1$ implies efficiency concerns, $\beta \to 0$ yields Cobb-Douglas utilities, and $\beta \to -\infty$ implies equity concerns (Leontief preferences).

Comparative statics in t In a meta-analysis of about 100 experiments, Engel (2011) shows that dictators' transfers increase in the transfer rate, i.e. as transfers become more efficient. This has been observed earlier by Andreoni and Miller (2002) but, for example, not by Fisman et al. (2007). The individual level analyses of Andreoni and Miller (2002) and Fisman et al. (2007) suggest that this inconsistency may be due to differences in subject heterogeneity. In both studies, the majority of subjects act consistently with CES altruism and can be weakly categorized into three standard cases of this utility function, namely selfish, perfect substitutes, and Leontief. Perfectly selfish preferences imply no reaction to changes in the transfer rate, but dictators increase transfers after increases of t if they consider the payoffs to be imperfect substitutes ($\beta > 0$), and they decrease transfers if they consider payoffs to be imperfect complements ($\beta < 0$).

Taking options reduce giving at the extensive and intensive margin Holding initial endowments constant, convexity of preferences implies that the extension of the dictator's option set to negative transfers does not affect the choice of a dictator unless she chooses the boundary solution of giving nothing in a standard dictator game. This prediction is implied by most models of giving, including CES altruism for $\beta < 1$, but falsified by a strand of studies on so-called taking games (List, 2007; Bardsley, 2008). Both List and Bardsley found that introducing options to take reduces the share of dictators who give positive amounts, though not always significantly. Furthermore, it reduces average amounts given by those who do give positive amounts, and leads to substantive accumulation at the most selfish option. Korenok et al. (2014) confirm these results. List (2007) and Cappelen et al. (2013b) obtain related results on real-effort versions of taking games. List (2007) and Bards-

Table 2.1: Stylized facts about distribution games

Comparative statics in t	The transfer can be either constant, increasing, or decreasing in the transfer rate.
Taking options reduce giving at the extensive and intensive margin	Holding endowments constant, extending the choice set of the dictator to the taking domain transforms some initial givers into takers and reduces average amounts given.
Incomplete crowding out	Reallocating endowment from the dictator to the recipient while holding the overall budget constant leads to a less than one-to-one reduction in the dictator's transfer.
Reluctant sharers	A substantial share of givers in the standard dictator game choose to sort out of the game when given the opportunity.
Outside option attractiveness	As the outside option becomes less attractive, fewer dictators sort out of the game. Nonsharers sort back in first followed by the least generous sharers and successively more and more generous sharers.

ley (2008) interpret the observed patterns in taking games as an indication that choice is menu dependent and, for example, Korenok et al. (2014) argue that taking might induce cold prickle in the sense of Andreoni (1995). Note that we in contrast argue that the initial assumption of convexity may be violated, as known, for example, from choice under risk.

Incomplete crowding out Reference independence of social preferences, as in CES altruism, implies the so-called *crowding out hypothesis* (Bolton and Katok, 1998): lump-sum transfers from dictator to recipient result in a dollar-for-dollar reduction in voluntary giving. The experimental results on dictator games with generalized endowments unanimously falsify this prediction. In both lab and field experiments, dictators reduce their transfers in response to reallocations of endowments to the recipient, but the observed reduction is significantly lower than predicted, a phenomenon referred to as incomplete crowding out (Bolton and Katok, 1998; Eckel et al., 2005; Korenok et al., 2012, 2013). These findings extend to the domain of taking games (Korenok et al., 2014) and to interactions where the budgets are not windfall but gen-

erated through either investment games or real effort tasks (Konow, 2000; Cappelen et al., 2007, 2010, 2013a; Almås et al., 2010; Ruffle, 1998; Oxoby and Spraggon, 2008; Jakiela, 2011, 2015). The evidence on dictator games with endowments generated in real effort tasks further suggests that the origin of initial endowments affects dictator behavior. Compared to a standard dictator game with windfall budget, the change to real effort budgets earned by the dictators themselves leads to a drastic reduction in the proportion of nonzero transfers (Cherry, 2001; Cherry et al., 2002; Cherry and Shogren, 2008; Oxoby and Spraggon, 2008; Jakiela, 2011, 2015; Hoffman et al., 1994). Cappelen et al. (2007) relate the observed endowment effects to social norms and, for example, Korenok et al. (2013) interpret the endowment effects as a sign that warm glow in the sense of Andreoni (1995) affects giving. Outside the literature on social preferences, endowment effects are mostly related to reference dependence of preferences (Kahneman et al., 1991; Tversky and Kahneman, 1991), which in turn will be predicted by our representation result.

Reluctant sharers & outside option attractiveness In sorting games convexity of preferences implies that a dictator cannot be strictly better off by opting out than by staying in. For, the dictator game offers a budget that is at least as high as the outside option. Convexity also implies that no dictator who transfers a positive amount in the dictator game will opt out, since for such a dictator the outside option must be strictly worse than the allocation she chose in the dictator game. Falsifying this prediction, Dana et al. (2006), Broberg et al. (2007), and Lazear et al. (2012) find that a substantive share ($20-60\%$) of their subjects in sorting games can be classified as reluctant sharers, i.e. as dictators who transfer a positive amount in the standard dictator game but given the opportunity rather opt out. As a result, the average amount shared significantly decreases when a sorting option is added to the standard dictator game. Lazear et al. (2012) also find that (i) making the outside option less attractive while holding the dictator game budget constant does reduce the number of dictators who opt out, but (ii) it also reduces the average amount shared. For, mostly nonsharers and reluctant sharers who share less generously in the dictator game reenter first when opting out becomes less attractive. DellaVigna et al. (2012) and Andreoni et al. (2017) obtain similar results in field experiments on charitable giving. Related to that, Cap-

pelen et al. (2017) observe a close interaction between the information the recipient receives about the origin of her payment and the transfers made by the dictators (in standard dictator games). There are again multiple proposals for capturing sorting theoretically. DellaVigna et al. (2012) suggest to model it by allowing for an aversion to "saying no" when asked about donations, which however does not capture the comparative statics observed by Lazear et al. (2012), while for example Andreoni and Bernheim (2009) and Ariely et al. (2009) propose to capture reluctancy by including image concerns. As indicated above, the falsified predictions are closely related to convexity, implying that non-convexity directly predicts reluctancy and sorting decisions.

Other extensions Other interesting variations of the standard dictator game include for example the usage of double blind procedures (e.g. Hoffman et al., 1996), extensions to risky environments (e.g. Krawczyk and Le Lec, 2010; Brock et al., 2013), and variations in the transparency of the relationship between dictator choices and outcomes (e.g. Dana et al., 2007). We do not discuss those in more detail here, as these studies have not been designed to primarily study the shape of social preferences, the scope of the present paper, but rather to study the shape of preferences in relation to uncertainty and transparency.

2.5 Payoff-based and welfare-based altruism: Foundation

In this section, we aim to identify families of utility functions representing interdependent preferences under widely accepted behavioral assumptions. Our analysis is partially based on assumptions that are comparably well-accepted in related work, for example on choice under risk, which we hope contributes to the applicability of our results. Yet, the analysis also differs in important ways, most notably by distinguishing contexts to express narrow bracketing. This provides a novel foundation for reference dependence without explicitly assuming the existence of reference points, as we discuss in detail below. We are not aware of directly comparable work on the foundation of interdependent preferences, but there exist axiomatic foundations of inequity aversion, e.g. Rohde (2010) and Saito (2013), that provide insightful foundations for the widely-used model of Fehr and Schmidt (1999). The difference is that these approaches explicitly use inequity-aversion axioms to establish a foundation

for this particular model—the objective was not to identify a general set of candidate models based on axioms not directly related to altruism or giving, which we attempt here.

2.5.1 Theoretical framework

Decision maker DM has to choose an option $x \in X$ where X is a convex subset of $\mathbb{R}^n$. Each option induces an n-dimensional outcome vector captured by $\pi : X \to \mathbb{R}^n$, with $n \geq 3$.[10] We will refer to π as a payoff function, but nothing in our theoretical analysis is specific to preferences over payoff profiles. For reasons clarified soon, we also say that π defines the "context" of the decision. We use Π to denote the set of payoff functions (and thus contexts) for which the behavioral assumptions are known to hold true. The image of π is $\pi[X] = \{\pi(x)|x \in X\}$.

DM has a preference ordering over the outcomes induced by options $x \in X$ that may depend on the context π. Amongst the many forms of context dependence, this allows for preferences to be reference dependent. For example, take two contexts π and π' and two pairs of options (x,y) and (x',y') such that the associated outcomes are pairwise identical: $\pi(x) = \pi'(x')$ and $\pi(y) = \pi'(y')$. By allowing for context dependence, we allow for the possibility that DM prefers x over y in context π but y' over x' in context π'. For example, outcomes below reference points may be ordered differently than outcomes above reference points. As reference points may change when the set of feasible outcomes (or, the context) changes, orderings of outcomes that are identical in absolute terms may change as a function of the context π. With our notation, we explicitly allow for such effects and any other form of context dependence.

Formally, the preference ordering on outcomes $\pi[X]$ is denoted as $\succsim_\pi$, with $\pi(x) \succsim_\pi \pi(y)$ indicating that outcome $\pi(x)$ is weakly preferred to outcome $\pi(y)$ in context π. Given π and $\succsim_\pi$, DM's preference relation R over option set X is straightforwardly defined as xRy if and only if $\pi(x) \succsim_\pi \pi(y)$, for all $x, y \in X$. As usual, the strict preference $\pi(x) \succ_\pi \pi(y)$ indicates $\pi(x) \succsim_\pi \pi(y)$ and

[10] Assuming the outcome vector has at least three dimensions simplifies some of the statements made below regarding existence of an additively separable utility representation. It is not crucial for the main result. If there was only one essential dimension, existence of an additively separable representation would be trivial, and if there were exactly two essential dimensions in the outcome vector, then existence of an additively separable representation would be ensured by additionally assuming the hexagon condition of Wakker (1989, p. 47).

$\pi(y) \not\succsim_\pi \pi(x)$. Given this notation, we impose the following assumptions.[11]

Assumption 4 (Framework)**.**

1. *Translatability:* $\pi, \pi' \in \Pi \Leftrightarrow$ there exists $c \in \mathbb{R}^n$ such that $\pi' = c + \pi$
2. *Outcome image is a cone:* $\pi[X]$ is a cone in $\mathbb{R}^n$, i.e. for all $x \in X$ and all $\lambda \in (0,1)$, there exists $x' \in X$ such that $\pi(x') = \lambda \pi(x)$
3. *Essentialness:* All $n \geq 3$ dimensions are essential, i.e. for all $i \leq n$ and each $\pi \in \Pi$, there exist $p, p' \in \pi[X]$ such that $p \succ_\pi p'$ with $p_{-i} = p'_{-i}$.

First, different payoff functions π and π' differ only by translation, i.e. by addition of constants to all outcome vectors, and in turn, all translations are possible. We refer to these additive constants as the "background income" vector, and the background income implicitly characterizes the context of the decision problem. Distinguishing such contexts is novel in relation to the literature and will allow us to state assumptions about responses to changes in background income or concurrent tasks, as discussed below. Second, the image of the set of options in the outcome space is a cone, i.e. we can think of X as a budget set of a consumer facing linear prices: for any $\lambda \in (0,1)$ and any option $x \in X$, an option x' is available satisfying $\pi(x') = \lambda \pi(x)$. This assumption implies that the set of options is rich, in the sense that the set of possible outcomes $\pi[X]$ has positive volume in $\mathbb{R}^n$, which helps us to establish uniqueness of the utility representation. Finally, essentialness requires that there are no redundant dimensions of the outcome vector from DM's perspective, i.e. DM does not ignore any of the dimensions, which is a necessary condition for uniqueness of the utility representation in all dimensions as well.

2.5.2 Axiomatic foundation of payoff-based and welfare-based altruism

We analyze the interplay of six axioms. The first two require that $\succsim_\pi$ is a continuous weak order, implying that it can be represented by a utility function. Separability (Axiom 3) ensures that an additively separable utility representation exists: if two options are equivalent in any dimension, then changing the value in this dimension equally for both options does not affect the preference

[11] Slightly abusing notation, we identify all $c \in \mathbb{R}^n$ with constant functions so the addition of functions and constants is well defined, i.e. for all $\pi, \pi' \in \Pi$, if $\pi' = \pi + c$ then $\pi'(x) = \pi(x) + c$ for all $x \in X$.

ordering between these options. Axioms 4–6 jointly define the functional form. Here, we will analyze the implications of scaling invariance (Axiom 4) in conjunction with either broad bracketing (Axiom 5) or narrow bracketing (Axiom 6).

Assumption 5 (Axioms)**.** For all $\pi \in \Pi$ and all $x, y \in X$:

1. *Weak order:* $\succsim_\pi$ is complete and transitive.

2. *Continuity:* If $\pi(x) \succ_\pi \pi(y)$, there exists $\varepsilon > 0$ such that $\pi(x') \succ_\pi \pi(y')$ for all $x' : d(x', x) < \varepsilon$ and all $y' : d(y', y) < \varepsilon$.

3. *Separability:* For any $x', y' \in X$ such that $\pi_{-i}(x) = \pi_{-i}(x')$ and $\pi_{-i}(y) = \pi_{-i}(y')$, as well as $\pi_i(x) = \pi_i(y)$ and $\pi_i(x') = \pi_i(y')$, we have $\pi(x) \succsim_\pi \pi(y)$ iff $\pi(x') \succsim_\pi \pi(y')$.

4. *Scaling invariance:* There exists a scaling-invariant context $\pi^0 \in \Pi$, i.e. for any $\lambda \in \mathbb{R} : \lambda > 0$, if $\pi^0(x) = \lambda \pi^0(x')$ and $\pi^0(y) = \lambda \pi^0(y')$, then $\pi^0(x) \succsim_{\pi^0} \pi^0(y) \Leftrightarrow \pi^0(x') \succsim_{\pi^0} \pi^0(y')$.

5. *Broad bracketing:* For any $\pi' \in \Pi$, if $\pi(x) = \pi'(x')$ and $\pi(y) = \pi'(y')$, then $\pi(x) \succsim_\pi \pi(y)$ implies $\pi'(x') \succsim_{\pi'} \pi'(y')$.

6. *Narrow bracketing:* For all $c \in \mathbb{R}^n$, $\pi(x) \succsim_\pi \pi(y)$ implies $(\pi + c)(x) \succsim_{\pi+c} (\pi + c)(y)$.

Briefly, let us discuss Axioms 3–6 to clarify to which extent they relate to standard or perhaps reasonable assumptions. Separability is also known as "independence of equal coordinates" (Wakker, 1989, p. 30). It implies additive separability of the utility function and closely relates to a broad range of standard assumptions: independence axioms in choice under risk (Wakker and Zank, 2002) or choice under uncertainty (Skiadas, 2013), "independence of irrelevant alternatives" in stochastic choice (Luce, 1959), and separability in social welfare functions (Piacquadio, 2017). Further, additive separability obtains in most utility representations discussed in the literature on altruistic giving, such as CES altruism (Andreoni and Miller, 2002), efficiency concerns (Charness and Rabin, 2002), and impure altruism (Andreoni, 1990; Korenok et al., 2013).[12]

[12] Violations of separability obtain in models of inequity aversion (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000).

Scaling invariance requires that DM's preferences over two options are robust to scaling the outcome vectors associated with these options. It implies that the utility function is homothetic, which again is satisfied by a broad range of utility functions discussed in the behavioral literature, including CES altruism, inequity aversion, prospect theoretical utilities, and nested CES functions. Scaling invariance is further supported by neuro-physiological evidence showing that the neural firing rate adapts to the scale of the choice problem (Padoa-Schioppa and Rustichini, 2014)[13] and a host of meta analyses showing that scaling differences between experiments are indeed choice-irrelevant overall. This applies in dictator games (Engel, 2011), ultimatum games (Oosterbeek et al., 2004; Cooper and Dutcher, 2011), trust games (Johnson and Mislin, 2011), and choice under risk (Wilcox, 2008, 2011, 2015).

Finally, broad and narrow bracketing describe behavior in response to changes in "context", which in our case are changes in background income $c \in \mathbb{R}^n$. Broad bracketing assumes that background income is fully factored in when decisions are made, while narrow bracketing assumes that background income is factored out. There is fairly strong evidence, from two strands of literature, that background income is indeed factored out. On the one hand, behavior was shown to be independent of socio-economic background variables such as income or wealth in experiments (Gächter et al., 2004; Bellemare et al., 2008, 2011) and in general (Easterlin, 2001). In conjunction with the wide range of results supporting narrow bracketing more generally (e.g. Read et al., 1999b, Rabin and Weizsäcker, 2009, Simonsohn and Gino, 2013), this suggests that narrow bracketing is a substantially more adequate behavioral assumption than broad bracketing.

Adaptive coding describes the neuro-economic observation that the neuronal representation of subjective values ("utilities") adapts to the range of values in any environment. This enables efficient adaptation to choice environments subject to the physical limitations in neuronal firing rates, and was first observed by Tremblay and Schultz (1999) and subsequently confirmed in a wide range of studies reviewed for example in Padoa-Schioppa and Rustichini (2014) and Camerer et al. (2017). Specifically, Padoa-Schioppa (2009) showed that the baseline activity of the cell encoding the value of a given

[13]The best option always has the maximal firing rate and the worst option always has the minimal firing rate, implying that choice is independent of scale after a transition period where the neural firing rate adapts to the scale of the decision problem. See Camerer et al. (2017) for a recent review of the evidence.

object generally represents the minimum of the value range, and the upper bound of the activity range of this "value cell" represents the upper bound of the value range. Implicitly, both the scale of the value range and background utility is factored out, yielding choice that satisfies scaling invariance and narrow bracketing simply as a result of the physical limitations in neuronal firing.

As usual, we say that a preference relation $\succsim_\pi$ is represented by a utility function $u_\pi : X \to \mathbb{R}$ if $\pi(x) \succsim_\pi \pi(y) \Leftrightarrow u_\pi(x) \geq u_\pi(y)$ for all $x, y \in X$. Proposition 4 establishes that, in conjunction with the other axioms, preferences compatible with broad bracketing are represented by CES altruism ("payoff-based altruism") and preferences compatible with narrow bracketing are represented by generalized prospect theoretical preferences ("welfare-based altruism").

Proposition 4. *Given Assumption 4, there exist* $\alpha \in \mathbb{R}^n$, $\beta \in \mathbb{R}$, $\delta \in \mathbb{R}^n$ *and* $r : \Pi \to \mathbb{R}^n$ *such that for all contexts* $\pi \in \Pi$,

$$\textit{Axioms 1,2,3,4,5} \quad \Leftrightarrow \quad \succsim_\pi \textit{ is represented by } \; u_\pi(x') = \sum_{i \leq n} \alpha_i \cdot v_i\big[\pi_i(x')\big],$$

$$\textit{Axioms 1,2,3,4,6} \quad \Leftrightarrow \quad \succsim_\pi \textit{ is represented by } \; u_\pi(x') = \sum_{i \leq n} \alpha_i \cdot v_i\big[\pi_i(x') - r_i(\pi)\big],$$

for all $x' \in X$, *where the reference points satisfy* $r(\pi^0 + c) = c$ *for all* $c \in \mathbb{R}^n$ *given the scaling invariant context* π^0, *and the value functions* $v_i : \mathbb{R} \to \mathbb{R}$ *satisfy*

$$v_i(p) \underset{\beta \neq 0}{=} \begin{cases} p^\beta/\beta, & \text{if } p \geq 0 \\ -\delta_i \cdot (-p)^\beta/\beta, & \text{if } p < 0 \end{cases} \qquad \text{and} \qquad v_i(p) \underset{\beta = 0}{=} \log(p).$$

The proof is relegated to the appendix. Existence of a continuous weak order (Axioms 1 and 2) implies that, in each context π, the preference relation $\succsim_\pi$ can be represented by some utility function $u_\pi : X \to \mathbb{R}$. Axiom 3 implies that an additively separable utility representation exists (Wakker, 1989), i.e. given context π, value functions $\{v_{\pi,i} : \mathbb{R} \to \mathbb{R}\}_{i \leq n}$ exist such that u_π with

$$u_\pi(x) = \sum_{i \leq n} v_{\pi,i}\big(\pi_i(x)\big) \tag{8}$$

represents $\succsim_\pi$. Broad bracketing implies that these value functions must be equivalent across contexts. Narrow bracketing requires that context shifts (changes in background income) are factored out, which implies that pay-

offs must be evaluated in relation to some unknown reference points. As a result, there exists a family of functions $\{v_i : \mathbb{R} \to \mathbb{R}\}_{i \leq n}$ and $r : \Pi \to \mathbb{R}^n$ such that

$$u_\pi(x) = \sum_{i \leq n} v_i\big(\pi_i(x)\big) \qquad\qquad u_\pi(x) = \sum_{i \leq n} v_i\big(\pi_i(x) - r_i(\pi)\big) \tag{9}$$

represent $\succsim_\pi$ for all $\pi \in \Pi$ in the cases of broad bracketing and narrow bracketing, respectively. With narrow bracketing, the utility function is equivalently expressed as

$$u_\pi(x) = \sum_{i \leq n} \alpha_i \cdot \big[r_i(\pi) + v_i\big(\pi_i(x) - r_i(\pi)\big)\big], \tag{10}$$

simply adding the reference points in all dimensions (or any other constant; given separability, the utility function is unique up to positive affine transformation). Formulation (10) may appear more intuitive if the reference points differ from zero.[14] Similarly, by uniqueness up to affine transformation, the weights (α_i) are unique up to scaling. A standard restriction here is to require that (α_i) adds up to 1. Finally, scaling invariance pins down the functional form of v_i. By scaling invariance, we know that, focusing on broad bracketing for simplicity here,

$$u_\pi(x) = \sum_{i \leq n} v_i\big(\pi_i(x)\big) \qquad \text{and} \qquad u_{\lambda\pi}(x) = \sum_{i \leq n} v_i\big(\lambda\pi_i(x)\big)$$

with $\lambda \in (0,1)$ both represent $\succsim_\pi$, and both being additively separable, this implies that they are positive affine transformations of one another. Hence, for all $i \leq n$,

$$v_i\big(\lambda\pi_i(x)\big) = a_i(\lambda) + b(\lambda) \cdot v_i\big(\pi_i(x)\big)$$

for some functions $a_i : \mathbb{R} \to \mathbb{R}$ and $b : \mathbb{R} \to \mathbb{R}_+$. By Assumption 4.2, the value function v_i is defined on an interval of positive length, by Axiom 2 it is continuous, and by 4.3 it is not equal to the constant function, which jointly implies that the unique solutions of this Pexider functional equation (Aczél,

[14] It expresses the idea that meeting one's reference point implies a utility exactly equal to the reference point (in case the value function is the power function in Proposition 4). Thus, for example, an individual being \$10 short of their reference point \$1,000,000 would enjoy a higher utility than an individual being \$10 short of their reference point \$20.

1966) are the power and logarithmic functions defined in Proposition 4.[15]

Two technical points appear worth noting. If there exists $x \in X$ such that $\pi^0(x) = 0$, where π^0 denotes the scaling-invariant context, then $\beta > 0$ obtains by continuity. Further, if we assume monotonicity, the parameters (α_i, δ_i) are guaranteed to be non-negative. While this appears plausible in many cases, it would rule out some phenomena resembling inequity aversion, the defining characteristic of which is that preferences are non-monotonic in the opponents' outcomes.[16]

2.5.3 Discussion

To summarize, broad bracketing induces a context-independent reference point of zero, yielding the well-known CES model of altruism in which payoffs are evaluated in absolute terms. This captures altruism as concern for the payoffs of others. Narrow bracketing implies that payoffs are evaluated in relation to reference points $r_i(\pi) = \pi_i(x) - \pi_i^0(x)$, where π^0 is the scaling invariant context existing by Axiom 4. This implies that altruism is a concern for the S-shaped welfares of others known from prospect theory, i.e. for the (individual) welfares they believe the others would derive from the various outcomes in single-person decision problems. Switching from broad bracketing to narrow bracketing is in this sense equivalent to switching from altruism as a concern for the payoff of others to altruism as a concern for the welfare of others.

Our analysis is related to studies of preferences in choice under risk, see for example Wakker and Tversky (1993) and Skiadas (2013). Most axioms in this branch of literature are similar to those imposed above, suggesting the possibility of constructing a general, unified foundation of behavior. The main difference of our analysis to this literature is the formal distinction of contexts. This is substantial, as it allows us to analyze narrow bracketing instead of translation invariance which endogenously yields reference depen-

[15] For illustrative purposes, assume v_i is also differentiable and let $a_i = 0$ (which removes the logarithmic solution). That is, $v_i(\lambda\,\pi_i) = b(\lambda) \cdot v_i(\pi_i)$, and after taking logarithms on both sides, we obtain for $\tilde{v}_i = \log v_i$ and $\tilde{b} = \log b$,

$$\tilde{v}_i(\lambda\,\pi_i) = \tilde{b}(\lambda) + \tilde{v}_i(\pi_i) \qquad \Rightarrow \qquad \tilde{v}_i'(\lambda\,\pi_i) \cdot \pi_i = \tilde{b}'(\lambda) \qquad \Rightarrow \qquad \tilde{v}_i'(\pi_i) = \beta/\pi_i$$

after taking the derivative with respect to λ and letting $\lambda = 1$. This differential equation has the solution $\tilde{v}_i(\pi_i) = \beta \log \pi_i + \alpha_i$ and reverting the logarithm we obtain $v_i(\pi_i) = \alpha_i \cdot \pi_i^{\beta}$.

[16] For example, inequity averse subjects prefer $(10,9)$ over $(11,20)$, or $(0,0)$ over $(1,9)$. Without monotonicity, welfare-based altruism can capture such preferences, and in this way, it can also capture rejections in ultimatum games.

dence, as discussed in the next paragraphs. The similarities are that analyses of choice under risk also work with existence of a weak order, continuity, and generally an independence assumption yielding additive separability across possible outcomes. Skiadas (2016) shows that a system of axioms including scaling invariance implies a form of CES preferences that is similar to CES altruism as characterized above, i.e. not to the reference-dependent (welfare-based) altruism, while one including translation invariance implies exponential rather than power utilities resembling constant absolute risk aversion. His results suggest that scaling invariance and translation invariance are mutually exclusive in axiomatic foundations, although both tend to be confirmed in behavioral meta studies, a conflict that is resolved with our context-based approach.

Following Read et al. (1999b), narrow bracketing refers to the phenomenon that concurrent decision problems are treated independently by decision makers, implying that other tasks simply provide a background income (the "context") that is factored out. Using this observation as part of an axiomatic foundation is novel and to be distinguished from translation invariance. Specifically, narrow bracketing operates between contexts (changes in background income) and translation invariance operates within contexts.[17] The distinction is noteworthy, as it is narrow bracketing, rather than translation invariance, that is backed by the behavioral and neuroeconomic evidence cited above.

Narrow bracketing implies reference dependence and the existence of reference points with the testable prediction that reference points move 1:1 as the background income changes. This result is substantial, as it generalizes existing axiomatic foundations of prospect theoretical utilities, which so far explicitly assume existence of a reference point, where the reference point is either an exogenously defined payoff vector (Wakker and Tversky, 1993; Wakker and Zank, 2002) or a well-defined option (Schmidt, 2003). Further, we link narrow bracketing and reference dependence based on axioms not related specifically to altruism or giving, underlining the link's generality and corroborating the observation that both narrow bracketing and reference dependence build on a wealth of empirical evidence (outside prospect theory, see for example Kőszegi and Rabin, 2007, 2009, for discussion). The refer-

[17] Translation invariance requires that if one pair of options yields payoff vectors π_x and π_y, and another pair of options yields $\pi_x + r$ and $\pi_y + r$ for some $r \in \mathbb{R}$, then the respective choices must be equivalent (see for example Skiadas, 2013). In contrast, narrow bracketing poses no restriction for different pairs of options.

ence points may reflect any (weighted) mean of status quo (Kahneman et al., 1991) and expectations (Kőszegi and Rabin, 2006b), since any such mean satisfies the above condition that adding a fixed vector of background incomes to all payoff profiles raises the vector of reference points by exactly that vector (assuming expectations are independent of the background income). Our model does not further restrict reference points, implying that different decision makers may have different reference points.

We have not imposed assumptions linking preferences across contexts in general, implying that preferences may change arbitrarily when say the image $\pi[X]$ changes in dimensions other than the background income. The existing literature analyzes how reference points may depend on say $\pi[X]$, usually in the form of expectation-based reference points as proposed by Kőszegi and Rabin (2006b), Falk et al. (2011) and Gill and Prowse (2012), or as a foundation of reciprocity, see Rabin (1993), Dufwenberg and Kirchsteiger (2004), and Falk and Fischbacher (2006). Implicitly, by allowing for expectation-based reference dependence, our simple model of altruism is compatible with these models of reciprocity, in particular because the "context" may be a function of previous moves by other players.

2.6 Implications for giving: Theory

In this section, we characterize the distributive decisions made by welfare-based altruists and analyze how they relate to the observations made in experiments. By context dependence, the reference points of dictator and recipient, r_1 and r_2, may be arbitrary functions of the game $\Gamma = \langle B_1, B_2, P_1, t \rangle$, but we know there exist reference points r_1 and r_2 such that the dictator's utility in a distribution game Γ is

$$u_\Gamma(p_1) = \frac{1}{\beta} \times \left\{ \begin{array}{ll} (p_1 - r_1)^\beta & \text{if } p_1 \geq r_1 \\ -\delta(r_1 - p_1)^\beta & \text{if } p_1 < r_1 \end{array} \right\} + \frac{\alpha}{\beta} \times \left\{ \begin{array}{ll} (t(B - p_1) - r_2)^\beta & \text{if } p_2(p_1) \geq r_2 \\ -\delta(r_2 - t(B - p_1))^\beta & \text{if } p_2(p_1) < r_2 \end{array} \right\}.$$

As above, α represents the degree of altruism, δ is the degree of loss-aversion, and β captures the trade-off between efficiency and equity concerns; $\frac{1}{1+\beta}$ is the elasticity of substitution between dictator's and recipient's well-being. With-

out loss of generality, we assume that δ is the same for both players, and for notational simplicity, we skip the limiting case $\beta = 0$.

The reference points contain the players' background incomes as additive constants by Proposition 4. We represent the background incomes by players' minimal payoffs, $\min p_1$ and $\min p_2$. Otherwise, the reference points may be arbitrarily complex functions of the game characteristics such as payoff function and option sets, which could be refined using additional axioms but is not the purpose of the present analysis. Its purposes are to clarify the intuition underlying choices made by welfare-based altruists and to demonstrate that the predictions are robust to variations in the context-dependence of reference points. We believe these two purposes are balanced well using a two-parametric family of functions describing how reference points change depending on the range of payoffs.

Specifically, each player's reference point is her minimal payoff $\min p_i$ ("background income") plus share $w_1 \in [0,1]$ of the amount she contributes to the cake to be redistributed ($B_i - \min p_i$) and share $w_2 \in [0,1]$ of the amount her partner contributes to the cake ($B_j - \min p_j$). We assume $w_1 \geq w_2$, i.e. that each player believes to be weakly more entitled to get some share of her contribution than of her partner's contribution.

Assumption 6. In game $\Gamma = \langle B_1, B_2, P_1, t \rangle$, the reference points satisfy, for some $w_1, w_2 \in [0,1]$ with $w_1 \geq w_2$,

$$r_1 = \min p_1 + w_1 \cdot (B_1 - \min p_1) + w_2 \cdot (B_2 - \min p_2/t),$$
$$r_2 = \min p_2 + w_2 \cdot t \cdot (B_1 - \min p_1) + w_1 \cdot (t \cdot B_2 - \min p_2).$$

This model contains status-quo-based reference points ($w_1 = w_2 = 0$) and strict expectations-based reference points ($w_1 + w_2 = 1$) as the most notable special cases, and by allowing for $w_1 + w_2 \in (0,1)$ all convex combinations are also included. Furthermore, the weights w_1 and w_2 have interpretations in light of different fairness ideals discussed in the literature (Konow, 2000; Cappelen et al., 2007). While the weight on the own contribution to the budget w_1 can be interpreted as measuring the degree to which a welfare-based altruist agrees with the libertarian fairness ideal demanding no redistribution ($w_1 = 1$, $w_2 = 0$), the weight on the other's contribution to the budget w_2 is a measure for the degree to which a welfare-based altruist agrees with the egalitarian fairness ideal demanding full redistribution eliminating all payoff

differences ($w_1 = w_2 = 0.5$). In the following we will therefore also refer to w_2 as the degree to which a welfare-based altruist feels social pressure to redistribute.

Dictators are welfare-based altruists denoted as $\Delta = (\alpha, \beta, \delta, w_1, w_2)$. Besides satisfiability of reference points ($w_1 + w_2 \leq 1$), we assume that dictators are imperfectly altruistic ($0 \leq \alpha \leq 1$), imperfectly efficiency concerned ($0 < \beta < 1$), and weakly loss averse ($\delta \geq 1$). Both $0 < \beta < 1$ and $\delta \geq 1$ are standard assumptions in, for example, prospect theoretical analyses, ensuring S-shaped utilities and avoiding loss seeking, which we therefore adopt as well. Weak altruism ($\alpha \leq 1$) is a standard assumption in analyses of social preferences and $\alpha \geq 0$ is assumed without loss of generality as egoism ($\alpha = 0$) is equivalent to spite ($\alpha < 0$) in the games we analyze.

Definition 3. Dictator $\Delta = (\alpha, \beta, \delta, w_1, w_2)$ is called **regular** if she exhibits imperfect altruism ($0 \leq \alpha \leq 1$), weak efficiency concerns ($0 < \beta < 1$), loss aversion ($\delta \geq 1$), and satisfiability ($w_1 + w_2 \leq 1$).

Proposition 5 formally characterizes giving of welfare-based altruists to provide the basic intuition. Our subsequent result will explore the relations to the stylized facts discussed above.

Proposition 5. *In a given distribution game Γ almost all regular dictators Δ can be classified as follows. Dictators with $\alpha^{1/\beta} < 1/t$ choose*

$$(p_1^*, p_2^*) = \begin{cases} \left(\frac{tB + c_\alpha r_1 - r_2/t}{c_\alpha + 1}, \frac{t\, c_\alpha (B - r_1) + r_2}{c_\alpha + 1} \right), & \text{if } \delta > \delta^+ \quad \textbf{(interior solution)} \\ (\max p_1, \min p_2), & \text{if } \delta < \delta^+ \quad \textbf{(egoistic solution)} \end{cases}$$

while dictators with $\alpha^{1/\beta} > 1/t$ choose

$$(p_1^*, p_2^*) = \begin{cases} \left(\frac{tB + c_\alpha r_1 - r_2/t}{c_\alpha + 1}, \frac{t\, c_\alpha (B - r_1) + r_2}{c_\alpha + 1} \right), & \text{if } \delta > \delta^- \quad \textbf{(interior solution)} \\ (\min p_1, \max p_2), & \text{if } \delta < \delta^- \quad \textbf{(altruistic solution)} \end{cases}$$

with

$$\delta^+ := c_\alpha^{\beta - 1} \left(\left(\frac{t(B - r_1)}{r_2} \right)^\beta - (c_\alpha + 1)^{1-\beta} \left(\frac{t(B - r_1)}{r_2} - 1 \right)^\beta \right)$$

$$\delta^- := c_\alpha^{1-\beta} \left(\frac{tB - r_2}{t r_1} \right)^\beta - (c_\alpha + 1)^{1-\beta} \left(\frac{tB - r_2}{t r_1} - 1 \right)^\beta$$

and $c_\alpha := (\alpha t^\beta)^{\frac{1}{1-\beta}}$.

That is, there are up to three types of welfare-based altruists: some give nothing or take all (choosing the lower bound), some give a bit (choosing an interior solution), and some give all (choosing the upper bound). In the interior solution, both reference points are satisfied, which implies that many possible decisions can be ruled out. Further, the types of welfare-based altruists are defined using simple thresholds (δ^-, δ^+) in terms of the degree of loss aversion δ. This allows us to rank dictators by their propensity to choose either of the corner solutions. While dictators with a low degree of loss aversion δ tend to have a high propensity to choose a corner solution, evaluating the extra costs of not satisfying a reference point as low, dictators with a high degree of loss aversion tend to pick an interior solution. The type of corner solution chosen by dictators with low δ depends on their degree of altruism α, the welfare function curvature β, and the transfer efficiency t. The altruistic corner solution becomes relevant only in games with efficiency gains from giving ($t > 1$) for dictators who have relatively high altruism weights α and/or strong efficiency concerns (high β). The thresholds (δ^+, δ^-) for actually choosing either corner solution when it is relevant have intuitive comparative statics in the preference parameters. The higher the altruism weight α, the lower the maximum δ for which the egoistic corner solution is chosen and the higher the maximum δ for which the altruistic corner solution is chosen. The stronger the dictator's efficiency concerns (the higher β), the higher the maximum δ for which an efficient corner solution is chosen and the lower the maximum δ for which an inefficient corner solution is chosen.

Since the interior solution and the thresholds δ^+ and δ^- are continuous in the game parameters $\langle B_1, B_2, P_1, t\rangle$ we can also characterize the comparative statics of behavior across different distribution games. The interior solution has very intuitive comparative statics in this respect: The recipient's payoff is decreasing in the dictator's reference point r_1, increasing in the recipient's reference point r_2 and budget B, and increasing in the transfer rate t. In conjunction with the similarly intuitive comparative statics of the thresholds δ^+ and δ^-, this directly predicts the stylized facts observed in the literature (Table 2.1). Proposition 6 establishes this formally, a detailed discussion follows. As above, we say a dictator is a "giver" if she transfers some of her endowment to the recipient, she is a "taker" if the net-transfer is negative, and comparing

two games, we say that the range of taking options is extended if $B = B_1 + B_2$ is held constant but the maximal dictator transfer $\max p_1$ increases.

Proposition 6. *Assume dictators $\Delta = (\alpha, \beta, \delta, w_1, w_2)$ are randomly distributed in $\mathbb{R}^5$ such that dictator Δ has positive density if and only if dictator Δ is regular. All "stylized facts" are implied.*

1. ***Non-convexity*** *In all games with $P_1 = [0, B]$, some dictators have non-convex preferences.*

2. ***Taking options reduce giving both at the extensive and intensive margin*** *Introducing a taking option turns some initial givers into takers and reduces average amounts given.*

3. ***Incomplete crowding out*** *Reallocating initial endowment from the dictator to the recipient results (in expectation) in a payoff increase for the recipient.*

4. ***Efficiency concerns*** *The recipient's payoff is weakly increasing in the transfer rate.*

5. ***Reluctant sharers*** *When an outside option is introduced, some initial givers switch to that option while the behavior of dictators who sort into the game stays unaffected.*

6. ***Social pressure givers*** *Ceteris paribus, higher susceptibility to social pressure (higher w_2) implies higher transfers in the interior solution but also a higher propensity to choose the outside option in a sorting game.*

Givers who become takers: Non-convexity of preferences One of the most distinctive characteristics of welfare-based altruism is the implied non-convexity of preferences that has important consequences for its theoretical predictions across distribution games that differ in the dictator's choice set, in particular comparing games with generalized endowments to taking games. The nature of this non-convexity and its consequences are illustrated in Figure 2.1.

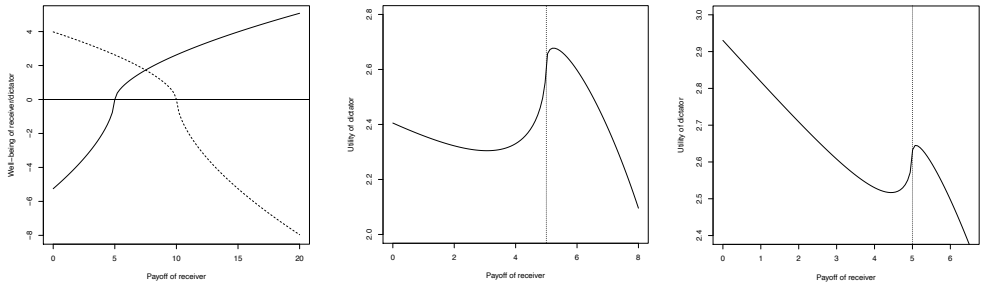
We consider a distribution game in which the dictator is asked to allocate a budget of 20 tokens between herself and the recipient at a transfer rate of $t = 1$.

Figure 2.1: Non-convexity of preferences and implications in taking games

(a) Welfares of recipient (solid) and dictator (dashed) with $\beta = 0.6$ and $\delta = 2$

(b) Utility of a dictator with $\alpha = 0.3$

(c) Utility of a dictator with $\alpha = 0.2$



Note: The dictator can choose to allocate x tokens to the recipient, where $x \in [0,20]$. The transfer rate is $t = 1$. The recipient's reference point is $r_2 = 5$ while the dictator's reference point is $r_1 = 10$. The dashed lines in (b) and (c) mark the recipient's reference point.

Suppose that the reference points are $r_2 = 5$ for the recipient and $r_1 = 10$ for the dictator. Figure 2.1a depicts the trade-off that the dictator faces between her own and the recipient's welfare. The more the dictator allocates to the recipient, the higher is the recipient's welfare (solid curve) but the lower is the dictator's own welfare (dashed curve). The individual welfares are steeper the closer the players are to their respective reference points. For recipient payoffs between 5 and 10, both the recipient and the dictator are in the gain domain, i.e. they achieve payoffs at least as high as their respective reference points, whereas for all other allocations one of them is in the loss domain. Figure 2.1b depicts the dictator's utility if her weight on the recipient's welfare is $\alpha = 0.3$. This dictator's utility function reaches its maximum at the interior solution where the transfer slightly exceeds the recipient's reference point. Figure 2.1c depicts the utility of a slightly less altruistic dictator ($\alpha = 0.2$). This dictator's optimal choice is the egoistic (corner) solution of allocating nothing to the recipient.

The S-shaped form of the individual welfare function implies that the deeper the recipient moves into the loss domain, the lower the marginal reduction in recipient welfare for any further token not allocated to him. In conjunction with weak altruism and the correspondingly S-shaped dictator welfare, this implies that dictator utility is not quasi-concave—it bends upwards once the recipient is sufficiently far below his reference point. Ceteris paribus, the lower the weight α that the dictator assigns to the recipient's wel-

fare, the earlier this minimum is reached and the more likely it is that the dictator's utility from choosing the lower bound exceeds her utility in the interior solution. As a result, dictator behavior is not generally continuous in the game parameters, which predicts the "preference reversals" observed in taking games.

To see this, have another look at Figure 2.1c, now assuming the recipient's reference point equates with his endowment ($B_2 = 5$ and $B_1 = 15$). That is, if the dictator allocates, say, 4 to the recipient, then she actually takes from his endowment. For simplicity, also assume that reference points are invariant to changes in the dictator's choice set (reference point movements are covered in the subsequent discussion) and suppose the dictator cannot take from the recipient's endowment. In this case, the dictator cannot implement an allocation with a recipient payoff below his reference point, to the left of the vertical dotted line in Figure 2.1c, and chooses the interior solution to the right. Now, as we extend the option set by allowing for taking one token from the recipient, allocations to the left of the vertical dotted line become admissible. Initially, upon extending the option set, the dictator's utility at the lower bound is decreasing. The recipient's welfare drops sharply and the dictator is concerned for his welfare. Upon further extending the option set into the taking domain, the dictator's utility reaches a minimum and starts to increase again. Eventually, the dictator prefers the lower bound to the interior solution and jumps to taking as much as possible. Such a "preference reversal" cannot be observed for the more altruistic dictator in Figure 2.1b as long as the recipient's payoff is restricted to be non-negative.

Decreasing the lower bound decreases expectations: Taking options Introducing a taking option decreases the recipient's minimal payoff, i.e. his background income. Regardless of whether the recipient has status-quo-based or expectations-based reference points, or a convex combination from the general class in Assumption 6, the recipient's reference point will consequentially decline. The reduction in the recipient's minimal payoff at the same time raises the surplus $B_2 - \min p_2/t$ he contributes, but generically (for all $w_1 < 1$) the first effect dominates. Loosely speaking, the recipient will be happy with less. In turn, the dictator's reference point weakly increases through her partial claim to the increasing surplus contributed by the recipient (if $w_2 > 0$). That is, after introducing taking options, the dictator asks for more. Both ef-

fects directly imply, at the intensive margin, that the dictator transfers less in the interior solution, which has the obvious comparative statics in reference points by Proposition 5. In addition, as the lower bound declines, defecting towards the lower bound becomes more attractive for the dictator (recall Figure 2.1) and with the increase of the own reference point, the interior solution becomes less attractive. As a result, at the extensive margin, dictators are more likely to pick the lower bound, and across the population, the share of regular dictators who choose the lower bound increases while the share of regular dictators who choose the interior solution decreases.

Shifting surplus claims: Generalized endowments Assume part of the dictator's endowment is reallocated to the recipient and the dictator cannot take any of it back, i.e. her budget correspondingly declines. Then, the dictator's background income is constant but the surplus she contributes ($B_1 - \min p_1$) decreases, while the recipient's background income increases and his surplus remains constant. As a result, the dictator's reference point declines and the recipient's reference point increases. By the comparative statics of the interior solution, the dictator thus allocates less to herself and more to the recipient at the interior solution, implying incomplete crowding out of endowment reallocations.

Avoiding high recipient expectations: Sorting options Lazear et al. (2012) call a dictator a "reluctant sharer" if she transfers a positive amount in a standard dictator game but sorts out when possible. That is, her utility from the interior solution is lower than her utility from the outside option $(\tilde{p}_1, 0)$—assuming the recipient is not informed about the dictator and her options if she sorts out. Remaining uninformed if the dictator sorts out, from the recipient's perspective literally nothing happens, both reference point and payoff are zero, and he remains welfare neutral. This removes the negative externality imposed by the recipient's expectations and may therefore be preferable for the dictator. To see this, assume reference points are just "satisfiable", i.e. $B = r_1 + r_2/t$, and the dictator chooses to satisfy them in the standard dictator game (as opposed to choosing the lower bound). The interior solution generates zero surplus for either player and consequentially zero utility. Then, sorting out is strictly preferable whenever $\tilde{p}_1 > r_1$. If we set $\tilde{p}_1 = B_1$ and start declining it, as in the experiment of Lazear et al. (2012), the condition

$\tilde{p}_1 > r_1$ is first violated for dictators with high reference points r_1, who transfer the least at the interior solution. These players are thus predicted to sort in first, regardless of how subjects mix status quo and expectations forming reference points, which corroborates the observation of Lazear et al. (2012) that the least generous dictators sort back in first.

Other games: Basic intuition The formal analysis presented in this section is restricted to the variety of distribution games as defined in section 2.4. Of course, the welfare-based altruism model derived in section 2.5 is applicable to a much broader class of games and many of the presented results and intuitions carry over to other games in which preferences for giving matter. In the following we informally discuss how our model may capture behavior in a small selection of related games that might be of interest to the reader.

Engelmann and Strobel (2004) experimentally investigate three player versions of the standard dictator game. The authors compare dictator choices across games in which they vary the coincidence of choices that are optimal with respect to different motives, in particular equality, efficiency, and maximin. While the authors interpret their findings as favouring models that include efficiency and maximin motives as suggested by Charness and Rabin (2002) over models that focus on equality motives like Fehr and Schmidt (1999) and Bolton and Ockenfels (2000), they also acknowledge that the relative importance attached to these motives seems to be quite sensitive to the details of the game. Welfare-based altruism may help explain this sensitivity. For, welfare-based altruists indeed seek to balance equality and efficiency via the preference parameter β, but in contrast to existing models, welfare-based altruists evaluate equality and efficiency not with respect to monetary payoffs but with respect to reference dependent individual welfares. The reference dependence induces sensitivity. A better understanding of how reference points depend on the details of the game might thus be the key to understanding the observed inconsistencies in the Engelmann-Strobel games just as in the distribution games analyzed above. More generally, note that welfare-based altruism is readily applicable to distribution games with more than two players.

The distribution games we focus on in this paper involve only one active player (the dictator) and therefore allow us to abstract from strategic concerns. Of course, the model is equally applicable to strategically more involved in-

teractions. While we leave most of these applications for future research, let us briefly discuss an application to one of the more prominent strategic games involving distributive concerns, namely the public goods game. Fischbacher et al. (2001) study the response functions of players in a four player standard linear public goods game using the strategy method. For every possible average contribution level of the other players, they elicit how much a given player wants to contribute herself. They find that the majority of their subjects can be either classified as free riders or conditional cooperators whose contributions to the public good are increasing in the average contribution of the other players. Furthermore, the authors identify a small group of subjects whose conditional contribution patterns are hump-shaped, i.e. first increasing and then decreasing in the others' average contribution. While the behavior of conditional cooperators follows naturally from the strategic complementarities that the game exhibits for payoff- or welfare-based altruists with weak efficiency concerns, compatibility with hump-shaped response functions is less obvious. Consider a player who observes a specific average contribution of the other players in her group. Suppose that player wants to contribute such that all players' reference points are fulfilled. However, based on her information she cannot be sure whether a given average contribution is due to a balanced contribution of each of the other players or, for example, a single player contributing above average. Depending on her beliefs about the other players' contributions, a welfare-based altruist may decrease her contribution as the average contribution exceeds some threshold if she believes only a low average is likely the result of an unbalanced contribution pattern at which she would want to step in to help out the highest contributor.

2.7 Implications for giving: Quantitative assessment

In this section, we quantitatively test welfare-based altruism on data from the very experiments discussed above. We examine whether welfare-based altruism indeed helps improve our understanding of giving in a statistically significant manner. Besides evaluating significance, this allows us to address three seemingly substantial concerns: Is the gain larger than two additional degrees of freedom (the reference points) allow to achieve anyways? Does it matter whether these degrees of freedom are spent on defining reference points, as predicted above, or perhaps on warm glow and cold prickle, or

envy and guilt, as suggested in the literature? Do these additional degrees of freedom facilitate overfitting?

Arguably, the match of theoretical predictions and empirical stylized facts for all distributions of reference points given "regularity" of dictator preferences strongly suggests that welfare-based altruism does capture giving reliably without the necessity of fine-tuning parameters. Yet, additional degrees of freedom tend to be an obstacle to robust fit (Hey et al., 2010), and to address this potential obstacle directly, our analysis will emphasize predictive adequacy over descriptive adequacy. For the lack of comparable analyses in the existing literature, we include a number of well-known models as benchmarks to provide some context.

2.7.1 The data

Table 2.2 summarizes the experiments we re-analyze. All of them are seminal studies run for the purpose of characterizing preferences underlying giving, rendering them adequate also for our purpose of validating utility representations of preferences. In total, we analyze behavior across 9 experiments, 83 treatments and 6500 observations. In relation to comparable studies of model validity, e.g. on choice under risk, this represents a very comprehensive data set, promising reliable results.

Table 2.2: The experiments re-analyzed to verify model adequacy

		Abbreviation	#Treatments	# Subjects	#Observations
Dictator games	Andreoni and Miller (2002)	AM02	8	176	1408
	Harrison and Johnson (2006)	HJ06	10	56	560
Generalized endowments					
	Cappelen et al. (2007)	CHST07	11	96	190
	Korenok et al. (2012)	KMR12	8	34	272
	Korenok et al. (2013)	KMR13	18	119	2142
Taking (and generalized endowments)					
	List (2007)	List07	3	120	120
	Bardsley (2008)	Bard08	6	180	180
	Korenok et al. (2014)	KMR14	9	106	954
Sorting	Lazear et al. (2012)	LMW12	8	94	518
Aggregate		Pooled	83	981	6578

To our knowledge, our data set includes all experiments on distribution games as analyzed above, i.e. with generalized endowments, taking, or sorting options, complete information, at least three treatments, manual entry of choices, and freely available data sets. The focus on experiments with at least three treatments facilitates statistically informative likelihood ratios but it precludes small experiments, most notably a seminal paper on sorting (Dana et al., 2006). The focus on games with complete information facilitates a unified theoretical treatment but precludes field experiments on charitable giving (such as DellaVigna et al., 2012) and experiments on moral wiggle room (Dana et al., 2007; van der Weele et al., 2014). The focus on games with manual choice entry simplifies out-of-sample predictions but precludes experiments with graphical user interfaces (Fisman et al., 2007). Finally, the focus on games with freely available data sets precludes the inclusion of experiments with real-effort tasks preceding a dictator game. However, as reviewed above, the main patterns in real-effort games resemble those in distribution games with generalized endowments and windfall budgets, three of which are included.

A notable difference between the analyzed distribution game experiments concerns the language used in the instructions for assigning the players' endowments. In standard dictator games (e.g. AM02 and HJ06), direct assignments are avoided by stating that "a number of tokens is to be divided", while in taking games (e.g. List07, Bard08, and KMR14), endowments are explicitly assigned prior to the choice task. This may provoke status quo and endowment effects (Samuelson and Zeckhauser, 1988; Kahneman et al., 1991) but to our knowledge it has not been discussed as a behavioral confound in preference analyses of (generalized) dictator games. Table A.2.1 in the appendix reviews the relevant passages in the experimental instructions and distinguishes between neutral language, where specific assignments of the endowments to either of the players are avoided, and loaded language, where initial endowments are specifically assigned or otherwise attributed to either of the players. Neutral language is typically used in standard dictator game experiments (AM02 and HJ06) and in sorting games (LMW12). Loaded language is typically used in experiments with generalized endowments or taking options. The hypothesis that such language differences affect the distribution of reference points and thus induce endowment effects as observed in other studies will be verified below and will be taken into account throughout the

entire analysis.

2.7.2 Heterogeneity and consistency of reference points

For the following analysis, we use the simplest formulation of reference points that seems conceivable, simplifying even in relation to Assumption 6, in order to rule out any biases in the results due to choosing functional forms.

Definition 4 (Simplified reference points)**.** Given game $\Gamma = \langle B_1, B_2, P_1, t \rangle$, the two players' reference points are

$$r_1 = w_1 \cdot B_1 + w_2 \cdot tB_2, \qquad r_2 = w_1 \cdot tB_2 + w_2 \cdot B_1.$$

Since the qualitative results hold regardless of the distribution of reference points, the specific assumption used here is largely irrelevant, but for any test of the model, some specification has to be used. The robustness checks in Appendix C explicitly show that alternative functional forms mapping endowments to reference points yield results very similar to those reported here. As above, we assume that they satisfy $w_1 \geq w_2$ such that subjects put higher weight on the role they end up playing in case their decision turns out to be payoff relevant, and we continue to allow that the weights $w_1, w_2 \in [0,1]$ do not necessarily add up to 1. The latter allows that subjects may be both altruistic givers and social pressure givers, thereby capturing the types observed by DellaVigna et al. (2012). Specifically, we speak of altruistic givers if $w_1 + w_2 < 1$, in which case satisfiability of reference points is fulfilled and dictators tend to pick (if $\alpha > 0$) interior solutions giving more than necessary to fulfill the recipient's reference point. In contrast, we speak of social-pressure givers if $w_1 + w_2 \geq 1$, which obtains if w_2 is sufficiently large, as the dictator is then unable to give more than "necessary" to both players, implying that she gives only to satisfy the reference points as good as possible. We fix the loss-aversion parameter at the conventional value $\delta = 2$ to remove a degree of freedom.

First, we examine heterogeneity of reference points within experiments (i.e. within subject pools) and consistency of reference point distributions across experiments (i.e. types of dictator games). We begin with examining consistency across experiments. For, the differences in the language used when assigning endowments potentially preclude consistency across experi-

ments, which might render the subsequent robustness analysis futile. Further, it would limit applicability of reference dependent concepts such as welfare-based altruism, or indeed any existing concept, to understand the behavioral reasons for differences in giving across experiments.

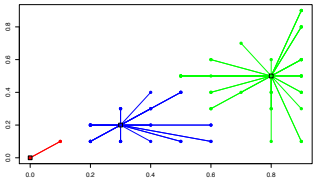
Formally, we estimate the individual reference points of all subjects in the largest experiment from each class of games: dictator games (AM02), games with generalized endowments (KMR13), sorting games (LMW12), and taking games (KMR14). To be precise, we estimate all four individual preference parameters for all subjects, as reference points cannot be estimated without controlling for altruism α and efficiency concerns β, but in the present subsection, we focus on the distributions of reference points. As the estimation procedure is standard maximum likelihood all details on optimization algorithms, generation of starting values, and cross-checking to ensure global optimality of estimates are relegated to the appendix. After estimating the reference point weights (w_1, w_2) for all subjects, we evaluate their structure in a cluster analysis by affinity propagation (Dueck and Frey, 2007). Figure 2.2 provides the results.

Consistently across data sets, three clusters of subjects are identified. The clusters tend to be of comparable size across experiments, each comprising at least 20% of the subjects in each case. In all cases, there is one group of subjects with endowment-independent reference points ($w_1 \approx w_2 \approx 0$), one group of subjects with "satisfiable" reference points where weights add up to less than one ($w_1 + w_2 < 1$), and one group of subjects with "excessive" reference points where weights add up to more than one ($w_1 + w_2 \geq 1$). The center of the second group moves a little between studies, but overall, the centers and sizes of the clusters are remarkably robust—and they fit received findings in the literature. The first group contains the "egoistic" subjects maximizing their pecuniary payoffs, a group comprising around one third of the subjects in all dictator game experiments. The members of the second and the third group comprise subjects that transfer tokens to the recipients either out of largely altruistic concerns (second group) or out of perceived social pressure (third group)—and further corroborating DellaVigna et al. (2012), these groups are similarly large.[18]

[18]Members of both the second and the third group react to the endowments induced via the experimental design. The difference is that the reference points of members in the second group do not eat up the entire budget, while the reference points of members in the third group cannot be satisfied jointly. The members of the third group transfer tokens aiming to

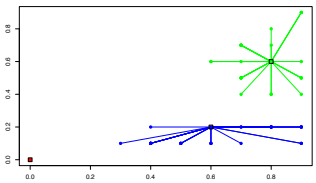
Figure 2.2: Distribution of reference point weights across types of dictator games

Dictator games (AM02)



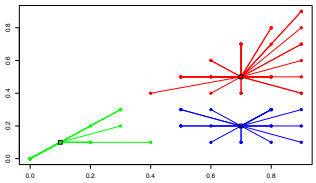
	w_1	w_2	Size	
Cluster 1	0.002	0.002	104/176	59%
Cluster 2	0.359	0.203	32/176	18%
Cluster 3	0.77	0.503	40/176	23%

Generalized endowments (KMR13)



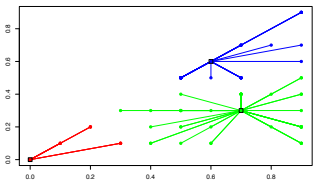
	w_1	w_2	Size	
Cluster 1	0	0	60/119	50%
Cluster 2	0.631	0.144	32/119	27%
Cluster 3	0.774	0.607	27/119	23%

Sorting games (LMW12)



	w_1	w_2	Size	
Cluster 1	0.106	0.084	32/94	34%
Cluster 2	0.71	0.215	20/94	21%
Cluster 3	0.717	0.571	42/94	45%

Taking games (KMR14)



	w_1	w_2	Size	
Cluster 1	0.049	0.032	37/106	35%
Cluster 2	0.69	0.267	39/106	37%
Cluster 3	0.657	0.593	30/106	28%

Note: For the largest experiments from each type of generalized dictator game, all individual reference point weights (w_1, w_2) are estimated, plotted with w_1 on the horizontal axis and w_2 on the vertical axis, and clustered by affinity propagation (Dueck and Frey, 2007). The centers and sizes of the three clusters identified in each case are provided in the respective tables to the right.

Result 4. *Across all four types of dictator games, there are three similarly-sized groups of subjects: subjects with endowment-independent reference points (mostly egoists), subjects with satisfiable reference points ("altruistic givers"), and subjects with non-satisfiable reference points ("social pressure givers").*

satisfy both players' reference points as good as possible, and in this sense, they react solely to the social pressure they perceive due to their (subjective) reference points. The members of the second group, however, react significantly weaker to the social pressure (i.e. to induced endowments), thanks to having smaller weights (w_1, w_2) and mainly decide how to transfer the (often substantial) residual amount after satisfying both reference points. In this sense, they are altruistic givers.

2.7.3 Significance and robustness of welfare-based altruism

Next, if reference dependence is a *robust* behavioral trait, then accounting for it improves both our descriptions and predictions of behavior across contexts. Besides being an informative test statistic, predictive adequacy is important also to improve policy recommendations and guide (behavioral) mechanism design. Given the data sets analyzed here, we can replicate out-of-sample predictions as used in such applications by making predictions across the types of dictator game experiments listed in Table 2.2.

In addition, if reference dependence is a *behavioral primitive*, then it improves on alternative ways of providing the implied degrees of freedom. Given the existing literature, there are two arguably natural extensions of CES altruism that have to be considered as benchmark models. The first benchmark extends CES altruism by warm glow and cold prickle, as proposed by Korenok et al. (2014):

$$u(\pi) = (1-\alpha_1-\alpha_2-\alpha_3)\cdot\pi_1^{\beta} + \alpha_1\,\pi_2^{\beta} + \alpha_2\cdot|B_1-\pi_1|_+^{\beta} - \alpha_3\cdot|B_2-\pi_2|_+^{\beta}, \qquad \text{(+ Warm Glow/Cold Prickle)}$$

where $|x|_+$ equates with x if $x>0$ and with 0 otherwise. Thus, $|B_1-\pi_1|_+$ captures the amount transferred by the dictator from her endowment (inducing "warm glow" which is independent of the amount received by the recipient), and $|B_2-\pi_2|_+$ captures the amount taken from the recipient's endowment (inducing "cold prickle"). The other benchmark extends CES altruism by motives of envy and guilt (Fehr and Schmidt, 1999) as proposed by Korenok et al. (2012).

$$u(\pi) = (1-\alpha_1-\alpha_2-\alpha_3)\cdot\pi_1^{\beta} + \alpha_1\,\pi_2^{\beta} - \alpha_2\cdot|\pi_1-\pi_2|_+ - \alpha_3\cdot|\pi_2-\pi_1|_+ \qquad \text{(+ Inequity Aversion)}$$

An attractive feature of these models is that they also contain four free parameters in total, in this respect equating with welfare-based altruism, which implies that these models can be estimated following the exact same procedure as welfare-based altruism. This way, we can ensure comparability of the results. All the technical details on likelihood maximization and statistical tests are provided in the appendix.

Table 2.3: Behavioral predictions across types of dictator game experiments

Calibrated on	Altruism is ...	Descriptive Adequacy	Predictive Adequacy	Details on predictions of ...			
				Dictator	Endowments	Taking	Sorting
Dictator Games	Payoff based (CES)	1460.9	8950.5	1343.4	4339	2353.3	914.7
	+ Warm Glow/Cold Prickle	1507.3^{--}	8854.6	1343	4218.4^{+}	2375.2	917.9
	+ Inequity Aversion	1234.6^{++}	8794.8^{++}	1217.1^{+}	4311.7	2360.7	905.3
	Welfare based	1146.6^{++}	8758^{++}	1279.8^{+}	4273.8^{+}	2316.6^{+}	887.7
	Welfare based (adj)	1146.6^{++}	8603.9^{++}	1263.9^{+}	4152.5^{++}	2300.8^{++}	888.2
Gen Endowments	Payoff based (CES)	2896.6	8752.9	4260.4	826.1	2613.8	1052.7
	+ Warm Glow/Cold Prickle	2395.5^{++}	8967.8^{--}	4289.6	954.5^{--}	2649.7	1074
	+ Inequity Aversion	2800.1^{+}	8916.4^{--}	4333.6^{-}	849.9	2663^{--}	1069.9^{--}
	Welfare based	2662.7^{++}	8416.7^{++}	4084.2^{++}	767.9^{+}	2565.9^{+}	998.7^{++}
	Welfare based (adj)	2662.7^{++}	7867.7^{++}	3985.8^{++}	637.1^{++}	2351^{++}	895.4^{++}
Taking Games	Payoff-based (CES)	1482.4	9700.7	3739.3	4466.7	579.7	914.9
	+ Warm Glow/Cold Prickle	1451.8	10252.5^{--}	4263.8^{--}	4408.7	592.8	987.2^{--}
	+ Inequity Aversion	1419.2^{+}	9736.7	3543.3^{++}	4698.2^{--}	576.6	918.5
	Welfare-based	1226.4^{++}	9499.7^{+}	3729.2	4343.2	568.5^{+}	858.8^{++}
	Welfare based (adj)	1226.4^{++}	9270.3^{++}	3633^{+}	4232.9^{++}	559.3^{++}	846.6^{++}
Aggregate	Payoff based (CES)	5839.8	27404.1	9343.1	9631.8	5546.8	2882.4
	+ Warm Glow/Cold Prickle	5354.6^{++}	28075^{--}	9896.5^{--}	9581.6	5617.8^{-}	2979.1^{--}
	+ Inequity Aversion	5453.9^{++}	27447.9	9094^{+}	9859.8^{--}	5600.4^{--}	2893.7
	Welfare based	5035.7^{++}	26674.4^{++}	9093.2^{++}	9385^{++}	5451^{++}	2745.2^{++}
	Welfare based (adj)	5035.7^{++}	25740.4^{++}	8883.6^{++}	9023.5^{++}	5212.2^{++}	2631.2^{++}

Note: For each type of dictator game experiment used to estimate the parameters (standard "Dictator games" in AM02, "Generalized endowments" in KMR13, "Taking Games" in KMR14), we report for each of the five models the in-sample fit ("Descriptive Adequacy"), the pooled out-of-sample fit by predicting all other experiments in Table 2.2 ("Predictive Adequacy"), and the detailed predictive adequacy for each type of experiments as distinguished in Table 2.2 (the four right-most columns). Plus and Minus signs indicate significance of differences of the Akaike Information Criterion (AIC) for each of the generalizations of the CES model to the CES model. The likelihood-ratio tests (Schennach and Wilhelm, 2016) are robust to misspecification and arbitrary nesting, and we distinguish significance levels of .05 ($^{+},^{-}$) and .01 ($^{++},^{--}$). In all cases, we cluster at the subject level to account for the panel character of the data.

We estimate all models on each of the three largest data sets, i.e. on standard dictator games (AM02), on games with generalized endowments (KMR13), and on games with taking options (KMR14), and predict behavior in all data sets listed in Table 2.2.[19] The results are summarized in Table 2.3. For completeness, we also provide the "Descriptive Adequacy", which is the Akaike information criterion of the in-sample fit, i.e. the sum of the absolute value of the log-likelihood and the number of parameters (in-sample, every reference point of every subject counts as a free parameter). Given the large number of parameters, the descriptive adequacy is of limited informational content on its own.

Our focus is on the "Predictive Adequacy", which is reported both on aggregate (column "Predictive Adequacy") and segregated by type of dictator game to be predicted (sets of columns "Details on predictions of ..."). In all cases, descriptive and predictive adequacies are reported for each of the four models discussed so far, payoff-based CES altruism, the extensions additionally allowing for either warm glow and cold prickle or envy and guilt, and the welfare-based altruism model. In addition, we report results from a robustness check allowing for variations in the strength of assignments of endowments, the model "Welfare based (adj)" that we discuss below. Finally, in the lower part of Table 2.3, all the numbers in the upper part are aggregated across all three in-sample data sets to provide the overall picture.

Descriptive adequacy Briefly, let us look at the in-sample fit (column "Descriptive Adequacy"). On aggregate, all generalized models significantly improve on the payoff-based CES model despite accounting for the additional parameters using AIC. The proposed model of welfare-based altruism is unique in that it improves highly significantly upon CES in all three contexts and in this sense represents the only robustly fitting model. Yet, the observation that on aggregate all three models do so suggests that perhaps they all capture differently important but significant facets of behavior. If so, this will show in their predictive adequacy.

[19] We do not consider predictions based on estimates from the sorting game experiment of LMW12, as their experimental design varies neither the transfer rate (fixed to $1:1$) nor the endowments of dictators and receivers, varying only the price for sorting out. This way, the preference parameter β, capturing the preference for efficiency and equity, is not identified and predictions are largely uninformative.

Predictive adequacy Evaluating robustness of the explanatory power (column "Predictive Adequacy") changes the picture substantially. Welfare-based altruism improves on CES' predictions in all contexts, regardless of the data set used for estimation, and mostly significantly so. That is, regardless of the context the model is fitted on and of the class of dictator game experiments to be predicted, the resulting goodness-of-fit is higher than that of the standard CES model, in all 3×4 cases, significantly so in $9/12$ cases, and always on aggregate.[20] The explanatory power of reference dependence in giving may therefore be considered robust.

At the other extreme, extending CES altruism by warm glow and cold prickle predicts behavior better than CES in only $3/12$ cases but worse than CES in $9/12$ cases. On aggregate, the alternative model's predictions are significantly worse than CES, and this obtains although warm glow and cold prickle seem to capture behavior (in-sample) in the case of generalized endowments best. This applies only in-sample, however, even predictions for the other experiments allowing for generalized endowments fit worse than CES (and all other models), suggesting that the extension allowing for warm glow and cold prickle does not capture a robust behavioral trait in the games analyzed here.

Finally, the extension allowing for envy and guilt ("inequity aversion") is in-between with respect to its descriptive and predictive adequacy. While it fits worse than welfare-based altruism in all contexts, both in-sample and out-of-sample, at least it does not overfit on aggregate and thereby it improves on warm glow and cold prickle. That is, on aggregate, accounting for envy and guilt does not yield predictions that are significantly worse than not doing so (as in the standard CES model). Nonetheless, predictions also do not improve on aggregate, suggesting that envy and guilt are actually not robust behavioral traits in giving—they allow to rationalize Leontief choices, but those are not robustly chosen.[21] Corroborating this observation, if we evaluate predictions

[20] Note that, as mentioned in the notes to all tables and in the appendix, we use the Schennach-Wilhelm likelihood ratio test throughout (Schennach and Wilhelm, 2016), clustered at the subject level. It is robust to misspecification of models, arbitrary nesting structures, and captures the panel character of the data with multiple observations per subject.

[21] In particular in the games with generalized (non-zero) endowments, the payoff-equalizing "Leontief" option happens to be rarely chosen (Korenok et al., 2013). For example, only $2/116$ subjects in KMR14 are strict Leontief types, whereas around 20% of the subjects are in standard dictator games (see AM02). In this context, predictions assuming that envy and guilt are behavioral factors fit poorly.

across all 4×3 cases, inequity aversion's predictions significantly improve on CES in $2/12$ cases, it predicts significantly worse in $4/12$ cases, and overall, its predictive adequacy is slightly worse than that of the payoff-based CES model.

Result 5. *Welfare-based altruism improves on CES altruism for all types of DG experiments, both descriptively (in-sample) and robustly (out-of-sample) highly significantly. None of the benchmark models does so in more than* $2/12$ *cases, corroborating the theoretical prediction that reference dependence is a causal factor in giving across contexts.*

Table 2.3 additionally informs on a robustness check accounting for the variation in language used assigning endowments (Table A.2.1 in the appendix). In this robustness check, we allow for homogeneous shifts in weights between experiments, by introducing a free parameter per set of predictions. Assuming the in-sample estimates of the weights are (w_1, w_2), we allow the out-of-sample weights to be (w_1^γ, w_2^γ), where the shift $\gamma \geq 0$ is homogeneous for all subjects. With $\gamma < 1$ all weights increase and with $\gamma > 1$ all weights decrease—reflecting stronger and weaker assignments, respectively. Introducing γ as a free parameter allows us to either strengthen or weaken weights homogeneously for all subjects. Naturally, this has no effect in-sample, but it has substantial effects out-of-sample—amounting to around 1000 points on the log-likelihood scale in total (yielding a drop from 26674.4 to 25740.4). This improvement is highly significant given the low number of additional parameters used, strongly underlining the initial hypothesis that the language used in experimental instructions is highly relevant in shaping behavior. The present analysis is neither suited nor intended to fully clarify the relevance of language used assigning endowments, but changes in language across experiments, which have not been explicitly discussed in the literature on generalized dictator games, are evidently not innocent choices in experimental design. This does not directly affect the above results, since acknowledging language differences as a factor shaping reference points only strengthens the case for welfare-based altruism, but such differences may be acknowledged more explicitly when designing and analyzing future experiments.

2.8 Conclusion

This paper contributes to the efforts in reorganizing models of the interdependence of preferences (List, 2009; Malmendier et al., 2014) that was initiated by a wave of distribution game experiments generalizing the standard dictator game allowing for non-trivial endowments (Bolton and Katok, 1998; Korenok et al., 2013), taking options (List, 2007; Bardsley, 2008), and sorting options (Dana et al., 2006; Lazear et al., 2012). The new observations were interpreted as being incompatible with observations from standard dictator games and in the existing literature a plethora of approaches have been proposed to capture them: menu dependent preferences and cold prickle to capture taking decisions, warm glow and social norms to capture endowment effects, image concerns and social pressure to capture sorting decisions. Considering this range of proposals simply to organize observations on giving under complete information, robustly applicable models of this most fundamental of economic activities appear to be out of reach (Korenok et al., 2014)—illustrating a surprisingly tight bound on economic modeling.

We propose an axiomatic approach toward modeling preferences over payoff profiles that resolves the persistent puzzle surrounding distributive decisions and differs from earlier work in four important ways. First, relying on an axiomatic foundation allows us to characterize a general family of utility functions representing interdependence of preferences. This identifies the class of candidate models. Second, we complement the axiomatic analysis by a comprehensive theoretic and econometric analysis of model validity across stylized facts and seminal laboratory experiments to provide a rigorous assessment of model adequacy. Third, as a technical innovation in the axiomatic derivation, we formally distinguish contexts, which allows us to formalize the notion of narrow bracketing as a property of preferences, and thus to establish a formally tight but ex-ante unsuspected link between four large literatures in behavioral economics: prospect theory (Kahneman and Tversky, 1979), narrow bracketing (Read et al., 1999b), altruism (Andreoni and Miller, 2002), and reference dependence (Kőszegi and Rabin, 2006b). Finally, our results reconcile a wide range of seemingly inconsistent experimental results with approaches and results from classical decision theory.

Implicitly, instead of constructing a utility function that fits as many stylized facts as possible, we derive a utility representation from established be-

havioral principles such as scaling invariance and narrow bracketing. This approach suggests applicability of our model that goes beyond the variety of distribution games analyzed in the paper. The theoretical predictions of behavior in these games, the tight relations to four major branches of behavioral economics, and the fact that welfare-based altruism directly formalizes the widespread notion that altruism is a concern for the welfare of others, while being derived from universal behavioral axioms not specific to altruism or distribution games, renders it a promising model for future work. Our econometric results on out-of-sample adequacy provide substantial validity in this respect, and both the model's generality and its quantitative adequacy open up a number of exciting avenues for future research.

These include experimental analyses of preferences and reference points, based on an axiomatically solid and econometrically validated model, theoretical analyses of utility representations under alternative axioms and of revealed preference with non-convexities (see also Halevy et al., 2017), empirical and theoretical analyses of behavioral welfare and preference laundering,[22] and, exploiting the relation to choice under risk, behavioral analyses of giving under incomplete information (as in Dana et al., 2007, and Andreoni and Bernheim, 2009) or in multilateral interactions. Due to the large extent of similarity of charitable giving and dictator behavior in the laboratory (Konow, 2010; Huck and Rasul, 2011; DellaVigna et al., 2012), a particularly immediate range of applications lies in structural analyses of charitable giving (DellaVigna, 2009; Card et al., 2011) generalizing, for example, the work of DellaVigna et al. (2012, 2016) and Huck et al. (2015).

[22]Letting all agents have equal weight, our analysis establishes a utilitarian welfare function which contains Rawls and Harsanyi as special cases (for $\beta \to -\infty$ and $\beta = 1$, respectively), where individual welfares are the prospect-theoretic utilities from single-person decision making. This provides an axiomatic foundation for preference laundering in welfare analyses (Goodin, 1986), i.e. to disregard concerns for others (such as envy) in behavioral welfare economics, which drastically affects policy recommendations (see also Piacquadio, 2017).

3 Fake news and information transmission

This chapter is based on joint work with Steffen Huck.

3.1 Abstract

We present a theoretical model to investigate how the presence of fake news affects information transmission from media outlets to economic agents. In a standard cheap talk framework we introduce uncertainty about the sender's (media outlet's) preferences. There are two types of media outlets. A fake news outlet wants to push the agent's belief to the maximum irrespective of the state of the world. A legitimate outlet wants to reveal the true state to the agent. We show that any informative perfect Bayesian equilibrium of our game is characterized by a threshold value. While the agent can perfectly separate amongst states below the threshold value, there is no separation amongst states above the threshold value. We determine the unique most informative threshold value for a general class of equilibria. Our results suggest that even if fake news are rare, their presence can have a substantial negative impact on the possibilities for information transmission.

3.2 Introduction

In recent years, non-traditional sources of information such as online news outlets and social media have gained a considerable amount of reach and influence (Newman et al., 2021). This development has led to a vast amount of information being widely available at virtually no cost. On the downside, these new sources of information are subject to far less rigorous curating and fact-checking procedures than traditional media which makes them less reliable and susceptible to misuse. Indeed, the spread of misinformation and fake news via online news sources has become a growing concern in both the economic and political sphere (see e.g. Newman et al., 2021; Mitchell et al., 2019; Lazer et al., 2018).

Fake news is fabricated or intentionally false information that is presented in a form that mimics reliable and fact-based news (Lazer et al., 2018; Allcott and Gentzkow, 2017). Fake news are often hardly distinguishable from factual news (Silverman and Singer-Vine, 2016), a problem that is exacerbated as

more and more sophisticated techniques such as deepfake algorithms become available to its fabricators. As a result, fake news pose a serious threat to the functioning of democratic societies which rely on citizens making factually informed decisions (Kuklinski et al., 2000).[23] Furthermore, they can prevent market forces from working properly and thereby limit a society's ability to achieve economically efficient outcomes.[24]

We present a theoretical model to investigate how the presence of fake news affects the transmission of information from media outlets to economic agents. Our model allows us to study how an agent's trust in media reports adjusts to the risk of being exposed to fake news. Furthermore, we are able to identify to what extent the presence of a fake news outlet has a negative spillover on the possibilities for a legitimate outlet to credibly transmit information.

Our model setup is based on Crawford and Sobel (1982)'s seminal work on strategic information transmission via cheap talk. A media outlet is privately informed about a state of the world. The state of the world is uniformly distributed on the unit interval. The media outlet sends a report to an agent. The agent then takes an action based on the information about the state she infers from the report. The agent's preferences are such that she wants her action to match the state of the world. We conceptualize fake news as reports that originate from an outlet that is strategically untrustworthy in the sense that it has an interest to push the agent towards holding extreme beliefs independent of the true state. In particular, we assume that the fake news outlet wants to induce the agent to take the maximal possible action. Fake news is contrasted to legitimate news which originates from an outlet which has preferences that are perfectly aligned with the agent's and therefore wants to reveal the true state. The agent cannot directly observe whether she is encountering a legitimate or a fake news outlet and has to update her beliefs about the type of media outlet she faces based on the report she receives.

We show that as long as the prior probability of meeting the fake news outlet is strictly above zero, there does not exist a fully informative perfect Bayesian equilibrium (PBE) in our game. Furthermore, any partially informative PBE takes the form of what we call a *threshold equilibrium*. In a thresh-

[23] See e.g. Allcott and Gentzkow (2017), Grinberg et al. (2019), Barrera et al. (2020), and Mocanu et al. (2015) for evidence on the prevalence, reach, and influence of fake news relating to political elections in the U.S. and Europe.

[24] See e.g. Kogan et al. (2020) for evidence on the impact of fake news in financial markets.

old equilibrium the agent is able to perfectly or partially separate amongst all states below a threshold value. Since the corresponding reports are never sent by the fake news outlet, they can credibly convey information about the state. However, there is no separation of states above the threshold value. The fake news outlet pools with the legitimate outlet on reports associated with states above the threshold value and since it wants to induce the maximal possible action irrespective of the state, all such reports must induce the same (maximal possible) belief about the state for the agent. The fake news outlet's equilibrium reporting strategy therefore "covers up" all information that the legitimate outlet may try to convey in order to separate amongst these states.

Since the fake news outlet's preferences are fully independent of the state, in our further analysis we restrict attention to state-independent reporting strategies for the fake news outlet. We are interested in identifying an upper bound for information transmission from the legitimate outlet to the economic agent in the presence of fake news. To that end, we concentrate on threshold equilibria in which the legitimate outlet perfectly separates all states below the threshold value. Without loss of generality, we focus on reporting strategies for the legitimate outlet in which it truthfully reports all states below the threshold value. We refer to such equilibria as *truthful threshold equilibria* (TTE). We show that in any TTE the agent follows all reports below the threshold value while she seizes to believe any report above the threshold value, reverting to taking an action equal to the threshold value for all such reports.

Conveniently, the informativeness of a TTE is measured directly by its threshold value. The higher the threshold value, the more states can be separated by the agent. We confirm existence and fully characterize potentially least and most informative TTE based on the legitimate outlet's reporting strategy for states above the threshold value. We find that the informativeness of a TTE does not depend on whether the legitimate outlet's reporting strategy perfectly separates or pools all states above the threshold value.

As is to be expected, the identified unique TTE threshold value is decreasing in the prior probability of meeting the fake news outlet. In the limit, as the prior probability of meeting the fake news outlet approaches one, the threshold value approaches 0.5, meaning that the legitimate outlet can separate only the lowest 50% of states. In the other extreme, as the prior probability of meeting the fake news outlet approaches zero, the TTE approaches full infor-

mativeness. In between, the threshold value is increasing first slowly, reaching a separation of close to 60% of states by the legitimate outlet as the agent encounters a fake news outlet 50% of the time, picking up speed the more likely it becomes for the agent to meet the legitimate outlet. At 90% probability of meeting the legitimate outlet, still, the highest 25% of states remain unseparated.

To get a first sense of the bounds on information transmission indicated by our model in a real world setting we can perform a back of the envelope calculation based on evidence collected by Allcott et al. (2019). The authors report that preceding the 2016 U.S. presidential elections about 40% of total news site engagements on facebook were with fake news sites. Translated to our model their finding indicates a 40% chance of meeting a fake news outlet on facebook which is associated with a TTE threshold value that allows legitimate outlets to separate only roughly 60% of states. Although this number can only serve as an extremely rough estimate for the upper bound of information transmission on social media, it still highlights the large negative spillover that the presence of fake news can have on the possibilities for information transmission from legitimate sources.

The paper proceeds as follows. In section 2 we discuss the related literature. In section 3 we present our model. In section 4 we discuss general equilibrim properties and present our characterization of TTE. Section 5 concludes.

3.3 Related literature

3.3.1 Strategic information transmission

We extend and vary Crawford and Sobel (1982)'s cheap talk framework in two important ways to capture the central characteristics of our fake news application. First, we introduce uncertainty about the media outlet's preferences. On the market for news agents do not perfectly observe whether it is in the best interest of a given news outlet to convey information truthfully or whether it has an incentive to mislead the agent. Furthermore, while Crawford and Sobel (1982) consider a sender with biased but state-dependent preferences over the agent's action, we look at the extreme case of a sender whose preferences are independent of the state of the world. This captures the central feature of fake news being fully detached from the truth.

Strategic information transmission when there is uncertainty about the sender's preferences was first studied by Sobel (1985) with a focus on reputation building in repeated interactions. In a model with binary state space and continuous action space Sobel (1985) introduces two types of senders. One is committed to truthful reporting and one is strategic with preferences that are diametrically opposed to the receiver's. More related to our model setup, Morris (2001) replaces the two sender types in Sobel (1985)'s model with two strategic senders, one who has preferences that are perfectly aligned with the receiver's and one who wants to induce the maximum action irrespective of the state. The unique informative equilibrium of the one-stage version of his model closely resembles the structure of our truthful threshold equilibria (TTE). The aligned sender reports truthfully while the misaligned sender always sends a high report irrespective of the state. As a result, the receiver's equilibrium action perfectly matches a low report while it only partially matches a high report. However, this equilibrium only exists if the probability of encountering the aligned sender is high enough. In contrast to Morris (2001) we consider a continuous state space. This allows us to look at a richer spectrum of partial information transmission and opens up the possibility to study the more subtle consequences of fake news in settings where there is only a small probability of being exposed to them.

Morgan and Stocken (2003) introduce uncertainty about sender preferences in a cheap talk framework with a continuous state space to study stock recommendations of financial analysts. They model both sender types as strategic players, one with perfectly aligned preferences and one with upward biased but state-dependent preferences. Similar to our results, the authors show that due to the upward bias in the preferences of the misaligned sender it is impossible to convey precise information about high states in equilibrium while it remains possible to transmit precise information about low states. Indeed, Morgan and Stocken (2003)'s semiresponsive equilibria are characterized by the same threshold value for information transmission as our TTE. However, in Morgan and Stocken (2003)'s framework there exist a multitude of partially informative equilibria with a different structure to our threshold equilibria which makes determining an upper bound on equilibrium information transmission unattainable. Another major difference to Morgan and Stocken (2003)'s analysis is that we do not restrict attention to pure reporting strategies. This allows us to characterize a threshold equilibrium that survives

commitment to truthful reporting by the aligned sender.

3.3.2 Fake news and media bias

Our paper is loosely related to an existing theoretical literature focusing on how competition in the news market influences supply side driven media bias (see e.g. Gehlbach and Sonin, 2014; Baron, 2006; Anderson and McLaren, 2012). A comprehensive survey of this literature is provided in Gentzkow et al. (2015). Roughly, these models consider settings in which news outlets face a tradeoff between attracting consumers with preferences for accuracy and biasing consumers' actions toward one side of the political spectrum. It is found that if the biasing motive of news outlets is strong enough, media bias can arise in equilibrium. However, strengthening competition generally reduces and eventually resolves equilibrium bias. The most striking difference to our framework is that the theoretical literature on competition and media bias models information transmission between news outlets and economic agents as Bayesian persuasion. In contrast to our cheap talk approach, it is assumed that before observing the state of the world news outlets announce and commit to a reporting strategy. Thus, consumers directly observe the bias of a media outlet and choose whether to consume news from that outlet based on this information. While focusing on the effects of competition, this literature fully abstracts from the agent's inference problem about the accuracy of news she observes. In contrast, we abstract from market forces and focus on the agent's inference problem. To the best of our knowledge, ours is the first paper that looks at the strategic effects of fake news on the possibilities for information transmission.

Kogan et al. (2020) present an empirical study on fake news in financial markets that nicely relates to our theoretical results. Exploiting a shock to traders' awareness of fake news on popular social media platforms for financial news, the authors investigate the effects of the presence of fake news on the extent to which traders react to the reports provided on these platforms. They find that trading volume and price volatility in response to both legitimate and fake news articles dropped significantly after traders gained awareness about the problem of fake news. Their observation corroborates our theoretical results on the negative spillover effects that the presence of fake news has on the transmission of information from legtitimate news out-

lets to economic agents.

3.4 The model

There is a state of the world, ω, uniformly distributed on $[0,1]$. There is an uninformed agent A and a fully informed media outlet. The agent takes an action, $x \in [0,1]$, with which she wants to match the state of the world. Specifically, we assume that her von Neumann-Morgenstern utility function is $u(x) = -(\omega - x)^2$. The fully informed media outlet writes a report $r \in [0,1]$. There are two types of media outlets. A legitimate outlet L has preferences that are fully aligned with the agent's, i.e. its von Neumann-Morgenstern utility function is $v_L(x) = -(\omega - x)^2$. A fake news outlet F wants to push the agent's decision to the extreme irrespective of the state of the world. Its von Neumann-Morgenstern utility is given by $v_F(x) = -(1-x)^2$. The prior probability of the media outlet being legitimate is denoted by $p \in (0,1)$.

The game proceeds as follows. First, the media outlet observes its type $t \in \{L,F\}$ and the state of the world $\omega \in [0,1]$. Then, the media outlet sends a report $r \in [0,1]$. After observing the media outlet's report, the agent chooses an action $x \in [0,1]$. Finally, all players' payoffs are realized.

A perfect Bayesian equilibrium (PBE) of the game is given by

- A family of reporting rules for each type of media outlet $t \in \{L,F\}$, $q_t(r|\omega)$, specifying for each $\omega \in [0,1]$ the density over reports $r \in [0,1]$. It must hold for any $\omega \in [0,1]$ and $t \in \{L,F\}$ that $\int_0^1 q_t(r|\omega)dr = 1$.
- An action rule for A, $x(r)$, specifying an action for each report $r \in [0,1]$.[25]
- A posterior belief for A, $\mu(\omega|r)$, specifying for each report $r \in [0,1]$ A's conditional belief about the distribution of states.

such that

(i) For each $r \in [0,1]$, A's action maximizes her expected utility given her posterior belief $\mu(\omega|r)$.

(ii) For each $\omega \in [0,1]$ and each r in the support of $q_t(r|\omega)$, r maximizes media outlet type t's expected utility given A's action rule $x(r)$.

[25] Note that since A's objective function is strictly concave in x, A will never use mixed strategies in equilibrium.

(iii) For each $r \in [0,1]$, $\mu(\omega|r)$ follows from Bayes' rule given $q_L(r|\omega)$ and $q_F(r|\omega)$.

Before turning to our model analysis in the next section, let us introduce some useful notation. Let $l(r)$ denote A's posterior belief about the probability of dealing with the legitimate outlet after receiving report r. Further, let $\mu_t(\omega|r)$ for $t \in \{L,F\}$ denote A's conditional belief about the distribution of states after receiving report r given she knew that she was dealing with a media outlet of type t.

3.5 Threshold equilibria

As in any model of cheap talk there always exist so-called babbling equilibria in which there is no transmission of information. If the agent ignores any report sent by the media outlet, neither type of outlet has an incentive to engage in informative reporting. Vice versa, if both media outlets employ an uninformative reporting strategy, the agent's best response is to take the same action corresponding to her prior expectation about the state of the world irrespective of the report she receives.

We call a perfect Bayesian equilibrium (PBE) of our model *informative* if there exist at least two reports $r, r' \in [0,1]$, each in the support of at least one media outlet type's reporting rule, r for state $\omega \in [0,1]$ and r' for state $\omega' \in [0,1]$, such that $x(r) \neq x(r')$. The following proposition determines the general structure of informative PBE in our model.

Proposition 7 (Threshold equilibria)**.** *Any informative PBE of our game is characterized by a threshold value $k(p) \in (0,1)$ such that the following statements hold:*

(i) The highest possible action that can be induced according to A's action rule is equal to the threshold value, i.e. $\max_r x(r) = k(p)$.

(ii) For any $\omega \in [0,1]$, F's reporting rule $q_F(r|\omega)$ is supported only on those $r \in [0,1]$ that induce an action equal to the threshold value, i.e. r such that $x(r) = k(p)$.

(iii) Any report $r' \in [0,1]$ such that r' is in the support of L's reporting rule $q_L(r|\omega)$ for some $\omega \in [k(p),1]$ induces A to take the same action $x(r') = k(p)$.

Proof.

Step 1: For any $\omega \in [0,1]$, if a report r' is in the support of $q_F(r|\omega)$, then $r' \in \arg\max_r x(r)$. If for some $\omega \in [0,1]$ F's reporting rule $q_F(r|\omega)$ would have support on any $r' \notin \arg\max_r x(r)$, F would have an incentive to deviate to excluding r' from the support of $q_F(r|\omega)$ and shifting all probability mass from r' to any $r'' \in \arg\max_r x(r)$.

Step 2: Define $k(p) := \max_r x(r)$. It must hold that $k(p) \in (0,1)$. Suppose $x(r' \in \arg\max_r x(r)) = 0$. Then the PBE cannot be informative since by $x \in [0,1]$ this implies $x(r') = x(r'')$ for all $r', r'' \in [0,1]$.

Suppose $x(r' \in \arg\max_r x(r)) = 1$. By Step 1 it must be the case that $q_F(r' \in \arg\max_r x(r)|\omega) = 1$ for all $\omega \in [0,1]$, which by Bayes' rule contradicts $\mu(\omega = 1|r' \in \arg\max_r x(r)) = 1$ and thus utility maximization of A.

Step 3: $\int_{r \in \arg\max_r x(r)} q_L(r|\omega)dr = 1$ for all $\omega \in [k(p),1]$. Suppose there exists $r' \notin \arg\max_r x(r)$ such that r' is in the support of $q_L(r|\omega')$ for some $\omega' \in [k(p),1]$. This implies $x(r') < k(p)$ such that L would have an incentive to adjust $q_L(r|\omega')$ by removing all probability mass from r' and reallocating it to any $r'' \in \arg\max_r x(r)$.

□

It follows from Proposition 7 that as long as the probability of encountering the legitimate outlet remains below one, there does not exist a fully informative PBE. In particular, there does not exist a PBE in which there is any separation amongst the highest states, i.e. the states above what we call the equilibrium threshold value $k(p)$ with $k(p) \in (0,1)$. The proposition states further that any informative PBE must have the structure of what we call a threshold equilibrium. In such an equilibrium, F only sends reports that induce A to take the maximal possible action. The same holds for L in any state $\omega \in [k(p),1]$. Thus, there exists no PBE in which states $\omega \in [k(p),1]$ can be separated from each other. In any such state the maximal possible action according to A's action rule is induced and this action is equal to the threshold value $k(p)$. Only for states below the threshold value, i.e. $\omega \in [0,k(p))$, does L have an incentive to induce actions below the maximal possible action $k(p)$ and since the corresponding reports are never sent by F, they can credibly convey information that enables the agent to separate them both from the states

above the threshold value and amongst each other.

Proposition 7 follows from the observation that as F wants to induce the highest possible action irrespective of the state, its reporting strategy needs to make sure that A does not infer different expectations about the state from different reports inside her strategy support. In particular, any report inside the support of any of F's reporting rules must induce A to hold the posterior belief that induces her to take the maximal action according to her action rule. Otherwise, F has an incentive to deviate from her reporting rule in any state for which this does not hold by shifting probability mass from a report that induces a lower action to a report that induces the maximal action. As a result, in any informative PBE separation can only happen amongst the states that L's corresponding reporting rules associate with reports that are never sent by F. Naturally, these states have to be on the lower end of the state space (below the threshold value) to make sure that F has no incentive to pool with L on the corresponding reports.

Since F's preferences are entirely independent of the state of the world, in the following we will restrict attention to state-independent reporting rules for F, i.e. $q_F(r|\omega) = q_F(r|\omega') =: q_F(r)$ for all $\omega, \omega' \in [0,1]$. By Bayes' rule a state-independent reporting strategy for F implies that if A knew she was dealing with the fake news outlet, her posterior expectation about the state of the world after receiving a report inside F's strategy support would be equal to her prior, i.e. $\int_0^1 \omega \mu_F(\omega|r) d\omega = \frac{1}{2}$ for all $r \in [0,1]$ such that r is in the support of F's reporting rule $q_F(r)$. Since we are mainly interested in characterizing the most informative PBE, we will further restrict attention to PBE in which L perfectly separates all states below the threshold value $k(p)$. Without loss of generality, we only consider such PBE in which L employs a truthful reporting rule for all states below $k(p)$, i.e. $q_L(r = \omega|\omega) = 1$ and $q_L(r \neq \omega|\omega) = 0$ for all $r \in [0, k(p))$ as well as $q_L(r \in [0,k(p))|\omega \in [k(p),1]) = 0$. We call such PBE *truthful threshold equilibria (TTE)*.

The following Lemma clarifies that in a TTE, F will never send reports $r \in [0,k(p))$. Therefore, any report $r \in [0,k(p))$ perfectly reveals the state and A will follow the report by choosing $x(r) = r$.

Lemma 2. *Suppose $q_F(r)$ and $x(r)$ are part of a TTE. Then $q_F(r)$ is only supported on $r \in [k(p),1)$. Furthermore, $x(r) = r$ for all $r \in [0,k(p))$.*

Proof. Since L employs a truthful reporting strategy for states $\omega \in [0,k(p))$,

it must follow by Bayes' rule that $\mu_L(\omega = r|r) = 1$ and $\mu_L(\omega \neq r|r) = 0$ for all $r \in [0, k(p))$. Now suppose there exists a report $r' \in [0, k(p))$ such that r' is in the support of $q_F(r)$. Since F's reporting rule is independent of the state, it must hold by Bayes' rule that $\int_0^1 \omega \mu_F(\omega|r')d\omega = \frac{1}{2}$. By Bayes' rule we therefore have

$$\int_0^1 \omega \mu(\omega|r')d\omega = \frac{p}{p+(1-p)q_F(r')}r' + \frac{(1-p)q_F(r')}{p+(1-p)q_F(r')}\frac{1}{2}.$$

Now by truthful reporting of L for states $\omega \in [0, k(p))$, r' cannot be in the support of $q_L(r|\omega)$ for any $\omega \in [k(p), 1]$. Therefore, it must hold for any $r'' \in [k(p), 1]$ such that r'' is in the support of $q_L(r|\omega)$ for some $\omega \in [k(p), 1]$ that $\int_0^1 \omega \mu_L(\omega|r'')d\omega > r'$. By Bayes' rule this implies $\int_0^1 \omega \mu(\omega|r'')d\omega > \int_0^1 \omega \mu(\omega|r')d\omega$. Thus, expected utility maximization of A requires $x(r'') > x(r')$ which by Proposition 7 contradicts r' being in the support of $q_F(r)$.

As a result, if A observes a report $r \in [0, k(p))$, she is sure to deal with the legitimate outlet and her posterior belief after observing r must be $\mu(\omega = r|r) = 1$ and $\mu(\omega \neq r|r) = 0$ such that her best response is $x(r) = r$. □

Conveniently, the informativeness of a TTE is directly linked to its threshold value $k(p)$. The higher $k(p)$, the more states can be perfectly separated from each other by the agent and the more informative the associated TTE. To understand the informativeness range of TTE in our game it is useful to characterize the two most extreme candidates for L's TTE reporting rules in terms of the information they would convey in the absence of the fake news outlet, i.e. for $p = 1$. As defined above, in a TTE L truthfully reports all states $\omega \in [0, k(p))$, i.e. $q_L(r = \omega|\omega) = 1$ and $q_L(r \neq \omega|\omega) = 0$ for all $\omega \in [0, k(p))$. Thus, we only distinguish L's reporting rules for $\omega \in [k(p), 1]$. In absence of the fake news outlet the least information would be conveyed if L's reporting rules were state-independent across states $\omega \in [k(p), 1]$. We formalize such reporting rules as $q_L(r|\omega)$ uniform on $[k(p), 1]$ for all $\omega \in [k(p), 1]$ and refer to the corresponding family of reporting rules as L's least separating truthful threshold (TT) reporting rules. In turn, the most information would be conveyed in absence of F if L's reporting rules were to perfectly separate all states $\omega \in [k(p), 1]$. We formalize such reporting rules as $q_L(r = \omega|\omega) = 1$ and $q_L(r \neq \omega|\omega) = 0$ for all $\omega \in [k(p), 1]$ and refer to the corresponding family of

reporting rules as L's most separating TT reporting rules.[26]

The following proposition shows existence of TTE both with most and least separating TT reporting rules for L. Furthermore, it shows that the informativeness of these two TTE is exactly the same.

Proposition 8. *For any $p \in (0,1)$ there exist both a TTE with least separating TT reporting by L and one with most separating TT reporting by L. The threshold for both of these equilibria is the same and given by $k(p) = \frac{1}{p}(1-\sqrt{1-p})$. Further, there does not exist a more informative TTE in which L employs her most, respectively least, separating TT reporting rules.*

Proof.

Step 1: Most informative TTE with least separating TT reporting rules for L. We show that the following strategies and posterior beliefs constitute a PBE for $k(p) = \frac{1}{p}(1-\sqrt{1-p})$[27]:

$$q_L(r = \omega|\omega) = 1 \text{ and } q_L(r \neq \omega|\omega) = 0 \text{ for all } \omega \in [0, k(p))$$
$$q_L(r|\omega) \text{ uniform on } [k(p),1] \text{ for all } \omega \in [k(p),1]$$

$$q_F(r|\omega) \text{ uniform on } [k(p),1] \text{ for all } \omega \in [0,1]$$

$$x(r) = \begin{cases} r & \text{for } r \in [0, k(p)) \\ k(p) & \text{for } r \in [k(p),1] \end{cases}$$

$$\mu(\omega = r|r) = 1 \text{ and } \mu(\omega \neq r|r) = 0 \text{ for all } r \in [0, k(p))$$
$$\mu(\omega|r \in [k(p),1]) = \begin{cases} \frac{1-p}{p(1-k(p))+1-p} & \text{for } \omega \in [0, k(p)) \\ \frac{1}{p(1-k(p))+1-p} & \text{for } \omega \in [k(p),1] \end{cases}$$

Consider first L. Given A's action rule, its reporting rule in any state $\omega \in [0, k(p)]$ achieves its maximum possible utility of $v_L(x = \omega) = 0$. Further-

[26] Note that concerning the informativeness of the resulting TTE the chosen formalizations of L's most and least separating TT reporting rules are largely without loss of generality in the sense that they only restrict TTE informativeness if we allow for unrestricted off-equilibrium-path beliefs.

[27] This threshold equilibrium is very similar to Morgan and Stocken (2003)'s characterization of size 1 semiresponsive equilibria in a related game.

more, its reporting rule in any state $\omega \in (k(p),1]$ yields $v_L(x=k(p)) = -(\omega - k(p))^2$. Given A's action rule, this is the maximum possible utility L can achieve in states $\omega \in (k(p),1]$.

Consider now F. Given A's action rule, its reporting rule in any state $\omega \in [0,1]$ achieves utility of $v_F(x=k(p)) = -(1-k(p))^2$. Given A's action rule, this is the maximum possible utility F can achieve in any state $\omega \in [0,1]$.

Consider now A. Her posterior beliefs and actions for $r \in [0,k(p))$ are pinned down by Lemma 2. Given the reporting rules of L and F, by Bayes' rule it follows that for $r \in [k(p),1]$, $l(r) = \frac{p(1-k(p))}{p(1-k(p))+1-p}$, $\mu_L(\omega|r)$ uniform on $[k(p),1]$, and $\mu_F(\omega|r)$ uniform on $[0,1]$. Thus, A's best response after receiving $r \in [k(p),1]$ which is equal to her posterior expectation about the state of the world is given by

$$x(r) = \frac{p(1-k(p))}{p(1-k(p))+1-p}\frac{1+k(p)}{2} + \frac{p(1-k(p))}{p(1-k(p))+1-p}\frac{1}{2}.$$

Proposition 7 requires this best response to be equal to $k(p)$, i.e.

$$\frac{p(1-k(p))}{p(1-k(p))+1-p}\frac{1+k(p)}{2} + \frac{p(1-k(p))}{p(1-k(p))+1-p}\frac{1}{2} = k(p) \qquad (11)$$

Solving the above equation for $k(p)$ reveals that $k(p) = \frac{1}{p}(1-\sqrt{1-p})$ is the unique $k(p) \in [0,1]$ for which equation 11 is fulfilled.

Step 2: Most informative TTE with most separating TT reporting rules for L. We show that the following strategies and posterior beliefs constitute a PBE for $k(p) = \frac{1}{p}(1-\sqrt{1-p})$:

$$q_L(r=\omega|\omega) = 1 \text{ and } q_L(r \neq \omega|\omega) = 0 \text{ for all } \omega \in [0,1]$$

$$q_F(r|\omega) =: q_F(r) \text{ for all } \omega \in [0,1] \text{ with}$$

$$q_F(r) = \begin{cases} 0 & \text{for } r \in [0,k(p)) \\ \frac{2p}{(1-p)(2k(p)-1)}(r-k(p)) & \text{for } r \in [k(p),1] \end{cases}$$

$$x(r) = \begin{cases} r & \text{for } r \in [0, k(p)) \\ k(p) & \text{for } r \in [k(p), 1] \end{cases}$$

$$\mu(\omega = r|r) = 1 \text{ and } \mu(\omega \neq r|r) = 0 \text{ for all } r \in [0, k(p))$$
$$\mu(\omega|r) = \begin{cases} \frac{(1-p)q_F(r)}{p+(1-p)q_F(r)} & \text{for } \omega \neq r \\ 1 & \text{for } \omega = r \end{cases} \text{ for all } r \in [k(p), 1]$$

Consider first L. Given A's action rule, its reporting rule in any state $\omega \in [0, k(p)]$ achieves its maximum possible utility of $v_L(x = \omega) = 0$. Furthermore, its reporting rule in any state $\omega \in (k(p), 1]$ yields $v_L(x = k(p)) = -(\omega - k(p))^2$. Given A's action rule this is the maximum possible utility L can achieve in states $\omega \in (k(p), 1]$.

Consider now F. Given A's action rule, its reporting rule in any state $\omega \in [0, 1]$ achieves utility of $v_F(x = k(p)) = -(1 - k(p))^2$. Given A's action rule this is the maximum possible utility F can achieve in any state $\omega \in [0, 1]$.

Consider now A. Her posterior beliefs and actions for $r \in [0, k(p))$ are pinned down by Lemma 2. Given the reporting rules of L and given that F uses a state-independent reporting rule, i.e. $q_F(r) := q_F(r|\omega)$ for all $\omega \in [0, 1]$, by Bayes' rule it follows that for $r \in [k(p), 1]$, $l(r) = \frac{p}{p+(1-p)q_F(r)}$, $\mu_L(\omega = r|r) = 1$, $\mu_L(\omega \neq r|r) = 0$, and $\mu_F(\omega|r)$ uniform on $[0, 1]$. Thus, A's best response after receiving $r \in [k(p), 1]$ which is equal to her posterior expectation about the state of the world is given by

$$x(r) = \frac{p}{p + (1-p)q_F(r)} r + \frac{(1-p)q_F(r)}{p + (1-p)q_F(r)} \frac{1}{2} \tag{12}$$
$$= \frac{1}{2} + \frac{p(r - \frac{1}{2})}{p + (1-p)q_F(r)}. \tag{13}$$

By Proposition 7 it must hold that in equilibrium $x(r) = k(p)$ for all $r \in [k(p), 1]$. We can now rewrite equation 12 to obtain

$$q_F(r) = \frac{2p}{(1-p)(2k(p) - 1)}(r - k(p)). \tag{14}$$

Finally, as it must hold that $\int_{k(p)}^{1} q_F(r)dr = 1$, we obtain

$$\int_{k(p)}^{1} \frac{2p}{(1-p)(2k(p)-1)}(r-k(p))dr = 1$$
$$\frac{p(k(p)-1)^2}{(1-p)(2k(p)-1)} = 1$$
$$k(p) = \frac{1}{p}(1-\sqrt{1-p}).$$

□

Note that by proving existence of a TTE in which L truthfully reports all states $\omega \in [0,1]$ (TTE with most separating TT reporting by L), Proposition 8 also establishes that allowing for commitment to truthful reporting by L cannot help to improve TTE informativeness.

Figure 3.1 illustrates how the TTE threshold value $k(p)$ as characterized by Proposition 8 evolves as the probability of meeting the legitimate outlet increases. As the prior probability of media outlet type L approaches one, the corresponding TTE approach full informativeness, i.e. $\lim_{p\to 1} k(p) = 1$. However, they do not become fully uninformative as the prior probability of media outlet type L approaches zero. Indeed, $\lim_{p\to 0} k(p) = 0.5$. This makes sense because as the probability of meeting the fake news outlet approaches one, A's best response to any report $r \in [k(p),1]$ must fall back on her prior expectation of the state of the world since F's reporting rule is fully uninformative. Furthermore, the higher the prior probability of media outlet type L, the more informative the TTE as characterized by Proposition 8.[28] Starting from $p \to 0$, the TTE threshold value is first slowly increasing, allowing A to separate close to 60% of states as $p = 0.5$, picking up speed the more likely it becomes for A to meet the legitimate outlet. At $p = 0.9$, still, the highest 25% of states remain unseparated from each other.

3.6 Conclusion

We present a simple cheap talk model to investigate how the presence of fake news affects the transmission of information from media outlets to economic agents. Our model analysis reveals that even a small probability of encountering a news report that originates from a fake news source has a substantial

[28] $\frac{\partial k(p)}{\partial p} = (4(1-p)+4\sqrt{1-p}-2p\sqrt{1-p})^{-1} > 0$ for $p \in (0,1)$

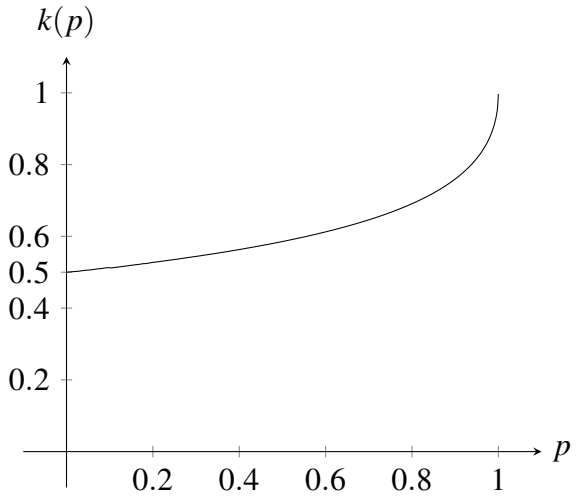


Figure 3.1: TTE threshold value $k(p)$.

negative impact on the amount of information that a legitimate source is able to credibly transmit in equilibrium. Our insights highlight the strategic effects of fake news on the reporting behavior of legitimate news outlets while taking into account the inference problem of economic agents who cannot observe the quality of the news they consume directly. While being at the very heart of the challenges that the presence of fake news poses for society, this inference problem has so far received very little attention in the theoretical literature on misinformation and fake news. The present paper provides a first step towards closing this gap.

A central assumption of our model is that the agent's updating behavior is fully rational. We assume that her posterior belief upon observing a news report is determined by Bayes' rule based on the equilibrium reporting strategies of media outlets. Of course, real world economic agents do not always update beliefs rationally. Indeed, there exists a large body of research demonstrating systematic biases in economic agents' belief updating behavior (see e.g. Ambuehl and Li, 2018; Esponda and Vespa, 2018; Barron et al., 2019; Enke, 2020). In particular, we consider understanding the effects of selection neglect or coarse reasoning as modeled by Jehiel and Koessler (2008) as well as Eyster and Rabin (2005) on the possibilites for information transmission in the presence of fake news a promising direction for future research.

A Appendix

The papers underlying the various chapters have profited from the input of colleagues and seminar participants at HU Berlin, WZB Berlin, DIW Berlin, TU Berlin, LMU Munich, University of Oxford, Frankfurt School of Finance and Management, DICE Düsseldorf, University of Manchester, University of Naples Frederico II, University of Exeter, University of Konstanz, University of Cologne, and NHH Bergen as well as conference participants in Utrecht (IMEBESS 2019), Kreuzlingen (theem 2019), Berlin (ESA 2018), Barcelona (IMEBESS 2017), and Lisbon (EEA/ESEM 2017). Financial support of the DFG (CRC TRR 190) is greatly appreciated.

A.1 Reference-dependent choice bracketing

A.1.1 Proof of Proposition 1 (Indifference curves)

Proof. The marginal rates of substitution for the broad and the narrow bracketer are

$$MRS(x_1,x_2) = \left.\frac{\partial u}{\partial x_1}\right|_{(x_1,x_2)} \left(\left.\frac{\partial u}{\partial x_2}\right|_{(x_1,x_2)}\right)^{-1} \text{ and}$$

$$\widetilde{MRS}(x_1,x_2) = \left.\frac{\partial u}{\partial x_1}\right|_{(x_1,r_2)} \left(\left.\frac{\partial u}{\partial x_2}\right|_{(r_1,x_2)}\right)^{-1}.$$

Thus, we obviously have $MRS(r_1,r_2) = \widetilde{MRS}(r_1,r_2)$. In this proof I focus on the case $\frac{\partial^2 u}{\partial x_1 \partial x_2} > 0$. The other two cases can are proven analogously. Consider pairs (r_1,x_2) with $x_2 > r_2$. The above expressions for the broad and narrow marginal rates of substitution reveal that the numerator of $\widetilde{MRS}(r_1,x_2)$ is equal to the numerator of $\widetilde{MRS}(r_1,r_2)$ and the denominator of $MRS(r_1,x_2)$ is equal to the denominator of $\widetilde{MRS}(r_1,x_2)$. Furthermore, by $\frac{\partial^2 u}{\partial x_1 \partial x_2} > 0$ we have that the numerator of $MRS(r_1,x_2)$ is larger than the numerator of $MRS(r_1,r_2)$. Together with $MRS(r_1,r_2) = \widetilde{MRS}(r_1,r_2)$ this implies $MRS(r_1,x_2) > \widetilde{MRS}(r_1,x_2)$ for all $x_2 > r_2$. Similar reasoning reveals that $MRS(r_1,x_2) < \widetilde{MRS}(r_1,x_2)$ for all $x_2 < r_2$. Now, consider pairs (x_1,r_2) with $x_1 > r_1$. The above expressions for the broad and narrow marginal rates of substitution reveal that the denominator of $\widetilde{MRS}(x_1,r_2)$ is equal to the denominator of $\widetilde{MRS}(r_1,r_2)$ and the numerator of $MRS(x_1,r_2)$ is equal to the numerator of $\widetilde{MRS}(x_1,r_2)$. Further-

more, by $\frac{\partial^2 u}{\partial x_1 \partial x_2} > 0$ we have that the denominator of $MRS(x_1, r_2)$ is larger than the denominator of $MRS(r_1, r_2)$. Together with $MRS(r_1, r_2) = \widetilde{MRS}(r_1, r_2)$ this implies that $MRS(x_1, r_2) < \widetilde{MRS}(x_1, r_2)$ for all $x_1 > r_1$. Similar reasoning reveals that $MRS(x_1, r_2) > \widetilde{MRS}(x_1, r_2)$ for all $x_1 < r_1$. Finally, the full claim presented in the proposition follows by convexity of preferences as implied by positive interactions. □

A.1.2 Proof of Proposition 2 (Narrow optimum)

Proof. Focus on the case $\frac{\partial^2 u}{\partial x_1 \partial x_2} > 0$. The proof for $\frac{\partial^2 u}{\partial x_1 \partial x_2} < 0$ proceeds analogously. Since x^* and $\tilde{x}$ are interior solutions and $r \neq x^*$, it must hold that $MRS(x_1^*, x_2^*) = \frac{p_1}{p_2}$, $\widetilde{MRS}(\tilde{x}_1, \tilde{x}_2) = \frac{p_1}{p_2}$, and $MRS(r_1, r_2) \neq \frac{p_1}{p_2}$.

Now, suppose $MRS(r_1, r_2) < \frac{p_1}{p_2}$. Since $MRS(r_1, r_2) = \widetilde{MRS}(r_1, r_2)$, this holds iff $\widetilde{MRS}(r_1, r_2) < \frac{p_1}{p_2}$. Since x^* and $\tilde{x}$ are interior solutions and $w = p_1 r_1 + p_2 r_2$, $MRS(r_1, r_2) < \frac{p_1}{p_2}$ and $\widetilde{MRS}(r_1, r_2) < \frac{p_1}{p_2}$ imply that $x_1^*, \tilde{x}_1 < r_1$ and $x_2^*, \tilde{x}_2 > r_2$. Thus, by Proposition 1 $\frac{\partial^2 u}{\partial x_1 \partial x_2} > 0 \Rightarrow MRS(x_1^*, x_2^*) > \widetilde{MRS}(x_1^*, x_2^*)$ and $MRS(\tilde{x}_1, \tilde{x}_2) > \widetilde{MRS}(\tilde{x}_1, \tilde{x}_2)$. As $\frac{p_1}{p_2} = MRS(x_1^*, x_2^*) > \widetilde{MRS}(x_1^*, x_2^*)$ and $MRS(\tilde{x}_1, \tilde{x}_2) > \widetilde{MRS}(\tilde{x}_1, \tilde{x}_2) = \frac{p_1}{p_2}$ it must therefore hold that $\frac{\partial^2 u}{\partial x_1 \partial x_2} > 0 \Rightarrow d(r, x^*) < d(r, \tilde{x})$.

Suppose instead $MRS(r_1, r_2) > \frac{p_1}{p_2}$. Since $MRS(r_1, r_2) = \widetilde{MRS}(r_1, r_2)$ this holds iff $\widetilde{MRS}(r_1, r_2) > \frac{p_1}{p_2}$. Since x^* and $\tilde{x}$ are interior solutions and $w = p_1 r_1 + p_2 r_2$, $MRS(r_1, r_2) > \frac{p_1}{p_2}$ and $\widetilde{MRS}(r_1, r_2) > \frac{p_1}{p_2}$ imply that $x_1^*, \tilde{x}_1 > r_1$ and $x_2^*, \tilde{x}_2 < r_2$. Thus, by Proposition 1 $\frac{\partial^2 u}{\partial x_1 \partial x_2} > 0 \Rightarrow MRS(x_1^*, x_2^*) < \widetilde{MRS}(x_1^*, x_2^*)$ and $MRS(\tilde{x}_1, \tilde{x}_2) < \widetilde{MRS}(\tilde{x}_1, \tilde{x}_2)$. As $\frac{p_1}{p_2} = MRS(x_1^*, x_2^*) < \widetilde{MRS}(x_1^*, x_2^*)$ and $MRS(\tilde{x}_1, \tilde{x}_2) < \widetilde{MRS}(\tilde{x}_1, \tilde{x}_2) = \frac{p_1}{p_2}$ it must therefore hold that $\frac{\partial^2 u}{\partial x_1 \partial x_2} > 0 \Rightarrow d(r, x^*) < d(r, \tilde{x})$. □

A.1.3 Proof of Proposition 3 (Exchange economy)

Proof. For any elements x and $\tilde{x}$ of the respective broad and narrow cores, we have $MRS^1(x^1) = MRS^2(x^2)$ and $\widetilde{MRS}^1(\tilde{x}^1) = \widetilde{MRS}^2(\tilde{x}^2)$.

Focus first on $\frac{\partial^2 u^i}{\partial x_1^i \partial x_2^i} > 0$ for $i = 1, 2$ and ω such that $MRS^1(\omega^1) > MRS^2(\omega^2)$. From Proposition 1 we know that since $r^i = \omega^i$, $MRS^i(\omega^i) = \widetilde{MRS}^i(\omega^i)$ for $i = 1, 2$. Therefore, $MRS^1(\omega^1) > MRS^2(\omega^2)$ implies $\widetilde{MRS}^1(\omega^1) > \widetilde{MRS}^2(\omega^2)$.

By $MRS^1(\omega^1) > MRS^2(\omega^2)$ and $\widetilde{MRS}^1(\omega^1) > \widetilde{MRS}^2(\omega^2)$ it must hold for any interior broad and narrow core allocations x and $\tilde{x}$, that $x_1^1 \geq \omega_1^1$, $x_2^1 \leq \omega_2^1$,

$\tilde{x}_1^1 \geq \omega_1^1$, and $\tilde{x}_2^1 \leq \omega_2^1$, with one of the two inequalities concerning x and $\tilde{x}$ holding strictly.

Now, consider any allocation y with $y_1^1 \geq \omega_1^1$ and $y_2^1 \leq \omega_2^1$, implying $y_1^2 \leq \omega_1^2$ and $y_2^2 \geq \omega_2^2$, where one of the two inequalities holds strictly. By Proposition 1 we have $MRS^1(y^1) < \widetilde{MRS}^1(y^1)$ and $MRS^2(y^2) > \widetilde{MRS}^2(y^2)$.

Thus, starting from the initial endowment allocation ω, increasing the amount of good 1 allocated to person 1 while decreasing the amount of good 2 allocated to person 1 reduces the difference between the broad marginal rates of substitution of persons 1 and 2 faster than the difference between the narrow marginal rates of substitution of persons 1 and 2. Therefore, it must hold that at any allocation in the broad core $x = (x^1, x^2)$, $\widetilde{MRS}^1(x^1) > \widetilde{MRS}^2(x^2)$ while at any allocation in the narrow core $\tilde{x} = (\tilde{x}^1, \tilde{x}^2)$, $MRS^1(\tilde{x}^1) < MRS^2(\tilde{x}^2)$, such that the Euclidean distance between the initial endowment allocation ω and any allocation in the broad core x, $d(x,\omega) = \sqrt{(\omega_1^1 - x_1^1)^2 + (\omega_2^1 - x_2^1)^2}$, is smaller than the Eucleadian distance between the intital endowment allocation ω and any allocation in the narrow core $\tilde{x}$, $d(\omega,\tilde{x}) = \sqrt{(\omega_1^1 - \tilde{x}_1^1)^2 + (\omega_2^1 - \tilde{x}_2^1)^2}$.

Focus now on $\frac{\partial^2 u^i}{\partial x_1^i \partial x_2^i} > 0$ for $i = 1,2$ and ω such that $MRS^1(\omega^1) < MRS^2(\omega^2)$. From Proposition 1 we know that since $r^i = \omega^i$, $MRS^i(\omega^i) = \widetilde{MRS}^i(\omega^i)$ for $i = 1,2$. Therefore $MRS^1(\omega^1) < MRS^2(\omega^2)$ implies $\widetilde{MRS}^1(\omega^1) < \widetilde{MRS}^2(\omega^2)$.

By $MRS^1(\omega^1) < MRS^2(\omega^2)$ and $\widetilde{MRS}^1(\omega^1) < \widetilde{MRS}^2(\omega^2)$ it must hold for any interior broad and narrow core allocations x and $\tilde{x}$, that $x_1^1 \leq \omega_1^1$ and $x_2^1 \geq \omega_2^1$, respectively $\tilde{x}_1^1 \leq \omega_1^1$ and $\tilde{x}_2^1 \geq \omega_2^1$, with one of each of the two inequalities holding strictly.

Now, consider any allocation y with $y_1^1 \leq \omega_1^1$ and $y_2^1 \geq \omega_2^1$, implying $y_1^2 \geq \omega_1^2$ and $y_2^2 \leq \omega_2^2$, where one of the two inequalities holds strictly. By Proposition 1 we have $MRS^1(y^1) > \widetilde{MRS}^1(y^1)$ and $MRS^2(y^2) < \widetilde{MRS}^2(y^2)$.

Thus, starting from the initial endowment allocation ω, decreasing the amount of good 1 allocated to person 1 while increasing the amount of good 2 allocated to person 1 reduces the difference between the broad marginal rates of substitution of consumers 1 and 2 faster than the difference between the narrow marginal rates of substitution of consumers 1 and 2. Therefore, it must hold that at any allocation in the broad core x, $\widetilde{MRS}^1(x) < \widetilde{MRS}^2(x)$ while at any allocation in the narrow core $\tilde{x}$, $MRS^1(\tilde{x}) > MRS^2(\tilde{x})$, such that the Euclidean distance between the initial endowment allocation ω and any allocation in the broad core x, $d(x,\omega) = \sqrt{(\omega_1^1 - x_1^1)^2 + (\omega_2^1 - x_2^1)^2}$, is larger

than the Eucleadian distance between the intital endowment allocation ω and any allocation in the narrow core $\tilde{x}$, $d(\omega,\tilde{x}) = \sqrt{(\omega_1^1 - \tilde{x}_1^1)^2 + (\omega_2^1 - \tilde{x}_2^1)^2}$.

The proof for $\frac{\partial^2 u^i}{\partial x_1^i \partial x_2^i} < 0$ for $i = 1,2$ proceeds analogously. □

A.2 Welfare-based altruism

A.2.1 Relation to social norms and "social appropriateness"

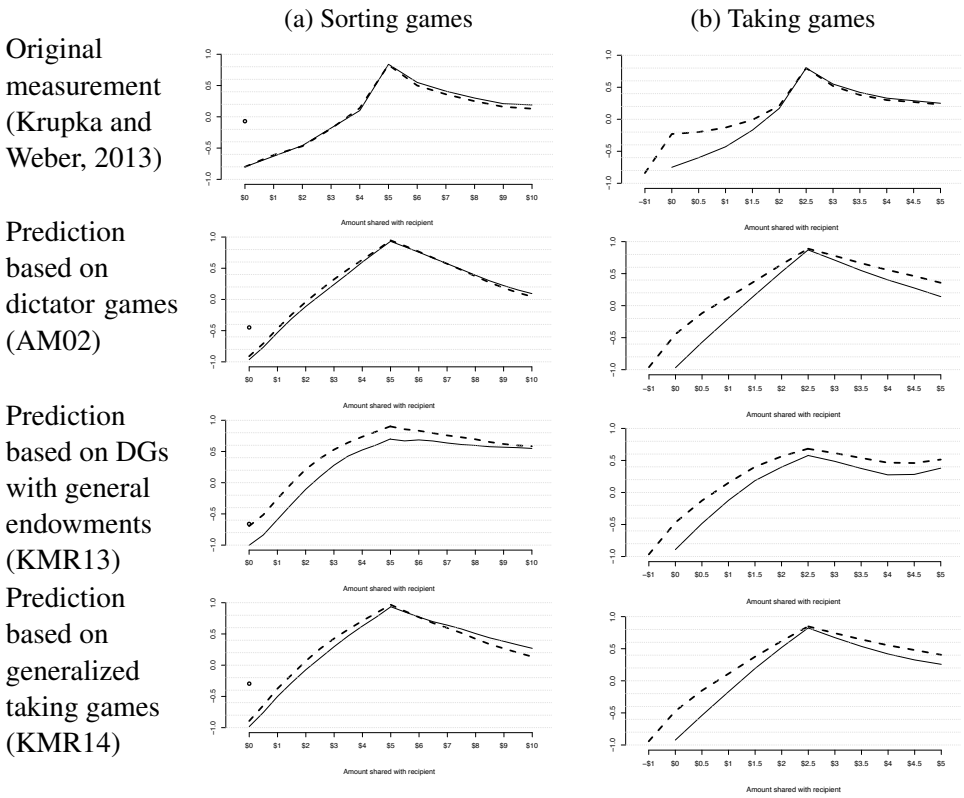
Starting with Krupka and Weber (2013), a growing literature relates giving observed in experiments to norm compliance. Subjects are assumed to have a common understanding of the "social appropriateness" of options, which in turn affects dictator behavior and is a function of the social norms applying in a given context. In a novel experimental design, Krupka and Weber measure social appropriateness by having (third) subjects play a coordination game—asking each subject how "socially appropriate" the available options are in the eyes of their co-players and paying a prize to all subjects picking the modal response. The mean of all appropriateness ratings is mapped into a measure $s_x \in [-1,1]$ for all options x, with $s_x = -1$ indicating highly inappropriate and $s_x = 1$ indicating highly appropriate options. Krupka and Weber then examine if a utility function of the form

$$u_x = \pi_x + \alpha s_x \tag{15}$$

fits behavior observed in earlier dictator game experiments, using the weight α as a free parameter. While statistical tests supporting the results are not provided, the plots in Krupka and Weber (2013) suggest a good fit after calibrating α. This finding has been interpreted as indicating that behavior is norm-guided, rather than being payoff or welfare concerned as assumed in earlier work. In the following, we clarify the relation of our findings to those of Krupka and Weber (2013) and subsequent work, to discuss how we may think of welfare-based altruism as a foundation of norm-guided giving.

To this end, let us recap two main results. Krupka and Weber convincingly demonstrate that experimental subjects are able to predict behavior in taking and sorting games, a feat that existing behavioral models struggled to achieve. We have shown that welfare-based altruism also allows to predict behavior, and hence our conjecture: the two are likely to correlate. A post-hoc straightforward approach would be to take our predictions of utility u_x across options, the respectively induced payoffs π_x, and to then compute social appropriateness s_x by inverting Eq. (15) for all options x. We skip this fairly unintelligible exercise and evaluate whether social appropriateness may be deduced from first principles.

Figure A.2.1: Relation of experimentally measured "social appropriateness" (Krupka and Weber) to the Rawlsian prediction following from our estimates



(c) Correlation between observed and predicted appropriateness

Predictions based on ...	Sorting games		Taking games	
	Spearman-ρ	p-value	Spearman-ρ	p-value
Dictator games (AM02)	0.641	(0.001)	0.738	(0)
Gen endowments (KMR13)	0.667	(0.001)	0.766	(0)
Taking games (KMR14)	0.644	(0.001)	0.751	(0)

Note: The "sorting games" compare appropriateness in a standard dictator game with endowments of 10 for the dictator and 0 for the recipient to appropriateness in a sorting game where the dictator game is succeeded by giving the dictator the option to sort out at costs of 1. The "taking games" compare appropriateness in a standard dictator game with endowments of 10 for the dictator and 5 for the recipient to appropriateness in a taking game where the dictator game may alternatively take one currency unit from the recipient's endowment. The plots follow Krupka and Weber: solid lines represent the social appropriateness in the standard dictator games and dashed lines represent social appropriateness in the sorting and taking games, respectively. The single "dot" in the sorting games reflects the appropriateness of sorting out.

Krupka and Weber (2013) interpret social appropriateness as reflecting the social norm that dictators facing a specific dictator game trade off with their self-interest. They argue that since their elicitation method (i) makes uninvolved subjects rate actions rather than outcomes and (ii) incentivizes subjects to rate in accordance with what they regard as a socially shared assessment, the resulting appropriateness ratings satisfy the two main characteristics of a social norm as defined by Elster (1989). These defining features of social norms are closely related to the "social contract" of Rawls (1971), which specifies a standard for social and distributive justice that "free and rational persons concerned to further their own interests would accept in an initial position of equality" (p. 11). The idea is that the members of a society would unanimously agree to the social contract if they met behind the "veil of ignorance", a hypothetical place where they are unaware of their positions in society (see also Konow, 2003). According to Rawls (1971) the social contract emerging in such a situation would prescribe a distribution that equalizes individual welfares unless inequality is to the advantage of the individual with the minimum welfare. While for obvious reasons an experimental test of Rawls hypothesis can never be perfect, Krupka and Weber's subjects share some central characteristics with Rawls' society members behind the veil of ignorance. They can be thought of as impartial since they are uninvolved while they are part of the same society as the involved players. Furthermore, they are incentivized to find an agreement instead of simply voicing their opinions. Therefore, looking at Krupka and Weber's social appropriateness ratings through the lense of our welfare-based altruism model allows us to test the Rawlsian hypothesis of social welfare being the minimum of all individual welfares, joint with the assertion that social appropriateness simply transforms social welfare to a scale ranging from highly inappropriate (-1) to highly appropriate (1).

Since our welfare-based approach directly builds on individual welfares v_1 and v_2, we are able to directly test the asserted Rawlsian link between appropriateness and welfares—simply by predicting individual welfares for all options in the sorting and taking games analyzed by Krupka and Weber, taking the minimum of v_1 and v_2 across options, and rescaling such that a measure ranging from -1 to $+1$ results. Specifically, we predict the social appropriateness ratings for both taking and sorting games analyzed by Krupka and Weber based on our estimates from each of the three experiments analyzed

before (AM02, MKR13, and KMR14). This yields 3×2 profiles of appropriateness ratings, which we then relate to the measurements of Krupka and Weber.[29] The results are reported in Figure A.2.1 and strongly corroborate the relation of social appropriateness and Rawlsian welfare asserted already by Krupka and Weber. The correlation between the out-of-sample predictions and the in-sample measurements of Krupka and Weber is very high, around 0.65 in sorting games and around 0.75 in taking games, regardless of the data set which the prediction is based on. We therefore conclude as follows.

Result 6. *Krupka and Weber's measure of social appropriateness strongly correlates with the Rawlsian notion of welfare, based on out-of-sample predictions of individual welfares derived from the above model of welfare-based altruism.*

That is, social appropriateness is founded in welfare concerns in the intuitive Rawlsian manner alluded to by Krupka and Weber. It seems futile to ask which came first, welfare concerns or social appropriateness/social norms, they rather appear to be two sides of the same coin. The received interpretation that giving reflects context-dependent social norms rather than more fundamental payoff and welfare concerns seems premature, but so would the opposite. From a practical point of view, both approaches seem to have distinctive strengths. Analyses relating behavior to social appropriateness need not be concerned with individual preferences and can focus on the picture at large. In turn, the behavioral foundation in welfare concerns has an independent axiomatic foundation in established behavioral principles, which greatly facilitates application across contexts, and the implied S-shape of individual welfares has been observed in many contexts, which promises reliable predictions and policy recommendations out-of-sample.

[29] Specifically, for each subject in our in-sample experiments (AM02, KMR13, KMR14), we determine the individual welfares if that subject would play either role, v_1 and v_2. We then assume that an impartial observer in the sense of Krupka and Weber determines appropriateness as follows: Across dictators, what is their average individual welfare from choosing x conditional on choosing x in the first place. Across recipients, what is their average individual welfare from getting x conditionally on being confronted with x in the first place (which is an empty condition, stated only for symmetry). The lesser of these conditional expectations is the unscaled Rawlsian appropriateness of each option, and rescaling to $[-1,1]$ across options yields our out-of-sample prediction for Krupka-Weber appropriateness.

A.2.2 Proof of Proposition 4

Step 1 Existence of a continuous, additively separable utility representation.

Axioms 1–2 imply existence of a continuous utility representation (see e.g. Rubinstein, 2012, chap. 4). In addition Axiom 3 implies existence of an additively separable utility representation, see Theorem III.4.1 in Wakker (1989) for each context $\pi \in \Pi$. That is, there exists a family of functions $\{v_{\pi,i} : \mathbb{R} \to \mathbb{R}\}_{\pi\in\Pi, i\leq n}$ such that $\pi(x) \succsim_\pi \pi(y) \Leftrightarrow u_\pi(x) \geq u_\pi(y)$ for all $x, y \in X$ and $\pi \in \Pi$ with

$$u_\pi(x') = \sum_{i\leq n} v_{\pi,i}\big(\pi_i(x')\big) \tag{16}$$

for all $x' \in X, \pi \in \Pi$. For later reference, Wakker's Theorem III.4.1 also establishes that all additively separable representations $\tilde{u}_\pi$ of $\succsim_\pi$ are positive affine transformations of one another. Also note that the representations obtained so far may be context dependent.

Step 2 Context independent (v_i) by narrow or broad bracketing.

We show that additionally assuming either Axiom 5 or Axiom 6 implies that there exists a family of functions $\{v_i : \mathbb{R} \to \mathbb{R}\}_{i\leq n}$ and $r : \Pi \to \mathbb{R}^n$ such that

$$u_\pi(x) = \sum_{i\leq n} v_i\big(\pi_i(x)\big) \qquad \text{(Broad bracketing)}$$

$$u_\pi(x) = \sum_{i\leq n} v_i\big(\pi_i(x) - r_i(\pi)\big) \qquad \text{(Narrow bracketing)}$$

represent $\succsim_\pi$ for all $\pi \in \Pi$.

Narrow bracketing: Fix any $\pi \in \Pi$, any $x \in X$. We show that if the preferences obey Axioms 1-3 and Axiom 6, then they admit the claimed representation for any function $r : \Pi \to \mathbb{R}^n$ with $r(\pi') = \pi'(x) - \pi(x)$ for all $\pi' \in \Pi$. Fix this r and any $\pi' \in \Pi$. By Assumption 4.1, $r(\pi') = \pi'(x') - \pi(x')$ for all $x' \in X$. Also note $r(\pi) = \mathbf{0}$. By narrow bracketing, we know that $\succsim_\pi$ is equivalent to $\succsim_{\pi'}$, and using the utility functions obtained in Step 1, this implies

$$\sum_{i\leq n} v_{\pi,i}\big(\pi_i(x)\big) \geq \sum_{i\leq n} v_{\pi,i}\big(\pi_i(y)\big) \quad \Leftrightarrow \quad \sum_{i\leq n} v_{\pi',i}\big(\pi'_i(x)\big) \geq \sum_{i\leq n} v_{\pi',i}\big(\pi'_i(y)\big) \tag{17}$$

for all $x, y \in X$. Since $\pi(x) = \pi'(x) - r_i(\pi')$ for all x by construction of r, this yields

$$\sum_{i\leq n} v_{\pi,i}\big(\pi'_i(x) - r_i(\pi')\big) \geq \sum_{i\leq n} v_{\pi,i}\big(\pi'_i(y) - r_i(\pi')\big) \quad \Leftrightarrow \quad \sum_{i\leq n} v_{\pi',i}\big(\pi'_i(x)\big) \geq \sum_{i\leq n} v_{\pi',i}\big(\pi'_i(y)\big)$$

for all $x, y \in X$. Since this holds true for all $\pi' \in \Pi$, the claim is established using $v_i = v_{\pi,i}$ for all $i \leq n$. Note again that u (and thus v) is unique up to positive affine transformation, and that for any π', if $\pi' = \pi + c$, then $r(\pi') = r(\pi + c) = r(\pi) + c$ by construction.

Broad bracketing: For each context π, fix value functions $(v_{\pi,i})$ representing $\succsim_\pi$ as obtained in Step 1. By broad bracketing, value functions $(\tilde{v}_\pi)_{\pi\in\Pi}$ representing preferences exist such that for all $x, x' \in X$ and all $\pi, \pi' \in \Pi$,

$$\pi(x) = \pi'(x') \qquad \Leftrightarrow \qquad \sum_{i\leq n} \tilde{v}_{\pi,i}\big(\pi_i(x)\big) = \sum_{i\leq n} \tilde{v}_{\pi',i}\big(\pi'_i(x')\big).$$

Given any such family $(\tilde{v}_{\pi,i})$, and using $P = \cup_{\pi\in\Pi}\pi[X]$, define the functions $\{v_i : P \to \mathbb{R}\}_{i\leq n}$ such that for all $p \in P$,

$$v_i(p_i) = \tilde{v}_{\pi,i}(\pi_i(x)) \qquad \text{for some } (\pi, x) : p = \pi(x).$$

Adequate (π, x) exist for all $p \in P$ by construction of P. By broad bracketing, this implies

$$v_i(p_i) = \tilde{v}_{\pi,i}(\pi_i(x)) \qquad \text{for all } (\pi, x) : p = \pi(x),$$

thus establishing that (v_i) allow to represent the preferences as claimed. Since all $(\tilde{v}_{\pi,i})$ must be positive affine transformations of $(v_{\pi,i})$, which are continuous, both $(\tilde{v}_{\pi,i})$ and (v_i) also are families of continuous functions.

Step 3 Normalize (v_i, r_i) in relation to π^0.
Fix the scaling-invariant context π^0, which exists by Axiom 4, we know from Step 2

$$u_{\pi^0}(x) = \sum_{i\leq n} v_i\big(\pi^0_i(x) - r_i(\pi^0)\big), \tag{18}$$

which in turn implies that we can translate (v_i) and (r_i) such that

$$u_{\pi^0}(x) = \sum_{i \leq n} v_i\big(\pi_i^0(x)\big), \tag{19}$$

i.e. such that $r_i(\pi^0) = 0$ for all $i \leq n$. Again, by $r(\pi) + c = r(\pi + c)$ for all $c \in \mathbb{R}^n$, this implies $r(\pi) = c$ if $\pi = \pi^0 + c$, for all $\pi \in \Pi$. Note that given this translation, we can analyze narrow bracketing and broad bracketing in a uniform manner when focusing on π^0 (i.e. we do not have to include r_i as $r_i(\pi^0) = 0$).

Step 4 Using scaling invariance to fix the functional form.
By Axiom 4, preferences in context π^0 are scaling invariant. That is, for all $\lambda > 0$, define $u_{\lambda\pi^0} : X \to \mathbb{R}$ such that

$$u_{\lambda\pi^0}(x) = \sum_{i \leq n} v_i\big(\lambda\pi_i^0(x)\big), \tag{20}$$

for all λ, x, and we obtain

$$u_{\lambda\pi^0}(x) \geq u_{\lambda\pi^0}(y) \Leftrightarrow \quad u_{\pi^0}(x) \geq u_{\pi^0}(y) \quad \Leftrightarrow \quad \pi^0(x) \succsim_{\pi^0} \pi^0(y). \tag{21}$$

By aforementioned Theorem III.4.1 of Wakker (1989) this implies that $u_{\lambda\pi^0}$ is a positive affine transformation of u_{π^0}, i.e. there exist $a : \mathbb{R}_+ \to \mathbb{R}$ and $b : \mathbb{R}_+ \to \mathbb{R}^n$ such that

$$v_i\big(\lambda\pi_i^0(x)\big) = v_i\big(\pi_i^0(x)\big) \cdot a(\lambda) + b_i(\lambda) \tag{22}$$

for all $i \in N$, $x \in X$, $\lambda > 0$. Now, define $X_i^+ = \{x \in X \,|\, \pi_i^0(x) > 0\}$ as well as $\tilde{\lambda} = \log\lambda$, $\tilde{v}_i : \mathbb{R} \to \mathbb{R}$ such that $\tilde{v}_i(\log p) = v_i(p)$ for all $p > 0$, and $\tilde{\pi}_i^0(x) = \log\pi_i^0(x)$ for all $x \in X_i^+$, which yields

$$\tilde{v}_i\big(\tilde{\lambda} + \tilde{\pi}_i^0(x)\big) = \tilde{v}_i\big(\tilde{\pi}_i^0(x)\big) \cdot a(\tilde{\lambda}) + b_i(\tilde{\lambda}). \tag{23}$$

By continuity of v_i we obtain continuity of $\tilde{v}_i$, and since the payoff image $\pi^0[X]$ is a cone in $\mathbb{R}^n$ with all dimensions being essential, it has positive volume in $\mathbb{R}^n$, i.e. $\pi_i^0[X]$ is an interval of positive length for all dimensions i. Hence, Theorem 1 and Corollary 1 of Aczél (1966, p. 150) imply that all

solutions of this (Pexider) functional equation satisfy either

$$\tilde{v}_i\big(\tilde{\pi}_i^0(x)\big) = \alpha \cdot \tilde{\pi}_i^0(x) + \gamma \qquad \text{or} \qquad \tilde{v}_i\big(\tilde{\pi}_i^0(x)\big) = \alpha \cdot e^{\beta\,\tilde{\pi}_i^0(x)} + \gamma$$

with $\alpha \neq 0$ and β, γ being arbitrary constants, and inverting the variable substitutions,

$$v_i(\pi_i^0(x)) = \alpha \cdot \log \pi_i^0(x) + \gamma \qquad \text{or} \qquad v_i(\pi_i^0(x)) = \alpha \cdot \big(\pi_i^0(x)\big)^{\beta} + \gamma.$$

To distinguish the constants from constants in other dimensions, we rewrite

$$v_i(\pi_i^0(x)) = \alpha_i^+ + \beta_i^+ \cdot \log \pi_i^0(x) \quad \text{or} \quad v_i(\pi_i^0(x)) = \alpha_i^+ \cdot \big(\pi_i^0(x)\big)^{\beta_i^+} + \gamma_i^+$$

for all $x \in X_i^+$. Next, define $X_i^- = \{x \in X \,|\, \pi_i^0(x) < 0\}$, and apply the same line of arguments to $-\pi_i^0(x)$ for all $x \in X_i^-$, which yields

$$v_i(\pi_i^0(x)) = \alpha_i^- + \beta_i^- \cdot \log\big(-\pi_i^0(x)\big) \quad \text{or} \quad v_i(\pi_i^0(x)) = -\alpha_i^- \cdot \big(-\pi_i^0(x)\big)^{\beta_i^-} + \gamma_i^-$$

for all $x \in X_i^-$, again with $\alpha_i^- \neq 0$ and β_i^-, γ_i^- being arbitrary constants.

Step 5 Using continuity and Eq. (22) to normalize the parameters.
In the following, we refer to the two possible forms of the value function v_i as power form and logarithmic form (in the obvious manner). By continuity, the logarithmic form is feasible only if $\pi_i(x) > 0$ for all $x \in X$, implying that the second branch is never taken. Hence, for all $i \leq n$ and all $x \in X$,

$$v_i(\pi_i^0(x)) = \alpha_i^+ + \beta_i^+ \cdot \log\big(\pi_i^0(x)\big),$$

and we can set $\alpha_i^+ = 0$ for all i by applying a positive affine transformation (recalling that the value functions are unique up to positive affine transformation). This establishes the claim for the logarithmic form in context π^0, noting that α_i^+ and β_i^+ are switched (for the logarithmic form) in the formulation of the proposition for notational convenience.

Regarding the power form of the value function, rescaling payoffs we

obtain

$$\forall x \in X_i^+ : \; v_i(\lambda\pi_i^0(x)) = \alpha_i^+ \cdot \big(\lambda\pi_i^0(x)\big)^{\beta_i^+} + \gamma_i^+ = \alpha_i^+ \cdot \big(\pi_i^0(x)\big)^{\beta_i^+} \cdot \lambda^{\beta_i^+} + \gamma_i^+$$
$$\forall x \in X_i^- : \; v_i(\lambda\pi_i^0(x)) = -\alpha_i^- \cdot \big(-\lambda\pi_i^0(x)\big)^{\beta_i^-} + \gamma_i^- = -\alpha_i^- \cdot \big(-\pi_i^0(x)\big)^{\beta_i^-} \cdot \lambda^{\beta_i^-} + \gamma_i^-,$$

which is compatible with Eq. (22) only if $\beta_i^+ = \beta_i^- = \beta$ and $\gamma_i^+ = \gamma_i^- = \gamma_i$ for all i. Given the latter, we can again set $\gamma_i^+ = \gamma_i^- = 0$ by a positive affine transformation. As a result, the claim for both the logarithmic form and the power form is established for context π^0.

Step 6 Extension to contexts $\pi \neq \pi^0$.
Narrow bracketing: Fix any $\pi \in \Pi$, and let $c \in \mathbb{R}^n$ such that $\pi = \pi^0 + c$. By Step 2, we know that

$$u_\pi(x) = \sum_{i \leq n} v_i\big(\pi_i(x) - r_i(\pi)\big)$$

represents $\succsim_\pi$ with v_i as characterized in the previous step and r_i as characterized in Steps 2 and 3. Since the representation is unique up to positive affine transformation, we can add arbitrary constants, and the claim is established for any context $\pi \in \Pi$.

Broad bracketing: By Step 2, the utility representation characterized in Step 5 applies uniformly to all contexts. □

A.2.3 Proofs of Propositions 5 and 6

A.2.3.1 Optimal choice of a regular dictator Δ in a given game Γ with $P_1 = [0, B]$

Note that since for this part of the proof the game Γ is kept fixed, we drop the game index on the utility function and write r_i instead of $r_i(\Gamma)$ for the reference points. Then dictator Δ's utility function in game Γ is given by

$$u(p_1) = \frac{1}{\beta} \times \begin{Bmatrix} (p_1 - r_1)^\beta & \text{if } p_1 \geq r_1 \\ -\delta(r_1 - p_1)^\beta & \text{if } p_1 < r_1 \end{Bmatrix} + \frac{\alpha}{\beta} \times \begin{Bmatrix} (p_2(p_1) - r_2)^\beta & \text{if } p_2(p_1) \geq r_2 \\ -\delta(r_2 - p_2(p_1))^\beta & \text{if } p_2(p_1) < r_2 \end{Bmatrix}$$

where $p_2(p_1) = t(B - p_1)$.

Step 1 Dictator Δ never chooses p_1 such that $p_1 < r_1$ and $p_2(p_1) < r_2$.

By satisfiability of reference points and $P_1 = [0, B]$ dictator Δ can always choose $p_1 \in P_1$ such that $p_1 \geq r_1$ and $p_2(p_1) \geq r_2$. This yields utility $u(p_1) = (p_1 - r_1)^\beta/\beta + \alpha(t(B - p_1) - r_2)^\beta/\beta \geq 0$ where the inequality follows by weak efficiency concerns ($0 < \beta < 1$). Choosing $p_1' \in P_1$ such that $p_1' < r_1$ and $p_2(p_1') < r_2$ instead yields utility $u(p_1') = -\delta(r_1 - p_1)^\beta/\beta - \alpha\delta(r_2 - t(B - p_1))^\beta/\beta < 0$.

Thus, we can restrict attention to the regions where at most one of the two players is in the loss-domain, i.e. does not reach her reference point. In the following we will first determine the local optima for dictator Δ in each of the three remaining regions. Then we can determine the global optimum by comparing utilities of the local optima.

Step 2 Local optimum in region 1: $p_1 \in [r_1, B - \frac{1}{t} r_2]$ ($\Leftrightarrow p_1 \geq r_1$ and $p_2(p_1) \geq r_2$)

The utility function that applies is

$$u^{(1)}(p_1) = (p_1 - r_1)^\beta + \alpha \cdot (t(B - p_1) - r_2)^\beta$$

Differentiating $u^{(1)}$ with respect to p_1 we get

$$\frac{du^{(1)}}{dp_1} = \beta (p_1 - r_1)^{\beta - 1} - \alpha \beta t (t (B - p_1) - r_2)^{\beta - 1}$$

which yields the first order condition

$$(p_1 - r_1)^{1-\beta} (t (B - p_1) - r_2)^{\beta - 1} = \frac{1}{\alpha t} \quad \Leftrightarrow \quad \frac{t (B - p_1) - r_2}{p_1 - r_1} = (\alpha t)^{\frac{1}{1-\beta}}.$$

and the solution

$$p_1^+(\Gamma) = \frac{B + c_\alpha r_1 - r_2/t}{c_\alpha + 1} \quad \text{and} \quad p_2^+(\Gamma) = \frac{t c_\alpha (B - r_1) + r_2}{c_\alpha + 1}$$

using $c_\alpha := (\alpha t^\beta)^{\frac{1}{1-\beta}}$. Note that for $p_1 = B - \frac{1}{t} r_2$ and $p_1 = r_1$ the above first order condition is not defined because the utility function exhibits kinks at these points. We have $p_1^+(\Gamma) = B - \frac{1}{t} r_2 = r_1$ iff satisfiability is binding, i.e. $B - r_1 - \frac{1}{t} r_2 = 0$. By satisfiability we have $p_1^+(\Gamma) \in [r_1, B - \frac{1}{t} r_2]$ for all regular dictators Δ. Furthermore, the second order condition for $p_1^+(\Gamma)$ to be a maximum reduces to

$$t^{2-\beta} c_\alpha (1 + c_\alpha)^{1-\beta} (t(B - r_1) - r_2)^\beta > 0,$$

which is fulfilled for $p_1^+(\Gamma)$ by satisfiability, weak efficiency concerns ($0 < \beta < 1$), and $\alpha, t > 0$. Overall, we thus have for the local optimum in region 1

$$p_1^{(*)} = p_1^+(\Gamma).$$

Step 3 Local optimum in region 2: $p_1 \in (B - \frac{1}{t} r_2, B]$ ($\Leftrightarrow p_1 \geq r_1$ and $p_2 < r_2$)

The utility function that applies is

$$u^{(2)}(p_1) = (p_1 - r_1)^\beta - \delta\alpha \cdot (r_2 - t(B - p_1))^\beta$$

Differentiating $u^{(2)}$ with respect to p_1 we obtain

$$\frac{du^{(2)}}{dp_1} = \beta (p_1 - r_1)^{\beta - 1} - \delta\alpha\beta t \, (r_2 - t \, (B - p_1))^{\beta - 1}$$

which yields the first order condition

$$(p_1 - r_1)^{1-\beta} (r_2 - t(B - p_1))^{\beta - 1} = \frac{1}{\delta\alpha t} \quad \Leftrightarrow \quad \frac{r_2 - t(B - p_1)}{p_1 - r_1} = (\delta\alpha t)^{\frac{1}{1-\beta}}$$

and the solution

$$p_1^{(2)}(\Gamma) = \frac{B - \delta^{\frac{1}{1-\beta}} c_\alpha r_1 - r_2/t}{1 - \delta^{\frac{1}{1-\beta}} c_\alpha} \quad \text{and} \quad p_2^{(2)}(\Gamma) = \frac{t \delta^{\frac{1}{1-\beta}} c_\alpha (r_1 - B) + r_2}{1 - \delta^{\frac{1}{1-\beta}} c_\alpha}.$$

By satisfiability we have $p_1^{(2)} \in (B - \frac{1}{t}r_2, B]$ iff

$$\delta^{\frac{1}{1-\beta}} c_\alpha \leq \frac{r_2}{t(B-r_1)} \Leftrightarrow \delta \leq \frac{1}{\alpha t^\beta} \left(\frac{r_2}{t(B-r_1)} \right)^{1-\beta}.$$

Using $\delta^{\frac{1}{1-\beta}} c_\alpha < 1$, the second order condition for $p_1^{(2)}(\Gamma)$ to be a maximum reduces to

$$t^{2-\beta} \delta^{\frac{1}{1-\beta}} c_\alpha (1 - \delta^{\frac{1}{1-\beta}} c_\alpha)^{1-\beta} (t(B-r_1) - r_2)^\beta < 0.$$

Thus, the second order condition does not hold for any $p_1^{(2)}(\Gamma) \in (B - \frac{1}{t}r_2, B]$ by satisfiability and weak efficiency concerns ($0 < \beta < 1$). It follows that the local optimum is either $p_1 = B - \frac{1}{t}r_2$ or $p_1 = B$ depending on whether $u^{(2)}(B - \frac{1}{t}r_2) \geq u^{(2)}(B)$, a condition which reduces to

$$\delta \geq c_\alpha^{\beta-1} \left(\left(\frac{t(B-r_1)}{r_2} \right)^\beta - \left(\frac{t(B-r_1)}{r_2} - 1 \right)^\beta \right).$$

Overall, we thus have for the local optimum in region 2

$$p_1^{(*)} = \begin{cases} B - \frac{1}{t}r_2 & \text{if } \delta \geq c_\alpha^{\beta-1} \left(\left(\frac{t(B-r_1)}{r_2} \right)^\beta - \left(\frac{t(B-r_1)}{r_2} - 1 \right)^\beta \right) \\ B & \text{else.} \end{cases}$$

Step 4 Local optimum in region 3: $p_1 \in [0, r_1)$ ($\Leftrightarrow p_1 < r_1$ and $p_2(p_1) \geq r_2$)

The utility function that applies is

$$u^{(3)}(p_1) = -\delta \cdot (r_1 - p_1)^\beta + \alpha \cdot (t(B-p_1) - r_2)^\beta$$

Differentiating $u^{(3)}$ with respect to p_1 we obtain

$$\frac{du^{(3)}}{dp_1} = \delta\beta(r_1 - p_1)^{\beta-1} - \alpha\beta t(t(B-p_1) - r_2)^{\beta-1}$$

which yields the first order condition

$$(r_1 - p_1)^{1-\beta} (t(B-p_1) - r_2)^{\beta-1} = \frac{\delta}{\alpha t} \quad \Leftrightarrow \quad \frac{t(B-p_1) - r_2}{r_1 - p_1} = \left(\frac{\alpha t}{\delta} \right)^{\frac{1}{1-\beta}}$$

and the solution

$$p_1^{(3)}(\Gamma) = \frac{B - \delta^{1-\beta} c_\alpha r_1 - r_2/t}{1 - \delta^{1-\beta} c_\alpha} \quad \text{and} \quad p_2^{(3)}(\Gamma) = \frac{t\delta^{1-\beta} c_\alpha (r_1 - B) + r_2}{1 - \delta^{1-\beta} c_\alpha}$$

By satisfiability we have $p_1^{(3)} \in [0, r_1)$ iff $\delta^{1-\beta} c_\alpha \geq \frac{tB - r_2}{tr_1} \Leftrightarrow \delta \leq \alpha t^\beta \left(\frac{tr_1}{tB - r_2}\right)^{1-\beta}$. The second order condition for $p_1^{(3)}(\Gamma)$ to be a maximum reduces to

$$\frac{1}{\delta^{1-\beta} c_\alpha} \left(\frac{r_1 - B + r_2/t}{1 - \delta^{1-\beta} c_\alpha}\right)^{\beta-2} > \left(\frac{r_1 - B + r_2/t}{1 - \delta^{1-\beta} c_\alpha}\right)^{\beta-2},$$

which by satisfiability does not hold for any $p_1^{(3)}(\Gamma) \in [0, r_1)$. It follows that the optimum is either $p_1 = 0$ or $p_1 = r_1$ depending on whether $u^{(3)}(0) \geq u^{(3)}(r_1)$, a condition which reduces to

$$\delta \leq c_\alpha^{1-\beta} \left(\left(\frac{tB - r_2}{tr_1}\right)^\beta - \left(\frac{tB - r_2}{tr_1} - 1\right)^\beta \right).$$

Overall, we thus have for the local optimum in region 3

$$p_1^{(*)} = \begin{cases} 0 & \text{if } \delta \leq c_\alpha^{1-\beta} \left(\left(\frac{tB - r_2}{tr_1}\right)^\beta - \left(\frac{tB - r_2}{tr_1} - 1\right)^\beta \right) \\ r_1 & \text{else.} \end{cases}$$

Step 5 Reducing the set of candidate solutions for the global optimum

Using weak efficiency concerns ($0 < \beta < 1$) and $\alpha, t \geq 0$ we have $u(p_1^+(\Gamma)) \geq u(B - \frac{1}{t} r_2)$ and $u(p_1^+(\Gamma)) \geq u(r_1)$ for all regular dictators Δ, a result which obtains by simple rearrangement of the two inequalities. Thus, the remaining candidate solutions for the overall utility maximizer are $p_1 = p_1^+(\Gamma)$, $p_1 = B$, and $p_1 = 0$.

Furthermore, we have $u(p_1^+(\Gamma)) \geq u(0)$ iff

$$\delta \geq c_\alpha^{1-\beta} \left(\frac{tB - r_2}{tr_1}\right)^\beta - (c_\alpha + 1)^{1-\beta} \left(\frac{tB - r_2}{tr_1} - 1\right)^\beta. \tag{24}$$

From weak efficiency concerns ($0 < \beta < 1$) we can conclude that

$$c_\alpha^{1-\beta} < (c_\alpha + 1)^{1-\beta}.$$

Define $f(x) = x^{\beta}$, then weak efficiency concerns ($0 < \beta < 1$) imply that f is subadditive in the domain $\mathbb{R}^+$, i.e. $f(a) + f(b) \geq f(a+b) \forall a, b \geq 0$. Thus, using satisfiability and letting $a = \frac{tB - r_2}{tr_1} - 1$ and $b = 1$, we have

$$f(a) + f(b) = \left(\frac{tB - r_2}{tr_1} - 1\right)^{\beta} + 1^{\beta} \geq \left(\frac{tB - r_2}{tr_1}\right)^{\beta} = f(a+b)$$

implying

$$\left(\frac{tB - r_2}{tr_1}\right)^{\beta} - \left(\frac{tB - r_2}{tr_1} - 1\right)^{\beta} \leq 1.$$

Suppose $c_{\alpha}^{1-\beta} \leq 1$. In this case we can conclude that the lower bound for δ defined in (24) is lower or equal 1, which by weak loss aversion ($\delta \geq 1$) implies $u(p_1^+(\Gamma)) \geq u(0)$. Note that $c_{\alpha}^{1-\beta} \leq 1$ by weak altruism ($0 \leq \alpha \leq 1$) always holds under no efficiency gains from giving ($t \leq 1$) such that in this case the candidate solutions for the overall utility maximizer reduce further to $p_1 = p_1^+(\Gamma)$ and $p_1 = B$.

Finally, we have $u(p_1^+(\Gamma)) \geq u(B)$ iff

$$\delta \geq c_{\alpha}^{\beta-1} \left(\left(\frac{t(B - r_1)}{r_2}\right)^{\beta} - (c_{\alpha} + 1)^{1-\beta} \left(\frac{t(B - r_1)}{r_2} - 1\right)^{\beta} \right). \tag{25}$$

Suppose $c_{\alpha}^{1-\beta} > 1$. In this case by a similar argument as above we can conclude that the lower bound for δ defined in (25) is lower or equal 1, which by weak loss aversion ($\delta \geq 1$) implies $u(p_1^+(\Gamma)) \geq u(B)$. We can therefore conclude that under efficiency gains from giving ($t > 1$) the candidate solutions for the overall utility maximizer reduce to $p_1 = p_1^+(\Gamma)$ and $p_1 = B$ in case $c_{\alpha}^{1-\beta} \leq 1$ while they reduce to $p_1 = p_1^+(\Gamma)$ and $p_1 = 0$ in case $c_{\alpha}^{1-\beta} > 1$.

Step 6 Global optimum

For the global optimum we have to distinguish the following two cases:

- Case 1: $c_{\alpha}^{1-\beta} \leq 1$

$$p_1^* = \begin{cases} p_1^+(\Gamma) & \text{if } \delta \geq \delta^+(\Gamma) \\ B & \text{else.} \end{cases}$$

with

$$\delta^+(\Gamma) := c_\alpha^{\beta-1}\left(\left(\frac{t(B-r_1)}{r_2}\right)^\beta - (c_\alpha+1)^{1-\beta}\left(\frac{t(B-r_1)}{r_2}-1\right)^\beta\right)$$

- Case 2: $c_\alpha^{1-\beta} > 1$

$$p_1^* = \begin{cases} p_1^+(\Gamma) & \text{if } \delta \geq \delta^-(\Gamma) \\ 0 & \text{else.} \end{cases}$$

with

$$\delta^-(\Gamma) := c_\alpha^{1-\beta}\left(\frac{tB-r_2}{tr_1}\right)^\beta - (c_\alpha+1)^{1-\beta}\left(\frac{tB-r_2}{tr_1}-1\right)^\beta$$

Note that under no efficiency gains from giving ($t \leq 1$) only case 1 applies.

A.2.3.2 Establishing the comparative statics

Step 1 Non-convexity In any game Γ with $P_1 = [0,B]$ there are dictators with non-convex preferences.

Fix a game Γ with $P_1 = [0,B]$. Consider a dictator Δ with $\delta \leq \bar{\delta}(\Gamma)$ where

$$\bar{\delta}(\Gamma) := c_\alpha^{\beta-1}\left(\frac{r_2(\Gamma)}{t(B-r_1(\Gamma))}\right)^{1-\beta}.$$

We have shown in step 3 of A.2.3.1 that the utility function of this dictator attains a minimum at $p_1 = p_1^{(2)}(\Gamma) \in [B-r_2/t, B]$ and has no other local extrema in that region. Furthermore, we have shown in step 2 of A.2.3.1 that her utility function attains a maximum at $p_1 = p_1^+(\Gamma) \in [r_1, B-r_2/t]$ and has no other local extrema in that region. Consider options a and b with $p_1^a = B$ and $p_1^b = p_1^+(\Gamma)$. Construct option c by choosing $\lambda \in [0,1]$ such that $p_1^c = \lambda p_1^a + (1-\lambda)p_1^b = p_1^{(2)}(\Gamma)$. Then, for dictator Δ in game Γ there exists an option d with $p_1^d \in (p_1^+(\Gamma), B)$ such that $u_\Gamma(p_1^a) \geq u_\Gamma(p_1^d)$ and $u_\Gamma(p_1^b) \geq u_\Gamma(p_1^d)$ but $u_\Gamma(p_1^c) < u_\Gamma(p_1^d)$. Since u_Γ represents dictator Δ's preferences in game Γ, this implies that her preferences are non-convex.

We still have to show that in any game Γ with $P_1 = [0,B]$ there exist regular dictators with $\delta \leq \bar{\delta}(\Gamma)$. For any transfer rate t specified by Γ we can find (α,β) satisfying $0 \leq \alpha \leq 1$ and $0 < \beta < 1$ such that $c_\alpha^{1-\beta} \leq 1 \Leftrightarrow c_\alpha^{\beta-1} \geq 1$.

Given such (α,β), for any endowments (B_1,B_2) specified by Γ, we can find (w_1,w_2) in accordance with satisfiability resulting in reference points $r_1(\Gamma) = w_1B_1 + w_2B_2$ and $r_2(\Gamma) = t(w_1B_2 + w_2B_1)$ such that $r_2(\Gamma)/t(B - r_1(\Gamma))$ is close enough to 1 to make $\bar{\delta}(\Gamma) \geq 1$. Thus, given such (α,β,w_1,w_2), we can conclude that there exist δ satisfying weak loss aversion ($\delta \geq 1$) such that $\delta \leq \bar{\delta}(\Gamma)$.

Step 2 Taking options reduce giving both at the extensive and intensive margin Introducing a taking option turns some initial givers into takers and reduces average amounts given.

Consider two games $\Gamma = \langle B_1,B_2,P_1,t\rangle$ and $\Gamma' = \langle B_1,B_2,P_1',t\rangle$ with $B_2 > 0$ that are equivalent in every dimension except the choice set of the dictator. In Γ the choice set is restricted to $P_1 = [0,\max p_1]$ with $\max p_1 = B_1$ and in Γ' the choice set is extended to $p_1' = [0,\max p_1']$ with $B_1 < \max p_1' \leq B_1 + B_2$.

Moving from Γ to Γ' the only game parameter that changes is the maximum payoff for the dictator which rises from $\max p_1 = B_1$ to $\max p_1'$. As a result of this rise, the minimum payoff for the recipient adjusts accordingly, i.e. it falls from $\min p_2 = t(B_1 + B_2 - \max p_1) = tB_2$ to $\min p_2' = t(B_1 + B_2 - \max p_1')$. Therefore, the utility functions of a regular dictator Δ in Γ and Γ' differ in the players' reference points. We have

$$r_2(\Gamma) = t\left(B_2 + (1 - w_1 + w_2)B_1 - (1 - w_1)\max p_1\right)$$

with $\frac{dr_2}{d\max p_1} = -t(1 - w_1) \leq 0$, and

$$r_1(\Gamma) = (w_1 - w_2)B_1 + w_2 \max p_1 \quad \text{with} \quad \frac{dr_1}{d\max p_1} = w_2 \geq 0,$$

where the inequalities follow from satisfiability. Thus, we have $r_2(\Gamma) \geq r_2(\Gamma')$ and $r_1(\Gamma) \leq r_1(\Gamma')$. Plugging in our reference points we get for the interior solution in game Γ

$$p_1^+(\Gamma) = (w_1 - w_2)B_1 + \frac{1 - w_1 + c_\alpha w_2}{c_\alpha + 1}\max p_1$$

and the derivative with respect to the maximum payoff of the dictator is given

by

$$\frac{dp_1^+}{d\max p_1} = \frac{1-w_1+c_\alpha w_2}{c_\alpha+1} \geq 0$$

where the inequality follows from $\alpha \geq 0$ and satisfiability. Thus, we have $p_1^+(\Gamma) \leq p_1^+(\Gamma')$. Note furthermore, that by satisfiability $\frac{dp_1^+}{d\max p_1} \leq 1$ implying that the interior solution is feasible for any regular dictator in Γ and Γ'.

In A.2.3.1 we specified the global optimum for games like Γ with $P_1 = [0,B]$. In games like Γ' where the choice set of the dictator is restricted to $P_1' = [0,\max p_1]$ with $\max p_1 < B$ the selfish corner solution $p_1 = B$ is not feasible. Thus, we have for $c_\alpha^{1-\beta} \leq 1$ (case 1):

$$p_1^* = \begin{cases} p_1^+(\Gamma) & \text{if } \delta \geq \hat{\delta}^+(\Gamma) \\ \max p_1 & \text{else.} \end{cases}$$

with

$$\hat{\delta}^+(\Gamma) := c_\alpha^{\beta-1}\left(\left(\frac{t(\max p_1 - r_1)}{r_2 - t(B-\max p_1)}\right)^\beta - (c_\alpha+1)^{1-\beta}\left(\frac{t(B-r_1)-r_2}{r_2-t(B-\max p_1)}\right)^\beta\right)$$

where the expression for $\hat{\delta}^+(\Gamma)$ follows from rearrangement of $u_\Gamma(p_1^+(\Gamma)) \geq u_\Gamma(\max p_1)$. Note that for $c_\alpha^{1-\beta} > 1$ (case 2) the specification of the global optimum is not affected by the restriction of the choice set because the altruistic corner solution $p_1 = 0$ is feasible in Γ'.

We consider this threshold $\hat{\delta}^+(\Gamma')$ such that in game Γ' among the regular dictators with $c_\alpha^{1-\beta} \leq 1$, those with $\delta < \hat{\delta}^+(\Gamma')$ choose the selfish corner solution $p_1 = \max p_1'$ while those with $\delta \geq \hat{\delta}^+(\Gamma')$ choose the interior solution $p_1 = p_1^+(\Gamma')$. We can rewrite it as

$$\hat{\delta}^+(\Gamma') := c_\alpha^{\beta-1}\left(\left(\frac{(1-w_2)\max p_1' - (w_1-w_2)B_1}{w_1\max p_1' - (w_1-w_2)B_1}\right)^\beta - (c_\alpha+1)^{1-\beta}\left(\frac{(1-w_2)\max p_1' - (w_1-w_2)B_1}{w_1\max p_1' - (w_1-w_2)B_1} - 1\right)^\beta\right).$$

Then the derivative with respect to $\max p_1'$ is given by

$$\frac{d\hat{\delta}^+}{d\max p_1'} = \frac{\beta(1-w_1-w_2)(w_1-w_2)B_1}{c_\alpha^{1-\beta}(w_1\max p_1' - (w_1-w_2)B_1)^2}\left((c_\alpha+1)^{1-\beta}\left(\frac{(1-w_2)\max p_1' - (w_1-w_2)B_1}{w_1\max p_1' - (w_1-w_2)B_1} - 1\right)^{\beta-1} - \left(\frac{(1-w_2)\max p_1' - (w_1-w_2)B_1}{w_1\max p_1' - (w_1-w_2)B_1}\right)^{\beta-1}\right).$$

From weak altruism, weak efficiency concerns, satisfiability, and $w_1 \geq w_2$ we can conclude that $\frac{d\hat{\delta}^+}{d\max p_1'} \geq 0$. Thus, we have $\hat{\delta}^+(\Gamma) \leq \hat{\delta}^+(\Gamma')$, implying that

weakly more regular dictators with $c_\alpha^{1-\beta} \leq 1$ prefer the selfish corner solution to the interior solution in Γ' compared to Γ.

Now, consider the threshold $\delta^-(\Gamma)$ such that in game Γ among the regular dictators with $c_\alpha^{1-\beta} > 1$, those with $\delta < \delta^-(\Gamma)$ prefer the altruistic corner solution $p_1 = 0$ while those with $\delta \geq \delta^-(\Gamma)$ prefer the interior solution $p_1 = p_1^+(\Gamma)$. We can rewrite the threshold as

$$\delta^-(\Gamma) = c_\alpha^{1-\beta}\left(\frac{(1-w_1)\max p_1 + (w_1-w_2)B_1}{w_2 \max p_1 + (w_1-w_2)B_1}\right)^\beta - (c_\alpha+1)^{1-\beta}\left(\frac{(1-w_1)\max p_1 + (w_1-w_2)B_1}{w_2 \max p_1 + (w_1-w_2)B_1} - 1\right)^\beta.$$

Then the derivative with respect to $\max p_1$ is given by

$$\frac{d\delta^-}{d\max p_1} = \frac{\beta(1-w_1-w_2)(w_1-w_2)B_1}{(w_2\max p_1 + (w_1-w_2)B_1)^2}\left(c_\alpha^{1-\beta}\left(\frac{(1-w_1)\max p_1 + (w_1-w_2)B_1}{w_2 \max p_1 + (w_1-w_2)B_1}\right)^{\beta-1} - (c_\alpha+1)^{1-\beta}\left(\frac{(1-w_1)\max p_1 + (w_1-w_2)B_1}{w_2 \max p_1 + (w_1-w_2)B_1} - 1\right)^{\beta-1}\right).$$

From weak altruism, weak efficiency concerns, satisfiability, and $w_1 \geq w_2$ we can conclude that $\frac{d\delta^-}{d\max p_1} \leq 0$. Thus, we have $\delta^-(\Gamma) \geq \delta^-(\Gamma')$ implying that weakly less regular dictators with $c_\alpha^{1-\beta} > 1$ prefer the altruistic corner solution to the interior solution in Γ' compared to Γ.

Using these results together with our results from A.2.3.1 we can show that comparing the choice of any regular dictator Δ in Γ to her choice in Γ' one of the following cases applies:

(i) Her choice switches from $p_1 = p_1^+(\Gamma)$ to $p_1 = p_1^+(\Gamma')$ where $p_1^+(\Gamma) \leq p_1^+(\Gamma')$.

(ii) Her choice switches from $p_1 = p_1^+(\Gamma)$ to $p_1 = \max p_1'$ where $p_1^+(\Gamma) < \max p_1'$.

(iii) Her choice switches from $p_1 = 0$ to $p_1 = p_1^+(\Gamma')$ where $0 \leq p_1^+(\Gamma')$.

(iv) Her choice remains at $p_1 = 0$.

First, we restrict attention to regular dictators with $c_\alpha^{1-\beta} \leq 1$. Note that in game Γ by satisfiability $r_2(\Gamma) \leq B_2$ such that there is no feasible choice for the dictator in which the recipient's reference point is not fulfilled. Thus, in game Γ these dictators all choose the interior solution $p_1 = p_1^+(\Gamma)$. Now consider the same dictators in Γ' and split them into two groups according to their loss aversion parameters. The dictators with $\delta \geq \hat{\delta}^+(\Gamma')$ choose $p_1 = p_1^+(\Gamma')$ in Γ'. The dictators with $\delta < \hat{\delta}^+(\Gamma')$ choose $p_1 = \max p_1'$ in Γ'.

Now, restrict attention to regular dictators with $c_\alpha^{1-\beta} > 1$. We split these dictators into three groups according to their loss aversion parameters. Consider first the dictators with $\delta \geq \delta^-(\Gamma)$. These dictators choose $p_1 = p_1^+(\Gamma)$ in Γ. Since $\delta^-(\Gamma) \geq \delta^-(\Gamma')$ they choose $p_1 = p_1^+(\Gamma')$ in Γ'. Second, consider the dictators with $\delta \in [\delta^-(\Gamma'), \delta^-(\Gamma))$. These dictators choose $p_1 = 0$ in Γ and switch to $p_1 = p_1^+(\Gamma')$ in Γ'. Third, consider the dictators with $\delta < \delta^+(\Gamma')$. These dictators choose $p_1 = 0$ both in Γ and in Γ'.

We still have to show that for any Γ and Γ' there exist regular dictators who give in Γ and switch to taking in Γ'. We show that for any Γ and Γ' there exist regular dictators with $p_1^+(\Gamma) < B_1$ and $\delta < \hat{\delta}^+(\Gamma')$, i.e. regular dictators who give at the interior solution in Γ and to whom case (ii) applies. We have $p_1^+(\Gamma) < B_1$ iff $c_\alpha(1-w_1)+w_2 > 0$. Thus, we have $p_1^+(\Gamma) < B_1$ for all regular dictators with $0 < w_1 < 1$ or $w_1 = 1$ and $w_2 > 0$. Now, for any transfer rate t specified by Γ and Γ' we can find (α, β) satisfying $0 < \alpha \leq 1$ and $0 < \beta < 1$ such that $c_\alpha^{1-\beta} < 1 \Leftrightarrow c_\alpha^{\beta-1} > 1$. Given such (α, β), for any endowments and choice set (B_1, B_2, P_1') specified by Γ' we have $((1-w_2)\max p_1' - (w_1 - w_2)B_1)/(w_1 \max p_1' - (w_1 - w_2)B_1) = 1$ for $w_1 = 1$ and $w_2 = 0$. Thus, by continuity of $\hat{\delta}^+$ we can always find $w_1 > 0$ and $w_2 \geq 0$ in accordance with satisfiability such that the expression is close enough to 1 to make $\hat{\delta}^+(\Gamma') > 1$ and given such $(\alpha, \beta, w_1, w_2)$, we can conclude that there exist δ satisfying weak loss aversion $(\delta \geq 1)$ such that $\delta < \hat{\delta}^+(\Gamma')$.

Finally, we need to show that for any Γ and Γ' there exist regular dictators who give more in Γ than in Γ'. We show that for any Γ and Γ' there exist regular dictators with $p_1^+(\Gamma), p_1^+(\Gamma') < B_1$ and $\delta \geq \hat{\delta}^+(\Gamma')$. As above we have $p_1^+(\Gamma) < B_1$ for all regular dictators with $0 < w_1 < 1$ or $w_1 = 1$ and $w_2 > 0$. Furthermore, we have $\frac{dp_1^+}{d\max p_1} = 0$ for $w_1 = 1$ and $w_2 = 0$. Thus, by continuity of $p_1^+(\Gamma)$ for any transfer rate t specified by Γ and Γ' we can find $0 < w_1 \leq 1$ and $w_2 \geq 0$ in accordance with satisfiability such that $p_1^+(\Gamma) < p_1^+(\Gamma') < B_1$. Since there is no upper bound on the loss aversion parameter of regular dictators given such (w_1, w_2) there always exist regular dictators with $\delta \geq \hat{\delta}^+(\Gamma')$.

Step 3 **Incomplete crowding out** Reallocating initial endowment from dictator to recipient results (in expectation) in a payoff increase for the recipient.

Consider two games $\Gamma = \langle B_1, B_2, P_1, t\rangle$ and $\Gamma' = \langle B_1', B_2', P_1', t\rangle$ without taking option, i.e. $P_1 = [0, B_1]$ and $P_1' = [0, B_1']$, where Γ' is generated from

Γ by reallocating initial endowment from the dictator to the recipient, i.e. $B_1 + B_2 = B_1' + B_2' = \bar{B}$ and $B_1 < B_1'$. Thus, comparing such games we can write the recipient's endowment as a function of the dictator's endowment, i.e. $B_2(B_1) = \bar{B} - B_1$.

Moving from Γ to Γ' the game parameters that change are the player's endowments and the maximum payoff for the dictator. The dictator's endowment falls from B_1 to B_1' while the recipient's endowment rises from $\bar{B} - B_1$ to $\bar{B} - B_1'$. Furthermore, the maximum payoff for the dictator falls from B_1 to B_1' such that the minimum payoff for the recipient rises from $\min p_2 = t(\bar{B} - B_1)$ to $\min p_2' = t(\bar{B} - B_1')$. Therefore, the utility functions of a regular dictator Δ in Γ and Γ' differ in the reference points of the dictator and the recipient. We have

$$r_1(\Gamma) = w_1 B_1 \quad \text{with} \quad \frac{dr_1}{dB_1} = w_1 \geq 0,$$

where the inequality follows from satisfiability, and

$$r_2(\Gamma) = t(\bar{B} - (1 - w_2)B_1) \quad \text{with} \quad \frac{dr_2}{dB_1} = -t(1 - w_2) \leq 0,$$

where the inequality follows from satisfiability and $t > 0$. Thus, we have $r_1(\Gamma) \geq r_1(\Gamma')$ and $r_2(\Gamma) \leq r_2(\Gamma')$. We can rewrite the interior solution as

$$p_1^+(\Gamma) = \frac{1 + c_\alpha w_1 - w_2}{c_\alpha + 1} B_1.$$

Taking the derivative with respect to the dictator's initial endowment we get

$$\frac{dp_1^+}{dB_1} = \frac{1 + c_\alpha w_1 - w_2}{c_\alpha + 1} \geq 0$$

where the inequality follows from imperfect altruism, weak efficiency concerns, and satisfiability. Thus, we have $p_1^+(\Gamma) \geq p_1^+(\Gamma')$.

Consider now the threshold for $\delta^-(\Gamma)$ such that in game Γ among the regular dictators with $c_\alpha^{1-\beta} > 1$, those with $\delta < \delta^-(\Gamma)$ choose the altruistic corner solution $p_1 = 0$ while those with $\delta \geq \delta^-(\Gamma)$ choose the interior solution $p_1 = p_1^+(\Gamma)$. We can rewrite the threshold as

$$\delta^-(\Gamma) = c_\alpha^{1-\beta} \left(\frac{1 - w_2}{w_1} \right)^\beta - (c_\alpha + 1)^{1-\beta} \left(\frac{1 - w_2}{w_1} - 1 \right)^\beta$$

and since the threshold is independent of B_1 we get $\frac{d\delta^-}{dB_1} = 0$. Thus, we have $\delta^-(\Gamma) = \delta^-(\Gamma') =: \delta^-$.

Using these results together with our results from A.2.3.1 we can show that comparing the choice of any regular dictator Δ in Γ to her choice in Γ' one of the following cases applies:

(i) Her choice switches from $p_1 = p_1^+(\Gamma)$ to $p_1 = p_1^+(\Gamma')$ where $p_1^+(\Gamma) \geq p_1^+(\Gamma')$.

(ii) Her choice remains at $p_1 = 0$.

Consider first only regular dictators with $c_\alpha^{1-\beta} \leq 1$. Since in neither Γ nor Γ' there is a feasible choice such that the reference point of the recipient is not fulfilled, these dictators all choose the respective interior solution in Γ and Γ'.

Now, consider regular dictators with $c_\alpha^{1-\beta} > 1$. We split these dictators into two groups according to their loss aversion parameters. Consider first the dictators with $\delta \geq \delta^-$. These dictators choose $p_1 = p_1^+(\Gamma)$ in Γ and $p_1 = p_1^+(\Gamma')$ in Γ'. Second, consider the dictators with $\delta < \delta^-$. These dictators choose $p_1 = 0$ both in Γ and Γ'.

Finally, we show that for any Γ and Γ' there exist regular dictators to whom case (i) applies in a strict sense, i.e. regular dictators whose choice in Γ' compared to Γ strictly increases the payoff of the recipient. For any transfer rate t specified by Γ and Γ' we can find $\alpha > 0$ and β satisfying weak altruism and weak efficiency concerns such that $c_\alpha^{1-\beta} \leq 1$, i.e. for any transfer rate t we can find regular dictators to whom case (i) applies. Furthermore, given such (α, β) we can always find (w_1, w_2) in accordance with satisfiability such that $dp_1^+/dB_1 > 0$.

Step 4 Efficiency concerns The recipient's payoff is weakly increasing in the transfer rate.

Consider two games $\Gamma = \langle B_1, B_2, P_1, t\rangle$ and $\Gamma' = \langle B_1, B_2, P_1, t'\rangle$ with $t < t'$, $P_1 = [0, \max p_1]$, and $B_1 \leq \max p_1 \leq B_1 + B_2$ which are equivalent in every dimension except the transfer rate.

The utility functions of a regular dictator Δ in Γ and Γ' differ only in the reference points of the recipient. His endowment is multiplied with t' instead of t and his minimal payoff increases from $\min p_2 = t(B - \max p_1)$ to

$\min p_2' = t'(B - \max p_1)$. We have

$$r_2(\Gamma) = t(B_2 + (1 - w_1 + w_2)B_1 - (1 - w_1)\max p_1)$$

with

$$\frac{dr_2}{dt} = B_2 + (1 - w_1 + w_2)B_1 - (1 - w_1)\max p_1 \geq 0$$

where the inequality follows by satisfiability and $\max p_1 \leq B_1 + B_2$. Thus, we have $r_2(\Gamma) \leq r_2(\Gamma')$. We can rewrite the interior solution as

$$p_1^+(\Gamma) = \frac{t((1 - w_1)\max p_1 + (w_1 - w_2)B_1) + (\alpha t)^{\frac{1}{1-\beta}} r_1(\Gamma)}{(\alpha t)^{\frac{1}{1-\beta}} + t}.$$

Taking the derivative with respect to the transfer rate we get

$$\frac{dp_1^+}{dt} = \frac{tc_\alpha \beta}{1-\beta}(r_1(\Gamma) - (1 - w_1)\max p_1 - (w_1 - w_2)B_1) = \frac{tc_\alpha \beta}{1-\beta}(w_1 + w_2 - 1)\max p_1 \leq 0$$

where the inequality follows from weak altruism, weak efficiency concerns, and satisfiability. Thus, we have $p_1^+(\Gamma) \geq p_1^+(\Gamma')$.

Consider now the threshold $\hat{\delta}^+(\Gamma)$ such that in a game Γ with $\max p_1 > B_1$ among the regular dictators with $c_\alpha^{1-\beta} \leq 1$, those with $\delta < \hat{\delta}^+(\Gamma)$ choose the selfish corner solution $p_1 = \max p_1$ while those with $\delta \geq \hat{\delta}^+(\Gamma)$ choose the interior solution $p_1 = p_1^+(\Gamma)$. We can rewrite this threshold as

$$\hat{\delta}^+(\Gamma) = \frac{1}{\alpha t^\beta}\left(\left(\frac{\max p_1 - r_1}{w_1 \max p_1 - (w_1 - w_2)B_1}\right)^\beta - \left(\left(\alpha t^\beta\right)^{\frac{1}{1-\beta}} + 1\right)^{1-\beta}\left(\frac{\max p_1 - r_1}{w_1 \max p_1 - (w_1 - w_2)B_1} - 1\right)^\beta\right).$$

Taking the derivative with respect to t we get

$$\frac{d\hat{\delta}^+}{dt} = \frac{\beta}{tc_\alpha^{1-\beta}}\left((c_\alpha + 1)^{-\beta}\left(\frac{\max p_1 - r_1}{w_1 \max p_1 - (w_1 - w_2)B_1} - 1\right)^\beta - \left(\frac{\max p_1 - r_1}{w_1 \max p_1 - (w_1 - w_2)B_1}\right)^\beta\right) \leq 0,$$

where the inequality follows from weak altruism, weak efficiency concerns, and satisfiability. Thus, we have $\hat{\delta}^+(\Gamma) \geq \hat{\delta}^+(\Gamma')$, implying that weakly more regular dictators with $c_\alpha^{1-\beta} \leq 1$ choose the selfish corner solution in Γ compared to Γ'.

Consider now the threshold $\delta^-(\Gamma)$ such that in game Γ among the regular dictators with $\alpha t^\beta > 1$, those with $\delta < \delta^-(\Gamma)$ choose the altruistic corner solution $p_1 = 0$ while those with $\delta \geq \delta^-(\Gamma)$ choose the interior solution

$p_1 = p_1^+(\Gamma)$. We can rewrite this threshold as

$$\delta^-(\Gamma) = \alpha t^\beta \left(\frac{(1-w_1)\max p_1 + (w_1 - w_2)B_1}{r_1} \right)^\beta - \left(\left(\alpha t^\beta\right)^{\frac{1}{1-\beta}} + 1 \right)^{1-\beta} \left(\frac{(1-w_1)\max p_1 + (w_1 - w_2)B_1}{r_1} - 1 \right)^\beta.$$

Taking the derivative with respect to t we get

$$\frac{d\delta^-}{dt} = \frac{\beta}{t} \left(c_\alpha^{1-\beta} \left(\frac{B_1 - (1-w_2)(\max p_1 - B_1)}{r_1} \right)^\beta - \frac{c_\alpha}{(c_\alpha + 1)^\beta} \left(\frac{B_1 - (1-w_2)(\max p_1 - B_1)}{r_1} - 1 \right)^\beta \right).$$

From weak altruism, weak efficiency concerns, and satisfiability we can conclude that $\frac{d\delta^-}{dt} \geq 0$. Thus, we have $\delta^-(\Gamma) \leq \delta^-(\Gamma')$, implying that weakly more regular dictators with $\alpha t^\beta > 1$ choose the altruistic corner solution in Γ' compared to Γ.

Step 5 Reluctant sharers When an outside option is introduced, some initial givers switch to that option while the behavior of dictators who sort into the game stays unaffected.

Consider two games $\Gamma = \langle B_1, B_2, P_1, t \rangle$ and $\Gamma' = \langle B_1, B_2, P_1', t \rangle$ with $B_1 > 0$, $B_2 = 0$, $P_1 = [0, B_1]$, and $P_1' = \{[0, B_1], \tilde{p}_1\}$ where $0.5B_1 < \tilde{p}_1 \leq B_1$, i.e. game Γ' is generated from game Γ by adding an outside option to the choice set of the dictator.

Since the two games differ only in the choice set of the dictator, which is equivalent in both games except for the extra outside option in game Γ', the utility functions of a regular dictator in Γ and Γ' are equivalent where the two choice sets overlap. Furthermore, since the dictator's information is not manipulated by the choice of the outside option, her reference point stays the same for the choice of the outside option. We have $r_1(\Gamma) = r_1(\Gamma') =: r_1$ with $r_1 = w_1 B_1$. However, since the outside option leaves the recipient completely uninformed about the choice of the dictator and the rules of the game, his reference point is zero for the outside option choice. We thus have for the reference point of the recipient $r_2(\Gamma) = r_2(\Gamma') =: r_2$ with

$$r_2 = \begin{cases} t w_2 B_1 & \text{if } p_1 \in [0, B_1] \\ 0 & \text{if } p_1 = \tilde{p}_1 \end{cases}$$

The utility of a regular dictator if she chooses the outside option is then given

by

$$u(\tilde{p}_1) = \begin{cases} \frac{1}{\beta}(\tilde{p}_1 - w_1 B_1)^\beta & \text{if } \tilde{p}_1 \geq w_1 B_1 \\ -\frac{\delta}{\beta}(w_1 B_1 - \tilde{p}_1)^\beta & \text{if } \tilde{p}_1 < w_1 B_1. \end{cases}$$

Since as noted above the utility functions of a regular dictator in Γ and Γ' are equivalent for $p_1 \in [0, B_1]$ we have $p_1^+(\Gamma) = p_1^+(\Gamma') =: p_1^+$ with

$$p_1^+ = \frac{1 + c_\alpha w_1 - w_2}{c_\alpha + 1} B_1.$$

and $\delta^+(\Gamma) = \delta^+(\Gamma') =: \delta^+$ with

$$\delta^+ = c_\alpha^{\beta-1} \left(\left(\frac{(1-w_1)B_1}{w_2} \right)^\beta - (c_\alpha + 1)^{1-\beta} \left(\frac{(1-w_1)B_1}{w_2} - 1 \right)^\beta \right).$$

Note first, that no regular dictator with $w_1 > \tilde{p}_1/B_1$ chooses the outside option. By satisfiability, such a dictator can always choose $p_1 \in [0, B_1]$ such that $p_1 \geq r_1$ and $p_2(p_1) \geq r_2$. This yields utility $u(p_1) = (p_1 - r_1)^\beta/\beta + \alpha(t(B_1 - p_1) - r_2)^\beta/\beta \geq 0$, where the inequality follows from weak efficiency concerns. Choosing $p_1' = \tilde{p}_1$ instead yields $u(\tilde{p}_1) = -\delta(w_1 B_1 - \tilde{p}_1)^\beta < 0$. In the following we restrict attention to dictators with $w_1 \leq \tilde{p}_1/B_1$. We have $u(p_1^+) < u(\tilde{p}_1)$ iff

$$\tilde{p}_1 > B_1 \left((c_\alpha + 1)^{\frac{1-\beta}{\beta}} (1 - w_1 - w_2) + w_1 \right) =: \tilde{p}_1^{min}$$

We show that for any Γ and Γ' there exist regular dictators with $\delta \geq \delta^+$ and $\tilde{p}_1^{min} < \tilde{p}_1$, i.e. regular dictators who choose the interior solution in Γ and the outside option in Γ'. For any transfer rate t specified by Γ and Γ' we can find (α, β) satisfying weak altruism and weak efficiency concerns such that $c_\alpha^{1-\beta} \leq 1$. Given such (α, β), for any dictator endowment B_1 specified by Γ and Γ' and any outside option payment $\tilde{p}_1$ specified by Γ' we have $\tilde{p}_1^{min} = 0.5 B_1$ for $w_1 = w_2 = 0.5$. Thus, by continuity of $\tilde{p}_1^{min}$ we can for any Γ and Γ' find (w_1, w_2) in accordance with satisfiability such that $\tilde{p}_1^{min} < B_1$. Since there is no upper bound on the loss aversion parameter of regular dictators, given such (w_1, w_2) there always exist regular dictators with $\delta \geq \delta^+$.

Step 6 Social pressure givers Ceteris paribus, higher susceptibility to social pressure implies higher recipient payoffs at the interior solution but also

a higher propensity to choose the outside option in a sorting game.

Higher susceptibility to social pressure corresponds to a higher weight on the opponent's endowment in the reference points, i.e. a higher w_2. We have

$$\frac{\partial p_1^+}{\partial w_2} = -\frac{t}{c_\alpha + t} B_1 < 0 \quad \text{and} \quad \frac{\partial \tilde{p}_1^{min}}{\partial w_2} = -\left(c_\alpha + 1\right)^{\frac{1-\beta}{\beta}} B_1 \leq 0$$

where the inequalities follow from weak altruism and weak efficiency concerns.

□

A.2.4 Details of the econometric specification

Technical details We estimate all parameters by maximum likelihood, and in each case, the likelihood is maximized by a combination of two algorithms: first, using the robust (gradient-free) NEWUOA algorithm (Powell, 2006; Auger et al., 2009), secondly a Newton-Raphson method to ensure convergence. In addition, we cross-test globality of the maxima using a large number of informed starting values. These starting values are derived from estimates for related models on the same data set or from the same model on other data sets. Since we estimated the same model on many different data sets and related models on the same data sets, we were able to generate many informed starting vectors helpful in examining globality of maxima via cross-testing. As is well-known from numerical non-linear maximization (see e.g. McCullough and Vinod, 2003), generating informed starting values is necessary to ensure global optimality, and it proved extremely helpful also in our case. We stopped cross-testing and generating new starting values once the estimates had converged across all optimization problems simultaneously, based on which we conclude that we approximated the global maxima.

We evaluate significance of differences between models using the Schennach-Wilhelm likelihood ratio test (Schennach and Wilhelm, 2016). This test is robust to both misspecification and arbitrary nesting of models, which is required to allow for the possibility that all models are misspecified and to acknowledge that the nesting structure at least out-of-sample is not necessarily well-defined. In addition, the Schennach-Wilhelm test allows us cluster at the subject level and to thus account for the panel character of the data. We indicate significance of differences between models distinguishing the conven-

tional level of 0.05 and the higher level of 0.01, which roughly implements the Bonferroni correction given four types of dictator game experiments we examine.

As many other experiments involving choice of numbers, responses in dictator games exhibit pronounced round-number patterns. We control for those using the focal choice adjusted logit model, exactly as derived and applied in Breitmoser (2017). The basic idea is that the roundedness of the number to be entered (to choose a given option) determine its "relative focality", which is captured by a focality index $\phi : X \to \mathbb{R}$. The idea that focality is a choice-relevant attribute of options next to utility follows from Gul and Pesendorfer (2001), and given standard axioms including positivity, independence of irrelevant alternatives and narrow bracketing, this implies a generalized logit model of the form

$$\Pr(x) = \frac{\exp\{\lambda u(x) + \kappa\phi(x)\}}{\sum_{x'} \exp\{\lambda u(x') + \kappa\phi(x')\}}. \tag{26}$$

This approach effectively captures round-number effects in stochastic choice, and in turn, simply ignoring the round-number effects as pronounced as in Dictator games was shown to yield substantially biased results in Breitmoser (2017).[30] To avoid spending any degree of freedom here, we use the same focality index as Breitmoser (2017)[31] and set κ equal to 0.8. Robustness checks on both choices are reported in Appendix C.

Capturing heterogeneity One of the more robust finding in behavioral economics is that subjects differ: They have heterogeneous preferences and differing precision in maximizing their preferences, and in addition, we suspect, they also have idiosyncratic reference points. Across subjects, these behavioral primitives are likely correlated. For example, a negative exponent β

[30] For example, in the experiment of Korenok et al. (2014), subjects mostly picked multiples of five, typically from option sets ranging from 0 to 20. The most pronounced interior mass points are at choosing payoffs of 10 for both, dictator and recipient. Estimating the reference points of subjects in this experiments without controlling for round number effects yields estimates of reference point 10 each, and in this case, the reference point simply helps to capture the round-number effect. Controlling for the round-number effects, the overall model fit improves drastically and less round-number inspired reference points (deviating from 10 each) are estimated.

[31] That is, multiples of 100 have focality level $\phi_x = 4$, other multiples of 50 have level 3, other multiples of 10 have level 2, other multiples of 5 have level 1, other integers have level 0, other multiples of 0.5 have level -1 and so on. The results are invariant to positive affine transformations of ϕ, i.e. shifting the level of or scaling ϕ does not affect the results.

in the CES utility function implies a flat utility function, and thus to maintain "average precision" in maximizing utility a larger logit-parameter λ is required. Hence, β and λ generally are negatively correlated. For a related observation in the context of risk aversion, see for example Wilcox (2008). The correlation structure itself is unknown, however, and in addition, functional form assumptions about the marginal distributions of parameters seem to be equally difficult to make in the present context. We have only little knowledge about the distribution of individual preferences in generalized dictator games, except that the altruism weight α is likely truncated at say $(-0.5, 0.5)$, and that the exponent β does not seem to comply with a simple continuous distribution (for example, Andreoni and Miller, 2002, estimate that some subjects have linear preferences with β close to 1, some have Cobb-Douglas with $\beta \approx 0$, and others are Leontief with $\beta \to -\infty$).

While somewhat adequate approximations exist for each of these issues, we chose to tackle heterogeneity in a non-parametric manner attempting to combine the strengths of continuous distributions ("random coefficients") and the generality of finite-mixture models (see e.g. McLachlan and Peel, 2004). In a first step, we estimate for each subject the model parameters (preferences α, β, precision λ, and reference point weights w_1, w_2) individually by maximum likelihood.[32] Then, for the predictions that most of our results rely on, we implement a finite mixture approach where each of the n subjects available in-sample has weight $1/n$ out-of-sample. That is, we model the out-of-sample subject pool to be characterized as a finite mixture of n components, each with prior weight $1/n$, where each component corresponds with one subject from the in-sample data set. For illustration, there are 106 subjects in KMR14. The in-sample estimation yields 106 parameter vectors denoted as $(p_1, p_2, \dots)$. This means that the prediction for the other experiments is that with probability $1/106$ a subject has vector p_1, with probability $1/106$ vector p_2 applies, and so on.

The main advantage of this approach that it allows us to capture distributions of parameters and their correlations without parametric assumptions.

[32] For numerical reasons, this step is split up into two substeps. First, we estimate individual preference and precision parameters for all reference point weights satisfying $w_1 \geq w_2$ on a grid of step-size 0.1. Secondly, we determine for each individual the likelihood maximizing reference point weights, taking the "smallest" reference point weights in cases of non-uniqueness (non-uniqueness occurs mainly for subjects consistently maximizing their pecuniary payoffs).

Any single parameter estimate is somewhat noisy, obviously, but since maximum likelihood estimates are approximately normally distributed, the errors overall cancel out and we obtain a fairly general description of the joint distribution of the individual parameters. The observed reliability of our out-of-sample predictions corroborates this approach. Finally, the approach is equally applicable to all models, also to the models accounting for say warm glow and cold prickle, or envy and guilt, and in this way it allows for an equally general treatment of heterogeneity across models.

Finally, to adjust for the differences in budgets between experiments and the (potential) differences in the weights of round numbers resulting from the differences in options sets, we allow all individual precision parameters λ and the round-number weight κ to be adjusted jointly across subjects when making predictions between experiments. These two scaling parameters are estimated from the data, but this rescaling is applied equally for all models and does therefore not affect the relative ranking. The likelihood-ratio tests of predictive adequacy also follow Schennach and Wilhelm (2016) as described above.

Table A.2.1: Instructions differ in the declaration and strength of assignment of endowments

Experiment	Instructions	Classification
AM02	"[...] you are asked to make a series of choices about how to divide a set of tokens between yourself and one other subject in the room."	neutral
HJ06	"[...] you are asked to make a series of choices about how to divide points between yourself and one other subject in the other room"	neutral
CHST07	"[...] you must decide how you want to divide the joint production between yourself and your opponent. In the example above the contributions of the two players to the joint production are 800 NOK and 200 NOK, respectively."	loaded
KMR12	"The blue player has to decide how much of $Y, a fixed amount of money, to pass to the green player and how much to keep for himself/herself. [...] In addition to the money passed by the blue player, the green player will also earn $X."	loaded
KMR13	"Blue will be asked to make a series of 18 choices about how to divide a set of tokens between herself and the Green player. [...] Each choice that Blue makes is similar to the following: Green has 15 points. Divide 50 tokens: HOLD [blank] @ 1 point(s) each, and PASS [blank] @ 2 point(s) each."	neutral (dictator) loaded (recipient)
List07	"Everyone in Room A and in Room B has been allocated $5. The person in Room A (YOU) has been provisionally allocated an additional $5. Participants in Room B have not been allocated this additional $5.[...] decide what portion, if any, of this $5 to transfer to the person you are paired with in Room B. You can also transfer a negative amount: i.e., you can take up to $1 from the person in Room B."	loaded
Bard08	"Each of you has been given GBP 6. [...] You can either leave payments unchanged, increase your own, by decreasing the other person's payment, or decrease your own, increasing the other person's payment."	loaded
KMR14	"In different scenarios you will decide what portion of your endowment to transfer to another participant in the room. Each scenario specifies how much money is in your endowment, how much money is in the OTHER endowment and the range of allowable transfers. In some scenarios you can also transfer a negative amount: i.e., you can take some of the OTHER endowment."	loaded
LMW12	"You will have to decide how to distribute €10 between yourself and the person."	neutral

A.2.5 Robustness checks in the econometric analysis

The purpose of this section is to show that the results are highly robust to variations in the three econometric assumptions: functional form for reference points (Assumption 6), relative focality of the numbers that may be entered (Footnote 31), extent of round-number effects ($\kappa = 0.8$ in Eq. (26)).

Result 7 (Summary of the robustness checks)**.**

- *We examine four different specifications clarifying how reference points change across contexts (see Definitions 5–7). In line with the theoretical prediction that welfare-based altruism improves model adequacy for all reference point specifications, both descriptive and predictive adequacy (in-sample and out-of-sample) improve highly significantly for all specifications. See Table A.2.2, panel "Aggregate".*
- *We examine two alternative specifications for factoring out round-number effects, the results are very similar for all specifications as shown. See Tables A.2.3 and A.2.4 in comparison to Table A.2.2.*
- *Throughout, we allow for* non-linear inequity aversion *as third benchmark model to extend payoff-based CES altruism. This extension fits substantially worse than the standard linear one examined above and hence was not reported in the paper. See the lines "+ Inequity Aversion (nonl)" in all the tables referenced above.*

A.2.5.1 Definitions

For clarity, we first repeat the (deliberately simplistic) base model from the main text.

Definition 5 (Welfare-based altruism (base model))**.** In game $\Gamma = \langle B_1, B_2, P_1, t \rangle$, using $w_1, w_2 \in [0,1]$,

$$r_1(\Gamma) = w_1 \cdot B_1 + w_2 \cdot tB_2$$
$$r_2(\Gamma) = w_2 \cdot B_1 + w_1 \cdot tB_2.$$

Our second robustness check is a model similar to Definition 5, but other endowments are weighed by transfer rate. This implicitly yields inequity averse reference points for $w_1 = w_2$ (scaled down or up if $w_1 + w_2 \gtrless 1$). It is

equivalent to Definition 5 if $t = 1$. By comparing it to Definition 5, we can evaluate if subjects take the transfer rate into account when forming reference points. Notable special cases are CES ($w_1 = w_2 = 0$), and inequity aversion/egalitarian ($w_1 = w_2 = 0.5$), strict libertarian ref points ($w_1 = 1, w_2 = 0$). Obviously, the model allows for a continuum in-between.

Definition 6 (Welfare-based altruism 2 (robustness check I))**.** In game $\Gamma = \langle B_1, B_2, P_1, t \rangle$, using $w_1, w_2 \in [0,1]$,

$$r_1(\Gamma) = w_1 \cdot B_1 + w_2 \cdot B_2$$
$$r_2(\Gamma) = w_2 \cdot tB_1 + w_1 \cdot tB_2.$$

Our second robustness check adapts the base model in Definition 5 by allowing for the background income to equate with the minimal payoff, rather than the outside-laboratory payoff.

Definition 7 (Welfare-based altruism 3 (robustness check II))**.** In game $\Gamma = \langle B_1, B_2, P_1, t \rangle$, using $w_1, w_2 \in [0,1]$,

$$r_1(\Gamma) = \min p_1 + w_1 \cdot (B_1 - \min p_1) + w_2 \cdot (tB_2 - \min p_2)$$
$$r_2(\Gamma) = \min p_2 + w_2 \cdot (B_1 - \min p_1) + w_1 \cdot (tB_2 - \min p_2).$$

Our final robustness check is the arguably most realistic model used in the theoretical analysis, weighing by transfer rate and using the minimal payoff as background income. This model usually fits best. It contains status-quo-based reference points ($w_1 = w_2 = 0$) and strict expectations-based reference points ($w_1 + w_2 = 1$) as the most notable special cases, and by allowing for $w_1 + w_2 \in (0,1)$ all convex combinations are also included.

Definition 8 (Welfare-based altruism 4 (robustness check III))**.** In game $\Gamma = \langle B_1, B_2, P_1, t \rangle$, using $w_1, w_2 \in [0,1]$,

$$r_1(\Gamma) = \min p_1 + w_1 \cdot (B_1 - \min p_1) + w_2 \cdot (B_2 - \min p_2/t),$$
$$r_2(\Gamma) = \min p_2 + w_2 \cdot t \cdot (B_1 - \min p_1) + w_1 \cdot (t \cdot B_2 - \min p_2).$$

As non-linear model of inequity aversion, we use the following straightforward extension of CES altruism.

Definition 9 (Non-linear inequity aversion)**.** Using the notation in the main text, non-linear inequity aversion is defined as follows:

$$u(\pi) = (1 - \alpha_1 - \alpha_2 - \alpha_3) \cdot \pi_1^{\beta} + \alpha_1 \, \pi_2^{\beta} - \alpha_2 \cdot |\pi_1 - \pi_2|_+^{\beta} - \alpha_3 \cdot |\pi_2 - \pi_1|_+^{\beta}.$$

(+ Inequity Aversion (nonl))

Finally, as simplified focality weights as robustness check for the standard focality weights described above (Footnote 31, which follows Breitmoser (2017)), we use the following.

Definition 10 (Simplified focality weights)**.** All numbers that are multiples of 5 have focality weight $\phi = 1$ in Eq. (26), all other numbers have focality weight $\phi = 0$.

A.2.5.2 Results

For the results of the robustness checks described above consider the tables on the following pages.

Table A.2.2: Predictions for standard focality weights and $\kappa = 0.8$ (results from main text)

Calibrated on	Altruism is ...	Descriptive Adequacy	Predictive Adequacy	Details on predictions of ...			
				Dictator	Endowments	Taking	Sorting
Aggregate	Payoff based (CES)	5839.8	27404.1	9343.1	9631.8	5546.8	2882.4
	+ Warm Glow/Cold Prickle	5354.6^{++}	28075^{--}	9896.5^{--}	9581.6	5617.8^{-}	2979.1^{--}
	+ Inequity Aversion	5453.9^{++}	27447.9	9094^{+}	9859.8^{--}	5600.4^{--}	2893.7
	+ Inequity Aversion (nonl)	5718.2^{+}	27435	9196.1	9811.9^{--}	5546.4	2880.6
	Welfare based	5035.7^{++}	26674.4^{++}	9093.2^{++}	9385^{++}	5451^{++}	2745.2^{++}
	Welfare based (adj)	5035.7^{++}	25740.4^{++}	8883.6^{++}	9023.5^{++}	5212.2^{++}	2631.2^{++}
	Welfare based 2	5181.4^{++}	26919.5^{++}	9108.6^{++}	9529.5	5473.3^{+}	2808.2^{++}
	Welfare based 2 (adj)	5181.4^{++}	26209^{++}	8852.9^{++}	9179.9^{++}	5393.2^{++}	2793^{++}
	Welfare based 3	5048.4^{++}	27064.9^{+}	9221.4	9640.7	5494.5^{+}	2708.2^{++}
	Welfare based 3 (adj)	5048.4^{++}	25920^{++}	8559.7^{++}	9306.4^{++}	5393.2^{++}	2670.7^{++}
	Welfare based 4	4936.9^{++}	26945^{++}	9308.3	9354.1^{++}	5493.6^{+}	2789^{++}
	Welfare based 4 (adj)	4936.9^{++}	25703.9^{++}	8594.5^{++}	9167.1^{++}	5286.7^{++}	2665.6^{++}
Dictator games	Payoff based (CES)	1460.9	8950.5	1343.4	4339	2353.3	914.7
	+ Warm Glow/Cold Prickle	1507.3^{--}	8854.6	1343	4218.4^{+}	2375.2	917.9
	+ Inequity Aversion	1234.6^{++}	8794.8^{++}	1217.1^{+}	4311.7	2360.7	905.3
	+ Inequity Aversion (nonl)	1314.9^{++}	8943.8	1271.4^{++}	4391.2^{--}	2357.8	923.3
	Welfare based	1146.6^{++}	8758^{++}	1279.8^{+}	4273.8^{+}	2316.6^{+}	887.7
	Welfare based (adj)	1146.6^{++}	8603.9^{++}	1263.9^{+}	4152.5^{++}	2300.8^{++}	888.2
	Welfare based 2	1146.4^{++}	8849.2^{+}	1276.4^{+}	4355.8	2325^{+}	892^{+}
	Welfare based 2 (adj)	1146.4^{++}	8585.3^{++}	1265.7^{+}	4119.5^{++}	2309.7^{++}	892^{+}
	Welfare based 3	1055^{++}	8818.4^{+}	1272.6^{+}	4336.7	2321.5^{+}	887.5^{+}
	Welfare based 3 (adj)	1055^{++}	8673.6^{++}	1255.2^{+}	4231.6	2307.8^{+}	880.5^{+}
	Welfare based 4	1050.9^{++}	8715.2^{++}	1268.8^{+}	4240.1^{++}	2324.2	882.1^{++}
	Welfare based 4 (adj)	1050.9^{++}	8662.1^{++}	1252.5^{+}	4219.8^{++}	2309.4^{+}	881.9^{++}
Gen Endowments	Payoff based (CES)	2896.6	8752.9	4260.4	826.1	2613.8	1052.7
	+ Warm Glow/Cold Prickle	2395.5^{++}	8967.8^{--}	4289.6	954.5^{--}	2649.7	1074
	+ Inequity Aversion	2800.1^{+}	8916.4^{--}	4333.6^{-}	849.9	2663^{--}	1069.9^{--}
	+ Inequity Aversion (nonl)	2923.3	8703.6^{+}	4235.2	824.5	2599.8^{+}	1044^{++}
	Welfare based	2662.7^{++}	8416.7^{++}	4084.2^{++}	767.9^{+}	2565.9^{+}	998.7^{++}
	Welfare based (adj)	2662.7^{++}	7867.7^{++}	3985.8^{++}	637.1^{++}	2351^{++}	895.4^{++}
	Welfare based 2	2769.6^{++}	8615.1^{++}	4157.7^{+}	819.4	2580.2	1057.8
	Welfare based 2 (adj)	2769.6^{++}	8312.5^{++}	3995.9^{++}	751.5^{++}	2521.6^{++}	1045.1
	Welfare based 3	2730^{++}	8626.1^{++}	4236.6	822.1	2606.2	961.2^{++}
	Welfare based 3 (adj)	2730^{++}	7928.2^{++}	3692.7^{++}	778.5^{+}	2524.5^{++}	934^{++}
	Welfare based 4	2662.7^{++}	8754.3	4319.1^{-}	782	2601.1	1052
	Welfare based 4 (adj)	2662.7^{++}	7710.2^{++}	3719.7^{++}	643.3^{++}	2413.4^{++}	935.2^{++}
Taking Games	Payoff-based (CES)	1482.4	9700.7	3739.3	4466.7	579.7	914.9
	+ Warm Glow/Cold Prickle	1451.8	10252.5^{--}	4263.8^{--}	4408.7	592.8	987.2^{--}
	+ Inequity Aversion	1419.2^{+}	9736.7	3543.3^{++}	4698.2^{--}	576.6	918.5
	+ Inequity Aversion (nonl)	1479.9	9787.7	3689.5	4596.1^{--}	588.8^{-}	913.3
	Welfare-based	1226.4^{++}	9499.7^{+}	3729.2	4343.2	568.5^{+}	858.8^{++}
	Welfare based (adj)	1226.4^{++}	9270.3^{++}	3633^{+}	4232.9^{++}	559.3^{++}	846.6^{++}
	Welfare-based 2	1265.5^{++}	9455.3^{++}	3674.5	4354.2^{+}	568.2	858.3^{++}
	Welfare based 2 (adj)	1265.5^{++}	9310.1^{++}	3590.4^{+}	4305.4^{++}	560.9^{+}	854.9^{++}
	Welfare-based 3	1263.4^{++}	9620.4	3712.2	4482	566.9	859.4^{++}
	Welfare based 3 (adj)	1263.4^{++}	9312.1^{++}	3603.2^{+}	4295.4^{+}	559.8^{+}	855.2^{++}
	Welfare-based 4	1223.4^{++}	9475.5^{++}	3720.4	4332^{+}	568.2^{+}	855^{++}
	Welfare based 4 (adj)	1223.4^{++}	9331.8^{++}	3620.6	4302.4^{+}	562.9^{++}	847.5^{++}

Table A.2.3: Predictions for simplified focality weights and $\kappa = 0.8$ (robustness check)

Calibrated on	Altruism is ...	Descriptive Adequacy	Predictive Adequacy	Details on predictions of ...			
				Dictator	Endowments	Taking	Sorting
Aggregate	Payoff based (CES)	5968.4	27868.5	10084.1	9676.9	5277.3	2830.1
	+ Warm Glow/Cold Prickle	5546.9^{++}	28922.2^{--}	10687^{--}	9846.6^{-}	5428^{--}	2960.7^{--}
	+ Inequity Aversion	5593.9^{++}	27994.9	9944	9772.7	5377.8^{--}	2900.4^{--}
	+ Inequity Aversion (nonl)	6128.5^{--}	28905.7^{--}	10752.5^{--}	9827^{--}	5353.9^{--}	2972.4^{--}
	Welfare based	4677.3^{++}	27288.5^{++}	9790.8^{++}	9560.8	5232.4^{+}	2704.5^{++}
	Welfare based (adj)	4677.3^{++}	26308.2^{++}	9618.3^{++}	9107.8^{++}	5000.7^{++}	2591.4^{++}
	Welfare based 2	5023.7^{++}	26894.6^{++}	9920.8^{+}	8986.9^{++}	5275.3	2711.6^{++}
	Welfare based 2 (adj)	5023.7^{++}	26240.1^{++}	9659.8^{++}	8759^{++}	5148^{+}	2683.3^{++}
	Welfare based 3	5258^{++}	27031.7^{++}	9843.9^{++}	9180.6^{++}	5270.6	2736.6^{++}
	Welfare based 3 (adj)	5258^{++}	26174.3^{++}	9591.9^{++}	8875.1^{++}	5133.9^{+}	2583.4^{++}
	Welfare based 4	5258^{++}	26772.1^{++}	9733.1^{++}	9174.7^{++}	5202.5^{+}	2661.8^{++}
	Welfare based 4 (adj)	5258^{++}	25472.9^{++}	8880^{++}	8933.2^{++}	5088.7^{++}	2581^{++}
Dictator games	Payoff based (CES)	1697.2	8998.3	1462.4	4387.3	2253.5	895.1
	+ Warm Glow/Cold Prickle	1715.9	9104.6^{--}	1502.8^{--}	4377.3	2304^{--}	920.4^{-}
	+ Inequity Aversion	1390.2^{++}	8834.1^{+}	1352^{+}	4313.9	2268.5	899.7
	+ Inequity Aversion (nonl)	1753^{--}	9117.8^{--}	1484.9^{-}	4418.9	2295.4^{--}	918.6^{-}
	Welfare based	1392^{++}	8807.9^{++}	1396.7^{++}	4335.2	2208.4^{++}	867.5
	Welfare based (adj)	1392^{++}	8473.5^{++}	1349.4^{++}	4080^{++}	2184.7^{++}	861^{+}
	Welfare based 2	1400.9^{++}	8757.5^{++}	1442.9	4170.8^{++}	2258	885.9
	Welfare based 2 (adj)	1400.9^{++}	8654.1^{++}	1437	4090.9^{++}	2248.6	879.1
	Welfare based 3	1392.3^{++}	8801^{++}	1391.2^{++}	4266^{++}	2270.1	873.7
	Welfare based 3 (adj)	1392.3^{++}	8548.6^{++}	1348.2^{++}	4071.4^{++}	2263.4	867.1^{+}
	Welfare based 4	1392.7^{++}	8707.2^{++}	1360.6^{+}	4234.8^{+}	2257.3	854.4
	Welfare based 4 (adj)	1392.7^{++}	8529.2^{++}	1356.5^{+}	4093.5^{++}	2241.9	838.8^{+}
Gen Endowments	Payoff based (CES)	2870.3	8828.2	4503	840.8	2441.2	1043.2
	+ Warm Glow/Cold Prickle	2438.7^{++}	9018.4^{--}	4518.8	936.9^{--}	2500.4^{--}	1062.4
	+ Inequity Aversion	2837.6	9057.8^{--}	4650.6^{--}	841.7	2504.8^{--}	1060.6^{-}
	+ Inequity Aversion (nonl)	2926.5^{--}	9003.2^{--}	4609.8^{--}	871.4^{--}	2453.2	1068.8^{--}
	Welfare based	2149.8^{++}	8650.6^{++}	4372.5^{++}	836.2	2448.7	993.1^{+}
	Welfare based (adj)	2149.8^{++}	8159.9^{++}	4308.6^{++}	703.3^{++}	2250.8^{++}	898.8^{++}
	Welfare based 2	2387^{++}	8763.2	4561.1	767^{++}	2443.9	991.2^{+}
	Welfare based 2 (adj)	2387^{++}	8321.2^{++}	4347.6^{+}	676.6^{++}	2329.4^{+}	969.1^{++}
	Welfare based 3	2636.8^{++}	8763.8	4544	762.5^{++}	2427.8	1029.4
	Welfare based 3 (adj)	2636.8^{++}	8216^{++}	4362.4^{++}	673^{++}	2299.8^{++}	882.2^{++}
	Welfare based 4	2586.3^{++}	8494.9^{++}	4401^{++}	774.9^{++}	2366^{++}	953^{+}
	Welfare based 4 (adj)	2586.3^{++}	7618.3^{++}	3757.5^{++}	696.4^{++}	2275.2^{++}	890.7^{++}
Taking Games	Payoff-based (CES)	1400.9	10041.9	4118.7	4448.7	582.6	891.9
	+ Warm Glow/Cold Prickle	1392.3	10799.2^{--}	4665.4^{--}	4532.4^{-}	623.5^{--}	977.9^{--}
	+ Inequity Aversion	1366.1^{+}	10103	3941.3^{+}	4617^{--}	604.6^{--}	940.1^{--}
	+ Inequity Aversion (nonl)	1448.9^{--}	10784.7^{--}	4657.8^{--}	4536.7^{--}	605.3^{--}	985^{--}
	Welfare-based	1135.5^{++}	9830^{+}	4021.5	4389.4	575.2	843.9^{+}
	Welfare based (adj)	1135.5^{++}	9676.3^{++}	3959.4^{++}	4323.5^{+}	564.2^{++}	830.7^{++}
	Welfare-based 2	1235.8^{++}	9374^{++}	3916.9^{++}	4049.1^{++}	573.4	834.6^{++}
	Welfare based 2 (adj)	1235.8^{++}	9266.3^{++}	3874.2^{++}	3990.5^{++}	569^{+}	834.1^{++}
	Welfare-based 3	1228.9^{++}	9466.8^{++}	3908.7^{++}	4152.1^{++}	572.7	833.4^{++}
	Welfare based 3 (adj)	1228.9^{++}	9411.2^{++}	3880.3^{++}	4129.7^{++}	569.6^{+}	833^{++}
	Welfare-based 4	1279.1^{++}	9569.9^{++}	3971.4	4164.9^{++}	579.1	854.4
	Welfare based 4 (adj)	1279.1^{++}	9326.9^{++}	3765.1^{++}	4142.3^{++}	570.5	850.5^{+}

Table A.2.4: Predictions for standard focality weights and $\kappa = 0.6$ (robustness check)

Calibrated on	Altruism is ...	Descriptive Adequacy	Predictive Adequacy	Details on predictions of ...			
				Dictator	Endowments	Taking	Sorting
Aggregate	Payoff based (CES)	5858.5	27706.5	9385.7	9895.7	5561.1	2864.1
	+ Warm Glow/Cold Prickle	5385.4^{++}	28483.9^{--}	10056.6^{--}	9809.1	5639^{-}	2979.2^{--}
	+ Inequity Aversion	5458.5^{++}	27412.6^{+}	9048^{+}	9844.8	5636.8^{--}	2882.9
	+ Inequity Aversion (nonl)	5703.4^{++}	27412.1^{+}	9166.5^{+}	9824.5	5548.6	2872.5
	Welfare based	5030.7^{++}	26719^{++}	9156.4^{+}	9359.3^{++}	5480.2^{+}	2723.1^{++}
	Welfare based (adj)	5030.7^{++}	25647.1^{++}	8894.5^{++}	8962.8^{++}	5193.7^{++}	2606.1^{++}
	Welfare based 2	5175.3^{++}	27115^{++}	9350.9	9488.8^{++}	5487.3^{+}	2788^{++}
	Welfare based 2 (adj)	5175.3^{++}	26237.4^{++}	9054^{++}	9037.9^{++}	5402.6^{++}	2752.8^{++}
	Welfare based 3	5015.1^{++}	26985.5^{++}	9207.2	9573.8^{+}	5505.2	2699.3^{++}
	Welfare based 3 (adj)	5015.1^{++}	25725.8^{++}	8509.8^{++}	9191.6^{++}	5401.6^{++}	2632.9^{++}
	Welfare based 4	4927.3^{++}	26759.6^{++}	9189.9	9334.5^{++}	5527	2708.2^{++}
	Welfare based 4 (adj)	4927.3^{++}	25558^{++}	8503.9^{++}	9130.5^{++}	5279.5^{++}	2654.1^{++}
Dictator games	Payoff based (CES)	1493.5	9087.2	1374.2	4442.4	2370.4	900.3
	+ Warm Glow/Cold Prickle	1533	9012.6	1369.2	4355.5^{+}	2378	910
	+ Inequity Aversion	1238.4^{++}	8835.5^{++}	1204.7^{++}	4341.1^{+}	2386.6	903
	+ Inequity Aversion (nonl)	1326.8^{++}	8999.9	1245.1^{++}	4472.8	2366.4	915.6^{-}
	Welfare based	1165.2^{++}	8725.1^{++}	1278^{++}	4256.8^{++}	2317.2^{+}	873.1
	Welfare based (adj)	1165.2^{++}	8486.5^{++}	1235.9^{++}	4077.4^{++}	2302.2^{++}	872.4
	Welfare based 2	1168.9^{++}	8738.7^{++}	1285.8^{++}	4245.4^{++}	2334.2^{+}	873.3
	Welfare based 2 (adj)	1168.9^{++}	8580.8^{++}	1256^{++}	4133.4^{++}	2321.1^{+}	871.8^{+}
	Welfare based 3	1066.6^{++}	8756.4^{++}	1270.3^{+}	4286.8^{+}	2321.5^{+}	877.7^{+}
	Welfare based 3 (adj)	1066.6^{++}	8556.3^{++}	1247.2^{++}	4124.1^{++}	2310^{++}	876.4^{+}
	Welfare based 4	1066.7^{++}	8690.2^{++}	1261.5^{++}	4225.2^{++}	2332.5	870.9^{+}
	Welfare based 4 (adj)	1066.7^{++}	8580.2^{++}	1238.9^{++}	4161.1^{++}	2312.8^{+}	868.9^{+}
Gen Endowments	Payoff based (CES)	2867	8696	4197.9	829.2	2613.8	1055.2
	+ Warm Glow/Cold Prickle	2383.2^{++}	9015.5^{--}	4311.2^{--}	961.3^{--}	2668	1075.1
	+ Inequity Aversion	2791.6^{+}	8899.2^{--}	4291.5^{--}	855.6	2681.2^{--}	1070.9^{-}
	+ Inequity Aversion (nonl)	2892	8677.4	4249.7	786.3^{++}	2595.8	1045.7
	Welfare based	2631.6^{++}	8479.8^{++}	4122.8^{+}	769.8^{+}	2586.6	1000.7^{++}
	Welfare based (adj)	2631.6^{++}	7884.6^{++}	4026^{++}	640.5^{++}	2329.2^{++}	890.3^{++}
	Welfare based 2	2731.8^{++}	8809.8^{--}	4348.1^{--}	813.5	2591	1057.2
	Welfare based 2 (adj)	2731.8^{++}	8468.9^{++}	4154	748.4^{++}	2522.8^{++}	1045
	Welfare based 3	2673.4^{++}	8750.8	4315.3^{--}	848.7	2621.3	965.5^{++}
	Welfare based 3 (adj)	2673.4^{++}	7915.3^{++}	3677^{++}	788.2^{+}	2532.8^{+}	918.8^{++}
	Welfare based 4	2626.2^{++}	8624.4	4242.4	776.7^{+}	2618.6	986.7^{++}
	Welfare based 4 (adj)	2626.2^{++}	7705.4^{++}	3715.2^{++}	647^{++}	2403.3^{++}	941.4^{++}
Taking Games	Payoff-based (CES)	1498.1	9923.3	3813.6	4624.1	576.9	908.6
	+ Warm Glow/Cold Prickle	1469.1	10455.7^{--}	4376.2^{--}	4492.4^{++}	593^{-}	994.1^{--}
	+ Inequity Aversion	1428.5^{+}	9677.9^{++}	3551.8^{++}	4648.1	568.9^{+}	909
	+ Inequity Aversion (nonl)	1484.7	9734.8^{+}	3671.7^{+}	4565.4	586.4^{-}	911.2
	Welfare-based	1234^{++}	9514.1^{++}	3755.7	4332.7^{++}	576.5	849.3^{++}
	Welfare based (adj)	1234^{++}	9277.5^{++}	3631.6^{++}	4243.8^{++}	561.3^{++}	842.4^{++}
	Welfare-based 2	1274.6^{++}	9566.5^{++}	3717	4429.9^{++}	562.2	857.5^{++}
	Welfare based 2 (adj)	1274.6^{++}	9187.6^{++}	3643^{++}	4153.4^{++}	557.7^{+}	835^{++}
	Welfare-based 3	1275.1^{++}	9478.3^{++}	3621.6^{++}	4438.3^{+}	562.3	856.1^{++}
	Welfare based 3 (adj)	1275.1^{++}	9255.1^{++}	3584.5^{++}	4278.2^{++}	557.2^{+}	836.7^{++}
	Welfare-based 4	1234.4^{++}	9445.1^{++}	3686	4332.5^{++}	575.9	850.7^{++}
	Welfare based 4 (adj)	1234.4^{++}	9273.4^{++}	3548.9^{++}	4321.1^{++}	562.2^{++}	842.8^{++}

References

Abeler, J., Falk, A., Goette, L., and Huffman, D. (2011). Reference points and effort provision. *American Economic Review*, 101(2):470–92.

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

Aczél, J. (1966). *Lectures on functional equations and their applications*, volume 19. Academic press.

Allcott, H. and Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2):211–36.

Allcott, H., Gentzkow, M., and Yu, C. (2019). Trends in the diffusion of misinformation on social media. *Research & Politics*, 6(2):2053168019848554.

Almås, I., Cappelen, A. W., Sørensen, E. Ø., and Tungodden, B. (2010). Fairness and the development of inequality acceptance. *Science*, 328(5982):1176–1178.

Ambuehl, S. and Li, S. (2018). Belief updating and the demand for information. *Games and Economic Behavior*, 109:21–39.

Anderson, S. P. and McLaren, J. (2012). Media mergers and media bias with rational consumers. *Journal of the European Economic Association*, 10(4):831–859.

Andreoni, J. (1990). Impure altruism and donations to public goods: A theory of warm-glow giving. *Economic Journal*, 100(401):464–477.

Andreoni, J. (1995). Warm-glow versus cold-prickle: the effects of positive and negative framing on cooperation in experiments. *Quarterly Journal of Economics*, 110(1):1–21.

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

Andreoni, J., Gravert, C., Kuhn, M. A., Saccardo, S., and Yang, Y. (2018). Arbitrage or narrow bracketing? on using money to measure intertemporal preferences. Technical report, National Bureau of Economic Research.

Andreoni, J. and Miller, J. (2002). Giving according to garp: An experimental test of the consistency of preferences for altruism. *Econometrica*, 70(2):737–753.

Andreoni, J., Rao, J. M., and Trachtman, H. (2017). Avoiding the ask: A field experiment on altruism, empathy, and charitable giving. *Journal of Political Economy*, 125(3):625–653.

Ariely, D., Bracha, A., and Meier, S. (2009). Doing good or doing well? image motivation and monetary incentives in behaving prosocially. *American Economic Review*, 99(1):544–555.

Auger, A., Hansen, N., Perez Zerpa, J., Ros, R., and Schoenauer, M. (2009). Experimental comparisons of derivative free optimization algorithms. *Experimental Algorithms*, pages 3–15.

Barberis, N. and Huang, M. (2001). Mental accounting, loss aversion, and individual stock returns. *The Journal of Finance*, 56(4):1247–1292.

Barberis, N. and Huang, M. (2009). Preferences with frames: a new utility specification that allows for the framing of risks. *Journal of Economic Dynamics and Control*, 33(8):1555–1576.

Barberis, N., Huang, M., and Santos, T. (2001). Prospect theory and asset prices. *The quarterly journal of economics*, 116(1):1–53.

Barberis, N., Huang, M., and Thaler, R. H. (2006). Individual preferences, monetary gambles, and stock market participation: A case for narrow framing. *American economic review*, 96(4):1069–1090.

Bardsley, N. (2008). Dictator game giving: altruism or artefact? *Experimental Economics*, 11(2):122–133.

Baron, D. P. (2006). Persistent media bias. *Journal of Public Economics*, 90(1-2):1–36.

Barrera, O., Guriev, S., Henry, E., and Zhuravskaya, E. (2020). Facts, alternative facts, and fact checking in times of post-truth politics. *Journal of Public Economics*, 182:104123.

Barron, K., Huck, S., and Jehiel, P. (2019). Everyday econometricians: Selection neglect and overoptimism when learning from others. Technical report, WZB Discussion Paper.

Becker, G. S. (1974). A theory of social interactions. *Journal of Political Economy*, 82(6):1063–1093.

Bellemare, C., Kröger, S., and van Soest, A. (2008). Measuring inequity aversion in a heterogeneous population using experimental decisions and subjective probabilities. *Econometrica*, 76(4):815–839.

Bellemare, C., Sebald, A., and Strobel, M. (2011). Measuring the willingness to pay to avoid guilt: Estimation using equilibrium and stated belief models. *Journal of Applied Econometrics*, 26(3):437–453.

Benartzi, S. and Thaler, R. H. (1995). Myopic loss aversion and the equity premium puzzle. *The Quarterly Journal of Economics*, 110(1):73–92.

Blanco, M., Engelmann, D., and Normann, H. T. (2011). A within-subject analysis of other-regarding preferences. *Games and Economic Behavior*, 72(2):321–338.

Blow, L. and Crawford, I. (2018). Observable consequences of mental accounting.

Bolton, G. E. and Katok, E. (1998). An experimental test of the crowding out hypothesis: The nature of beneficent behavior. *Journal of Economic Behavior & Organization*, 37(3):315–331.

Bolton, G. E. and Ockenfels, A. (2000). Erc: A theory of equity, reciprocity, and competition. *American Economic Review*, pages 166–193.

Breitmoser, Y. (2013). Estimation of social preferences in generalized dictator games. *Economics Letters*, 121(2):192–197.

Breitmoser, Y. (2017). Discrete choice with representation effects. CRC TRR 190 Working Paper.

Broberg, T., Ellingsen, T., and Johannesson, M. (2007). Is generosity involuntary? *Economics Letters*, 94(1):32–37.

Brock, J. M., Lange, A., and Ozbay, E. Y. (2013). Dictating the risk: Experimental evidence on giving in risky environments. *American Economic Review*, 103(1):415–437.

Brown, J. R., Kling, J. R., Mullainathan, S., and Wrobel, M. V. (2008). Why don't people insure late-life consumption? a framing explanation of the under-annuitization puzzle. *American Economic Review: Papers & Proceedings*, 98(2):304–09.

Bruhin, A., Fehr, E., and Schunk, D. (2018). The many faces of human sociality: Uncovering the distribution and stability of social preferences. *Journal of the European Economic Association (forthcoming)*.

Camerer, C., Babcock, L., Loewenstein, G., and Thaler, R. (1997). Labor supply of new york city cabdrivers: One day at a time. *The Quarterly Journal of Economics*, 112(2):407–441.

Camerer, C., Cohen, J., Fehr, E., Glimcher, P., and Laibson, D. (2017). Neuroeconomics. In Kagel, J. and Roth, A., editors, *Handbook of Experimental Economics*, volume 2, pages 153–217. Princeton University Press.

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

Camerer, C. F., Ho, T.-H., and Chong, J.-K. (2004). A cognitive hierarchy model of games. *Quarterly Journal of Economics*, 119(3):861–898.

Cappelen, A. W., Halvorsen, T., Sørensen, E. Ø., and Tungodden, B. (2017). Face-saving or fair-minded: What motivates moral behavior? *Journal of the European Economic Association*, 15(3):540–557.

Cappelen, A. W., Hole, A. D., Sørensen, E. Ø., and Tungodden, B. (2007). The pluralism of fairness ideals: An experimental approach. *American Economic Review*, 97(3):818–827.

Cappelen, A. W., Moene, K. O., Sørensen, E. Ø., and Tungodden, B. (2013a). Needs versus entitlements—an international fairness experiment. *Journal of the European Economic Association*, 11(3):574–598.

Cappelen, A. W., Nielsen, U. H., Sørensen, E. Ø., Tungodden, B., and Tyran, J.-R. (2013b). Give and take in dictator games. *Economics Letters*, 118(2):280 – 283.

Cappelen, A. W., Sørensen, E. Ø., and Tungodden, B. (2010). Responsibility for what? fairness and individual responsibility. *European Economic Review*, 54(3):429–441.

Card, D., DellaVigna, S., and Malmendier, U. (2011). The role of theory in field experiments. *The Journal of Economic Perspectives*, 25(3):39–62.

Chambers, C. P. and Echenique, F. (2009). Supermodularity and preferences. *Journal of Economic Theory*, 144(3):1004–1014.

Charness, G. and Rabin, M. (2002). Understanding social preferences with simple tests. *Quarterly Journal of Economics*, pages 817–869.

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

Cherry, T. L. (2001). Mental accounting and other-regarding behavior: Evidence from the lab. *Journal of Economic Psychology*, 22(5):605–615.

Cherry, T. L., Frykblom, P., and Shogren, J. F. (2002). Hardnose the dictator. *American Economic Review*, 92(4):1218–1221.

Cherry, T. L. and Shogren, J. F. (2008). Self-interest, sympathy and the origin of endowments. *Economics Letters*, 101(1):69–72.

Choi, J. J., Laibson, D., and Madrian, B. C. (2009). Mental accounting in portfolio choice: Evidence from a flypaper effect. *American Economic Review*, 99(5):2085–95.

Cooper, D. J. and Dutcher, E. G. (2011). The dynamics of responder behavior in ultimatum games: a meta-study. *Experimental Economics*, 14(4):519–546.

Cox, J. C., Friedman, D., and Gjerstad, S. (2007). A tractable model of reciprocity and fairness. *Games and Economic Behavior*, 59(1):17–45.

Crawford, V. P. and Sobel, J. (1982). Strategic information transmission. *Econometrica: Journal of the Econometric Society*, pages 1431–1451.

Dana, J., Cain, D. M., and Dawes, R. M. (2006). What you don't know won't hurt me: Costly (but quiet) exit in dictator games. *Organizational Behavior and Human Decision Processes*, 100(2):193–201.

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

De Bruyn, A. and Bolton, G. E. (2008). Estimating the influence of fairness on bargaining behavior. *Management Science*, 54(10):1774–1791.

DellaVigna, S. (2009). Psychology and economics: Evidence from the field. *Journal of Economic Literature*, 47(2):315–372.

DellaVigna, S., List, J. A., and Malmendier, U. (2012). Testing for altruism and social pressure in charitable giving. *Quarterly Journal of Economics*, 127:1–56.

DellaVigna, S., List, J. A., Malmendier, U., and Rao, G. (2016). Estimating social preferences and gift exchange at work. NBER Working Paper No. 22043.

Dueck, D. and Frey, B. J. (2007). Non-metric affinity propagation for unsupervised image categorization. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE.

Dufwenberg, M., Heidhues, P., Kirchsteiger, G., Riedel, F., and Sobel, J. (2011). Other-regarding preferences in general equilibrium. *The Review of Economic Studies*, 78(2):613–639.

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

Easterlin, R. A. (2001). Income and happiness: Towards a unified theory. *Economic Journal*, 111(473):465–484.

Eckel, C. C., Grossman, P. J., and Johnston, R. M. (2005). An experimental test of the crowding out hypothesis. *Journal of Public Economics*, 89(8):1543–1560.

Ellis, A. and Freeman, D. J. (2020). Revealing choice bracketing. *arXiv preprint arXiv:2006.14869*.

Ellis, A. and Piccione, M. (2017). Correlation misperception in choice. *American Economic Review*, 107(4):1264–92.

Elster, J. (1989). Social norms and economic theory. *Journal of Economic Perspectives*, 3(4):99–117.

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

Engelmann, D. and Strobel, M. (2004). Inequality aversion, efficiency, and maximin preferences in simple distribution experiments. *American Economic Review*, 94(4):857–869.

Enke, B. (2020). What you see is all there is. *The Quarterly Journal of Economics*, 135(3):1363–1398.

Enke, B. and Zimmermann, F. (2018). Correlation neglect in belief formation. *The Review of Economic Studies*, 0:1–20.

Esponda, I. and Vespa, E. (2018). Endogenous sample selection: A laboratory study. *Quantitative Economics*, 9(1):183–216.

Eyster, E. and Rabin, M. (2005). Cursed equilibrium. *Econometrica*, 73(5):1623–1672.

Eyster, E. and Weizsäcker, G. (2016). Correlation neglect in portfolio choice: Lab evidence.

Falk, A., Becker, A., Dohmen, T., Enke, B., Huffman, D. B., and Sunde, U. (2018). Global evidence on economic preferences. *The Quarterly Journal of Economics*, doi: 10.1093/qje/qjy013:1–48.

Falk, A. and Fischbacher, U. (2006). A theory of reciprocity. *Games and Economic Behavior*, 54(2):293–315.

Falk, A., Goette, L., and Huffman, D. (2011). Reference points and effort provision. *The American Economic Review*, 101(2):470–492.

Fallucchi, F. and Kaufmann, M. (2021). Narrow bracketing in work choices. *arXiv preprint arXiv:2101.04529*.

Fehr, E. and Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics*, pages 817–868.

Fischbacher, U., Gächter, S., and Fehr, E. (2001). Are people conditionally cooperative? evidence from a public goods experiment. *Economics letters*, 71(3):397–404.

Fishburn, P. C. (1965). Independence in utility theory with whole product sets. *Operations Research*, 13(1):28–45.

Fishburn, P. C. (1967). Interdependence and additivity in multivariate, unidimensional expected utility theory. *International Economic Review*, 8(3):335–342.

Fishburn, P. C. (1970). *Utility theory for decision making*. Research Analysis Corporation.

Fisman, R., Kariv, S., and Markovits, D. (2007). Individual preferences for giving. *American Economic Review*, 97(5):1858–1876.

Fleurbaey, M. and Maniquet, F. (2011). *A theory of fairness and social welfare*, volume 48. Cambridge University Press.

Gabaix, X. (2014). A sparsity-based model of bounded rationality. *The Quarterly Journal of Economics*, 129(4):1661–1710.

Gächter, S., Herrmann, B., and Thöni, C. (2004). Trust, voluntary cooperation, and socio-economic background: survey and experimental evidence. *Journal of Economic Behavior and Organization*, 55(4):505–531.

Galperti, S. (2019). A theory of personal budgeting. *Theoretical Economics*, 14(1):173–210.

Gehlbach, S. and Sonin, K. (2014). Government control of the media. *Journal of public Economics*, 118:163–171.

Gentzkow, M., Shapiro, J. M., and Stone, D. F. (2015). Media bias in the marketplace: Theory. In *Handbook of media economics*, volume 1, pages 623–645. Elsevier.

Gilboa, I., Postlewaite, A., and Schmeidler, D. (2010). The complexity of the consumer problem and mental accounting.

Gill, D. and Prowse, V. (2012). A structural analysis of disappointment aversion in a real effort competition. *The American Economic Review*, 102(1):469–503.

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

Goodin, R. E. (1986). Laundering preferences. In *Foundations of social choice theory*. Cambridge University Press Cambridge.

Gorman, W. M. (1959). Separable utility and aggregation. *Econometrica*, 27(3):469–481.

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

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., and Lazer, D. (2019). Fake news on twitter during the 2016 us presidential election. *Science*, 363(6425):374–378.

Gul, F. and Pesendorfer, W. (2001). Temptation and self-control. *Econometrica*, 69(6):1403–1435.

Halevy, Y., Persitz, D., Zrill, L., et al. (2017). Parametric recoverability of preferences. *Journal of Political Economy*.

Harless, D. W., Camerer, C. F., et al. (1994). The predictive utility of generalized expected utility theories. *Econometrica*, 62(6):1251–1289.

Harrison, G. W. and Johnson, L. T. (2006). Identifying altruism in the laboratory. In *Experiments Investigating Fundraising and Charitable Contributors*, pages 177–223. Emerald Group Publishing Limited.

Hey, J. D., Lotito, G., and Maffioletti, A. (2010). The descriptive and predictive adequacy of theories of decision making under uncertainty/ambiguity. *Journal of Risk and Uncertainty*, 41(2):81–111.

Hoffman, E., McCabe, K., Shachat, K., and Smith, V. (1994). Preferences, property rights, and anonymity in bargaining games. *Games and Economic Behavior*, 7(3):346–380.

Hoffman, E., McCabe, K., and Smith, V. L. (1996). Social distance and other-regarding behavior in dictator games. *American Economic Review*, 86(3):653–660.

Hsiaw, A. (2018). Goal bracketing and self-control. *Games and Economic Behavior*, 111:100–121.

Huck, S. and Rasul, I. (2011). Matched fundraising: Evidence from a natural field experiment. *Journal of Public Economics*, 95(5):351–362.

Huck, S., Rasul, I., and Shephard, A. (2015). Comparing charitable fundraising schemes: Evidence from a natural field experiment and a structural model. *American Economic Journal: Economic Policy*, 7(2):326–69.

Jakiela, P. (2011). Social preferences and fairness norms as informal institutions: experimental evidence. *American Economic Review*, 101(3):509–513.

Jakiela, P. (2015). How fair shares compare: Experimental evidence from two cultures. *Journal of Economic Behavior & Organization*, 118:40–54.

Jehiel, P. and Koessler, F. (2008). Revisiting games of incomplete information with analogy-based expectations. *Games and Economic Behavior*, 62(2):533–557.

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

Kahneman, D., Knetsch, J. L., and Thaler, R. (1986). Fairness as a constraint on profit seeking: Entitlements in the market. *American Economic Review*, pages 728–741.

Kahneman, D., Knetsch, J. L., and Thaler, R. H. (1990). Experimental tests of the endowment effect and the coase theorem. *Journal of political Economy*, 98(6):1325–1348.

Kahneman, D., Knetsch, J. L., and Thaler, R. H. (1991). Anomalies: The endowment effect, loss aversion, and status quo bias. *Journal of Economic Perspectives*, 5(1):193–206.

Kahneman, D. and Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47:278.

Keeney, R. L. and Raiffa, H. (1993). *Decisions with multiple objectives: preferences and value trade-offs*. Cambridge university press.

Koch, A. K. and Nafziger, J. (2016). Goals and bracketing under mental accounting. *Journal of Economic Theory*, 162:305–351.

Koch, A. K. and Nafziger, J. (2020). Motivational goal bracketing: An experiment. *Journal of Economic Theory*, 185:104949.

Kogan, S., Moskowitz, T. J., and Niessner, M. (2020). Fake news: Evidence from financial markets. *Available at SSRN 3237763*.

Konow, J. (2000). Fair shares: Accountability and cognitive dissonance in allocation decisions. *American Economic Review*, 90(4):1072–1091.

Konow, J. (2003). Which is the fairest one of all? a positive analysis of justice theories. *Journal of Economic Literature*, 41(4):1188–1239.

Konow, J. (2010). Mixed feelings: Theories of and evidence on giving. *Journal of Public Economics*, 94(3):279–297.

Korenok, O., Millner, E. L., and Razzolini, L. (2012). Are dictators averse to inequality? *Journal of Economic Behavior & Organization*, 82(2):543–547.

Korenok, O., Millner, E. L., and Razzolini, L. (2013). Impure altruism in dictators' giving. *Journal of Public Economics*, 97:1–8.

Korenok, O., Millner, E. L., and Razzolini, L. (2014). Taking, giving, and impure altruism in dictator games. *Experimental Economics*, 17(3):488–500.

Kőszegi, B. and Matějka, F. (2020). Choice simplification: A theory of mental budgeting and naive diversification. *The Quarterly Journal of Economics*, 135(2):1153–1207.

Kőszegi, B. and Rabin, M. (2006a). A model of reference-dependent preferences. *The Quarterly Journal of Economics*, 121(4):1133–1165.

Kőszegi, B. and Rabin, M. (2006b). A model of reference-dependent preferences. *The Quarterly Journal of Economics*, 121(4):1133–1165.

Kőszegi, B. and Rabin, M. (2007). Reference-dependent risk attitudes. *American Economic Review*, 97(4):1047–1073.

Kőszegi, B. and Rabin, M. (2009). Reference-dependent consumption plans. *American Economic Review*, 99(3):909–936.

Krawczyk, M. and Le Lec, F. (2010). Give me a chance! an experiment in social decision under risk. *Experimental Economics*, 13(4):500–511.

Kritikos, A. and Bolle, F. (2005). Utility-based altruism: evidence from experiments. In *Psychology, Rationality and Economic Behaviour*, pages 181–194. Springer.

Krupka, E. L. and Weber, R. A. (2013). Identifying social norms using coordination games: Why does dictator game sharing vary? *Journal of the European Economic Association*, 11(3):495–524.

Kuklinski, J. H., Quirk, P. J., Jerit, J., Schwieder, D., and Rich, R. F. (2000). Misinformation and the currency of democratic citizenship. *The Journal of Politics*, 62(3):790–816.

Kumar, A. and Lim, S. S. (2008). How do decision frames influence the stock investment choices of individual investors? *Management Science*, 54(6):1052–1064.

Lazear, E. P., Malmendier, U., and Weber, R. A. (2012). Sorting in experiments with application to social preferences. *American Economic Journal: Applied Economics*, 4(1):136–163.

Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., et al. (2018). The science of fake news. *Science*, 359(6380):1094–1096.

Lian, C. (2020). A theory of narrow thinking.

List, J. (2007). On the interpretation of giving in dictator games. *Journal of Political Economy*, 115(3):482–493.

List, J. (2009). Social preferences: Some thoughts from the field. *Annual Review of Economics*, 1(1):563–583.

Luce, R. D. (1959). *Individual Choice Behavior: A Theoretical Analysis*. John Wiley and sons.

Malmendier, U., te Velde, V. L., and Weber, R. A. (2014). Rethinking reciprocity. *Annual Review of Economics*, 6(1):849–874.

Mas-Colell, A., Whinston, M. D., Green, J. R., et al. (1995). *Microeconomic theory*, volume 1. Oxford university press New York.

McCullough, B. D. and Vinod, H. D. (2003). Verifying the solution from a nonlinear solver: A case study. *American Economic Review*, 93(3):873–892.

McLachlan, G. and Peel, D. (2004). *Finite mixture models*. John Wiley & Sons.

Mitchell, A., Gottfried, J., Stocking, G., Walker, M., and Fedeli, S. (2019). Many americans say made-up news is a critical problem that needs to be fixed. *Pew Research Center*, 5:2019.

Mocanu, D., Rossi, L., Zhang, Q., Karsai, M., and Quattrociocchi, W. (2015). Collective attention in the age of (mis)information. *Computers in Human Behavior*, 51:1198–1204.

Morgan, J. and Stocken, P. C. (2003). An analysis of stock recommendations. *RAND Journal of economics*, pages 183–203.

Morris, S. (2001). Political correctness. *Journal of political Economy*, 109(2):231–265.

Mu, X., Pomatto, L., Strack, P., and Tamuz, O. (2020). Background risk and small-stakes risk aversion. *arXiv preprint arXiv:2010.08033*.

Newman, N., Fletcher, R., Schulz, A., Andi, S., Robertson, C., and Nielsen, R. K. (2021). Reuters institute digital news report 2021. *Available at SSRN: https://ssrn.com/abstract=3873260*.

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

Oxoby, R. J. and Spraggon, J. (2008). Mine and yours: Property rights in dictator games. *Journal of Economic Behavior & Organization*, 65(3):703–713.

Padoa-Schioppa, C. (2009). Range-adapting representation of economic value in the orbitofrontal cortex. *The Journal of Neuroscience*, 29(44):14004–14014.

Padoa-Schioppa, C. and Rustichini, A. (2014). Rational attention and adaptive coding: a puzzle and a solution. *American Economic Review*, 104(5):507–513.

Penczynski, S., Sitzia, S., and Zheng, J. (2020). Compound games, focal points, and the framing of collective and individual interests.

Piacquadio, P. G. (2017). A fairness justification of utilitarianism. *Econometrica*, 85(4):1261–1276.

Powell, M. (2006). The newuoa software for unconstrained optimization without derivatives. *Large-Scale Nonlinear Optimization*, pages 255–297.

Rabin, M. (1993). Incorporating fairness into game theory and economics. *The American Economic Review*, 83(5):1281–1302.

Rabin, M. and Weizsäcker, G. (2009). Narrow bracketing and dominated choices. *American Economic Review*, 99(4):1508–43.

Rawls, J. (1971). *A theory of justice*. Harvard University Press.

Read, D., Antonides, G., Van den Ouden, L., and Trienekens, H. (2001). Which is better: Simultaneous or sequential choice? *Organizational Behavior and Human Decision Processes*, 84(1):54–70.

Read, D. and Loewenstein, G. (1995). Diversification bias: Explaining the discrepancy in variety seeking between combined and separated choices. *Journal of Experimental Psychology: Applied*, 1(1):34.

Read, D., Loewenstein, G., and Kalyanaraman, S. (1999a). Mixing virtue and vice: Combining the immediacy effect and the diversification heuristic. *Journal of Behavioral Decision Making*, 12(4):257–273.

Read, D., Loewenstein, G., and Rabin, M. (1999b). Choice bracketing. *Journal of Risk and Uncertainty*, 19(1-3):171–97.

Rohde, K. I. (2010). A preference foundation for fehr and schmidt's model of inequity aversion. *Social Choice and Welfare*, 34(4):537–547.

Rubinstein, A. (2012). *Lecture notes in microeconomic theory: the economic agent*. Princeton University Press.

Ruffle, B. J. (1998). More is better, but fair is fair: Tipping in dictator and ultimatum games. *Games and Economic Behavior*, 23(2):247–265.

Saito, K. (2013). Social preferences under risk: equality of opportunity versus equality of outcome. *American Economic Review*, 103(7):3084–3101.

Samuelson, W. and Zeckhauser, R. (1988). Status quo bias in decision making. *Journal of Risk and Uncertainty*, 1(1):7–59.

Schennach, S. and Wilhelm, D. (2016). A simple parametric model selection test. *Journal of the American Statistical Association*.

Schmidt, U. (2003). Reference dependence in cumulative prospect theory. *Journal of Mathematical Psychology*, 47(2):122–131.

Silverman, C. and Singer-Vine, J. (2016). Most americans who see fake news believe it, new survey says. *BuzzFeed news*, December 6.

Simonsohn, U. and Gino, F. (2013). Daily horizons: evidence of narrow bracketing in judgment from 10 years of mba admissions interviews. *Psychological science*, 24(2):219–224.

Skiadas, C. (2013). Scale-invariant uncertainty-averse preferences and source-dependent constant relative risk aversion. *Theoretical Economics*, 8(1):59–93.

Skiadas, C. (2016). Scale or translation invariant additive preferences. Unpublished manuscript.

Smith, V. L. (1976). Experimental economics: Induced value theory. *The American Economic Review*, 66(2):274–279.

Sobel, J. (1985). A theory of credibility. *The Review of Economic Studies*, 52(4):557–573.

Stracke, R., Kerschbamer, R., and Sunde, U. (2017). Coping with complexity–experimental evidence for narrow bracketing in multi-stage contests. *European Economic Review*, 98:264–281.

Strotz, R. H. (1957). The empirical implications of a utility tree. *Econometrica*, pages 269–280.

Strotz, R. H. (1959). The utility tree–a correction and further appraisal. *Econometrica*, pages 482–488.

Thaler, R. H. (1999). Mental accounting matters. *Journal of Behavioral decision making*, 12(3):183–206.

Thaler, R. H., Tversky, A., Kahneman, D., and Schwartz, A. (1997). The effect of myopia and loss aversion on risk taking: An experimental test. *The Quarterly Journal of Economics*, 112(2):647–661.

Topkis, D. M. (1998). *Supermodularity and complementarity*. Princeton university press.

Tremblay, L. and Schultz, W. (1999). Relative reward preference in primate orbitofrontal cortex. *Nature*, 398(6729):704–708.

Tversky, A. and Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, 211(4481):453–458.

Tversky, A. and Kahneman, D. (1991). Loss aversion in riskless choice: A reference-dependent model. *Quarterly Journal of Economics*, pages 1039–1061.

van der Weele, J. J., Kulisa, J., Kosfeld, M., and Friebel, G. (2014). Resisting moral wiggle room: how robust is reciprocal behavior? *American Economic Journal: Microeconomics*, 6(3):256–264.

Vorjohann, P. (2020). Reference-dependent choice bracketing. AEA RCT Registry. September 10. https://doi.org/10.1257/rct.6419-1.0.

Wakker, P. and Tversky, A. (1993). An axiomatization of cumulative prospect theory. *Journal of Risk and Uncertainty*, 7(2):147–175.

Wakker, P. P. (1989). *Additive representations of preferences: A new foundation of decision analysis*, volume 4. Springer Science & Business Media.

Wakker, P. P. (2010). *Prospect theory: For risk and ambiguity*. Cambridge university press.

Wakker, P. P. and Zank, H. (2002). A simple preference foundation of cumulative prospect theory with power utility. *European Economic Review*, 46(7):1253–1271.

Wilcox, N. (2008). Stochastic models for binary discrete choice under risk: A critical primer and econometric comparison. In Cox, J. C. and Harrison, G. W., editors, *Risk aversion in experiments*, volume 12 of *Research in experimental economics*, pages 197–292. Emerald Group Publishing Limited.

Wilcox, N. T. (2011). Stochastically more risk averse: A contextual theory of stochastic discrete choice under risk. *Journal of Econometrics*, 162(1):89–104.

Wilcox, N. T. (2015). Error and generalization in discrete choice under risk. ESI Working Paper 15-11.

Zhao, S., López Vargas, K., Friedman, D., and Gutierrez Chavez, M. A. (2020). Ucsc leeps lab protocol for online economics experiments. Available at SSRN: https://ssrn.com/abstract=3594027 or http://dx.doi.org/10.2139/ssrn.3594027.

Erklärung zu Selbstständigkeit und Hilfsmitteln

Hiermit erkläre ich, dass ich die Dissertation selbständig und nur unter der Verwendung der angegebenen Hilfen und Hilfsmittel angefertigt habe.

Ich bezeuge durch meine Unterschrift, dass meine Angaben über die bei der Abfassung meiner Dissertation benutzten Hilfsmittel, über die mir zuteil gewordene Hilfe sowie über frühere Begutachtungen meiner Dissertation in jeder Hinsicht der Wahrheit entsprechen.

Exeter, den 6. Januar 2022.

Kumulative Dissertation: Erklärung zu Ko-Autoren, Eigenanteil und Publikationsstatus

Lfd. Nr	Titel der Einzelarbeit	Namen der Ko-Autoren	Erklärung zum Eigenanteil	Publikations-status
1	Reference-dependent choice bracketing	Keine	100%	Nicht eingereicht
2	Welfare-based altruism	Yves Breitmoser	Einordnung in die Literatur (90%), Erarbeitung des Modells (50%), Theoretische Analyse der Modellimplikationen (100%), Formulierung des Manuskripts (50%)	Eingereicht
3	Fake news and information transmission	Steffen Huck	Einordnung in die Literatur (100%), Erarbeitung des Modells (30%), Theoretische Analyse der Modellimplikationen (70%), Formulierung des Manuskripts (100%)	Nicht eingereicht