UNIVERSITAT DE BARCELONA

Essays on Belief-Based Utility

Theo Saroglou

PhD in Economics

Essays on Belief-Based Utility

Theo Saroglou

UNIVERSITAT DE
BARCELONA

PhD in Economics

Thesis title:

Essays on Belief-Based Utility

PhD candidate:

Theo Saroglou

Advisors:

Alessandro De Chiara
Alexandrina Stoyanova

Date:

November 2024

UNIVERSITAT DE BARCELONA

This work was financially supported by the Spanish Ministry of Science, Innovation and Universities, Spain [grant number PID2020-114040RB-I00]; the Catalan Research Agency [grant number 2021SGR00678]; and the European Union's Horizon Europe Research and Innovation Programme under Grant Agreement number 101095175 (SUSTAIN-WELL project). Views and opinions expressed are however those of the author only and do not necessarily reflect those of the European Union.

Acknowledgements

The original plan, the one I had as a child, was to become a truck driver. That plan failed terribly, so today I am submitting my doctoral thesis. For this failure, I have a lot of people to blame—people who encouraged me and taught me to dream big. Of course, there are also those who questioned me, but I have been advised against including a letter to them here. Unfortunately, I cannot recount 30 years of important people in my life and express my deep gratitude to all of them. I must focus on thanking those who made this PhD dissertation possible over the past few years.

First, I would like to thank Ció Patxot. Without her persistent support, I would not be writing this letter. I am grateful to my coauthor and advisor, Alessandro De Chiara, for his support at every step of this work and for helping me build a dissertation that reflects my interests. I thank my advisor, Alexandrina Stoyanova—the most prolific feedback provider in Barcelona—and thank God that she doesn't get to give feedback on this letter too. I also want to thank my coauthor, Ester Manna, for her valuable input and calming presence during our lab experiments, and my coauthor, Claudia Cerrone, for always being able to find the cracks in the experimental design. I am grateful to Joël van der Weele for his valuable input, his hospitality during my stay at the University of Amsterdam, and for introducing me to this research field many years ago. Finally, I thank Jordi Roca, whose ability to quickly assist in any scenario makes me think he might be more than one person. Without his help, I would not be able to find my way to the bathroom, let alone a PhD degree.

The final year of my PhD coincided with a particularly difficult time in my life, but it also filled me with gratitude for the people in it. I am deeply thankful to my family: my mother, Sofia; my brother, Konstantinos; and above all, my sister, Anna, who endured me more than anyone else. I could not have done this without her generous support. I also want to extend my gratitude to Elena, who always believes in me more than I do; to Pamela, who knows how to be the right person at the right time; to Stefanos, who I will never get tired of arguing with; and to Ece, Jean, Kevin, Luke, Mattia, Ole, and Rachel, who took care of me during a time of need. Finally, I want to express my appreciation and love to my oldest and dearest friends: Olga, who belongs to her own league of support; and Giorgos, Marianthi, Panos, and Tryfonas, the kind of friends one always needs but rarely deserves.

In my signature style, I am probably forgetting someone. However, I am keenly aware of how little we accomplish on our own. Therefore, I thank everyone at the University of Barcelona and all the friends and family around the world. Even if your name has not been mentioned, know that you played a key role in this journey.

Abstract

This dissertation investigates belief-based utility, the notion that beliefs provide direct utility beyond their informational role, through three experimental studies. Chapter 1 introduces belief-based utility, explaining its implications and the cognitive strategies it involves. It lays the groundwork for understanding how beliefs interact with decision-making and social dynamics. Chapter 2 explores the non-monotonic relationship between self-deception and altruistic self-image, revealing distinct patterns of bias depending on the participants' motivations. This study uses a novel experimental design to categorize participants and quantifies altruistic self-image using a proxy, uncovering opposing biases that arise under similar belief structures. Chapter 3 examines social image concerns in task choices within an organizational context, demonstrating how social image motives can lead to inefficient decision-making and suboptimal task selection. Chapter 4 delves into expectations, merit beliefs, and redistribution preferences. It shows that individuals' expectations of future economic performance shape their redistribution preferences. Together, these studies provide novel insights into the cognitive mechanisms underlying belief formation and their implications for individual and societal welfare. This research contributes to the literature on behavioral and experimental economics by uncovering the complex interplay between beliefs, utility, and decision-making.

Keywords

Belief-Based Utility, Self-Deception, Altruism, Social Image, Redistribution Preferences, Merit Beliefs; Motivated Reasoning, Expectations, Self-Image, Social Preferences, Inequality, Overconfidence, Prosocial Behavior

Table of Contents

List of Tables

List of Figures

1 An Introduction to Belief-Based Utility

In this chapter, I introduce the concept of belief-based utility, summarize the various ways in which beliefs can enter the utility function, and clarify the distinctions between motivated cognition and heuristics. Sections 1.4, 1.5, and 1.6 explore the implications of belief-based utility and the cognitive strategies associated with it. The aim of this chapter is to establish a foundational understanding of belief-based utility, which is essential for following and critically evaluating the three experimental studies presented in this dissertation.

1.1 Background: Beliefs & Economics

Traditionally, in Economics, beliefs are treated as a form of strategic information, entering the utility function as probabilities. These probabilities reflect the rational expectations an individual holds regarding the present or future state of the world. In this framework, believing that most people are just is functionally equivalent to knowing that most people will act justly in a strategic interaction. This interpretation, which treats beliefs as information, implies that more accurate beliefs are inherently preferable. However, could it not be that believing one lives in a society of just individuals provides intrinsic satisfaction? Such a belief could bring about positive feelings, such as security, fraternity, and trust. When presented with information that challenges the belief that people are generally just, the informational perspective suggests that an individual should feel less trust and security. Why, then, would one relinquish these positive feelings without resistance? An individual may prefer to bear the costs associated with holding an inaccurate belief rather than give up these positive sentiments. Belief-based utility proposes that beliefs are a direct source of utility and that this utility need not depend on their accuracy (Molnar and Loewenstein, 2022).

Attempts to incorporate subjectivity into beliefs, such as Subjective Expected Utility (Savage, 1972) and extensive research in the economics of information (Ackerloff, 1970; Spence, 1978; Stiglitz and Weiss, 1981), treat beliefs exclusively as aids in decision-making. This perspective implies that individuals strive to maintain the most accurate beliefs and, if biases are present, they should not systematically skew in one direction (Molnar and Loewenstein, 2022).

Nevertheless, individuals derive value from their beliefs in a variety of ways. People gain utility directly from their self-image across various qualities (Steele, 1988), from their expectations about future outcomes (Loewenstein, 1987), and even from the internal consistency of their beliefs (Eyster, 2002). Experimental findings indicate that individuals seek information not only when it aids decision-making but also when they expect it to bolster self-esteem (Golman et al., 2022). These examples illustrate ways in which individuals value their beliefs independently of other sources of utility. Additionally, we are influenced by the beliefs of others; for instance, individuals prefer holding views aligned with those of people around them (Golman et al., 2016) and experience discomfort if they suspect others' beliefs are inaccurate (Molnar and Loewenstein, 2020).

Early economic thought treated beliefs as a direct source of utility. As far back as Smith (1759), economists argued that beliefs were not merely probabilities that assist in utility maximization (a broader historical context is provided in Lowenstein, 1992). As behavioral economics progressed, the focus shifted towards heuristics and cognitive biases, departing from the strict framework of rational expectations. Nonetheless, belief-based utility has re-emerged as a key research area, with Schelling (1988) being among the first to conceptualize beliefs as akin to goods or assets that individuals might consume or invest in. His essay, *The Mind as a Consuming Organ*, encapsulates this view, suggesting that people consume beliefs and information even when such consumption lacks strategic benefit.

1.2 How we care about beliefs

1.2.1 Functional and Affective Value of Beliefs

Beliefs can be categorized into two general types: affective and functional (Bénabou and Tirole, 2002; Bénabou and Tirole, 2016).

Following the framework of Schelling (1988), affective beliefs imply direct consumption, encompassing self-esteem, social-image concerns, hope, and anxiety. Affective beliefs also enter the utility function as anticipatory emotions—feelings directed toward future outcomes based on the current state, such as dread, fear, and hope.

Functional beliefs, by contrast, support goals like motivation and commitment (Bénabou and Tirole, 2002, 2004; Carrillo and Mariotti, 2000). For instance, if my goal is to increase gym attendance and improve health, framing my choices around the identity of being a ”gym person” helps reinforce my commitment. By questioning, “Would a gym person make this choice?” I reinforce alignment with my goal. Another functional use of beliefs occurs in the facilitation of deception; individuals can more convincingly persuade others if they believe their own misrepresentations (Hippel and Trivers, 2011; Schwardmann and Van der Weele, 2019).

Although affective and functional beliefs serve distinct purposes, they can be modeled similarly when examining behavior. Bénabou and Tirole (2011) highlights that the greatest inefficiencies in decision-making often arise when self-esteem has consumption value, as efforts to maintain such beliefs can lead to a ”hedonic treadmill” effect. In this state, individuals prioritize belief protection over welfare improvement, which may ultimately foster sacred values or even taboos at both personal and societal levels (Bénabou, 2013). For example, Callen et al. (2016) found that one-quarter of Pakistani participants would forgo a large payment rather than anonymously thank the U.S. government. At the group level, ostracism often serves as the most potent tool available for defending collective beliefs.

1.2.2 Intrapersonal & Interpersonal Beliefs

In the traditional economic view, new information should prompt individuals to adjust their beliefs. However, people tend to prefer consistency in their beliefs. Beyond the social stigma associated with changing beliefs (Allgeier et al., 1979), individuals often go to considerable lengths to justify past mistakes (Eyster, 2002) and derive utility directly from maintaining consistency (Falk and Zimmermann, 2011). Justifying past mistakes frequently requires holding biased beliefs and making suboptimal decisions in the present. A further distinction between the traditional perspective on beliefs in economics and belief-based utility is in the importance placed on the beliefs of others. In conventional economics, the beliefs of others are relevant mainly insofar as they affect strategic interactions. For example, in the prisoner's dilemma, a conventional analysis would focus on whether the other prisoner might confess. However, belief-based utility suggests that we also care about whether the other prisoner perceives us as someone who would betray a partner by confessing. Additionally, we value consistency between our beliefs and those of people around us (Golman et al., 2016). However, we tend to experience a stronger aversion to others' beliefs when we are convinced they are wrong, rather than simply different from ours (Molnar and Loewenstein, 2020).

1.3 Motivated cognition vs. heuristics

It is essential to distinguish heuristics and bounded rationality from motivated beliefs. Bénabou and Tirole (2016) identify three primary indicators of motivated cognition. First, motivated cognition exhibits endogenous directionality: when interpreting signals in a biased way, the bias should not be direction-neutral. For instance, if someone misinterprets signals regarding their intelligence, these errors should be consistently biased in one direction (typically upward), rather than being evenly distributed on both sides of the signal. Second, unlike bounded rationality, motivated cognition is not a byproduct of limited sophistication. While cognitive biases such as the endowment effect or loss aversion tend to diminish among more sophisticated individuals (Frederick, 2005), ideologically motivated biases are often stronger in those with higher intelligence (Kahan, 2013; Kahan et al., 2017). Third, motivated thinkers are driven by emotional factors rather than constraints in their perceptual abilities.

1.4 Motivated Cognition & Side-Effects

Psychology has significantly contributed to the study of motivated beliefs, though the topic has been contentious within the discipline. In a seminal study, Kunda (1987) examined self-serving biases and demonstrated through a series of experiments that individuals favor beliefs implying positive outcomes. One mechanism supporting such beliefs is the differential threshold for information confirmation—self-serving beliefs require less information to be validated (Ditto and Lopez, 1992). For a defense of motivated beliefs and a summary of the psychological debate at the time, see Kunda, 1990. Although the exact mechanisms of self-deception remain debated, the psychological benefits of self-deception and its value in social deception and status management are widely recognized (Chance and Norton, 2015). Incorporating beliefs as direct sources of utility within the utility function can help explain various psychological phenomena. Anticipatory emotions, such as fear or anxiety, contribute to time inconsistency and observed overreactions to small probabilities (Caplin and Leahy, 2001). Common violations of the independence axiom in intertemporal choice (Loewenstein, 1987) can be reconciled by accounting for anticipatory emotions within the utility framework. For example, individuals may savor or dread a future outcome, leading them to delay positive experiences or hasten negative ones. These anticipatory emotions can be intensely felt; in a study by Berns et al. (2006), some participants opted for a stronger electric shock immediately rather than wait for a milder one later. Beliefs need not be self-image-relevant to impact decision-making. When individuals invest in an outcome, they develop stakes-dependent beliefs (Kunda, 1987). Stakes-dependent beliefs resemble the sunk-cost fallacy, where personal investment in a positive outcome inclines individuals to uphold the expectation of a favorable result. Laboratory and field studies have shown that simply "hitching one's horse" to any outcome can prompt strategies to preserve the belief that the chosen path is worthwhile. Importantly, stakes-dependent beliefs persist even when individuals are involuntarily tied to an outcome; the bias remains whether or not they choose the outcome themselves (Kunda, 1987; Babcock et al., 1995; Di Tella et al., 2007; Mijović-Prelec and Prelec, 2010; Mayraz, 2011). An illustrative example is found in Di Tella et al. (2007), where Argentinian squatters who received homes through a lottery held more favorable beliefs about the housing market than those who continued to occupy spaces illegally.

1.4.1 Implications at the individual level

Overconfidence merits special attention due to its prevalence and complex intersections with motivated beliefs. Although extensively studied as a side effect of motivated beliefs, overconfidence is not always problematic. Research indicates that moderate overconfidence can yield benefits, including higher remarriage rates and improved psychological health (Puri and Robinson, 2007; Korn et al., 2014). Notably, unlike many other biases arising from motivated reasoning, overconfidence exhibits gender differences: men tend to overestimate their performance more than women (Thaler, 2021). Several side effects of motivated beliefs discussed in this section also depend on a degree of overconfidence. While motivated beliefs can indeed lead to overconfidence, overconfidence may exist independently of motivated beliefs or self-image concerns. Burks et al. (2013) challenge the notion that overconfidence stems from self-image concerns, instead suggesting that social image concerns drive this phenomenon. Their findings reveal that overconfident individuals persistently sought information about their intelligence test scores, even when the probability of receiving a negative signal was high. However, the assumption that self-image concerns prevent individuals from seeking information with a high chance of negative outcomes has been contested. For instance, in Eil and Rao (2011a), participants were more likely to seek a signal if they had previously received a positive one, even in the absence of social-image considerations. These findings suggest that an inflated self-image is resilient; individuals seek information not despite unfavorable odds, but because they believe their chances of receiving positive feedback remain high. Overconfidence naturally arises within the context of motivated reasoning. Motivated reasoning creates self-reinforcing cycles, enabling individuals to inhabit a reality that validates their beliefs more frequently than objective circumstances would. Motivated beliefs appear in various health-related settings. For terminally

ill patients, accepting mortality can be cognitively more challenging than fostering self-deceptive optimism about survival prospects, leading many to choose self-deception (Echarte et al., 2016). In such cases, excessive hope may result in overtreatment and, ultimately, regret (Finkelstein et al., 2021). Motivated beliefs also contribute to reduced participation in preventive care (Schwardmann, 2019). Similarly, parents of children on the autism spectrum often recognize symptoms later than more objective observers around them (Sicherman et al., 2021). These examples illustrate how motivated beliefs can directly impact health outcomes. Indirect damage may also occur through conspiracy theories, ideological resistance to preventive health measures, and related factors. Studying self-deception helps us evaluate its costs and potentially identify strategies to mitigate or eliminate them. Gneezy et al. (2020) demonstrate that simply altering the timing of incentive-related information can reduce image-preserving biases among investment advisors. This research provides insight into conflicts of interest—a complex issue in which one of the most common remedies, conflict disclosure, may paradoxically worsen the situation (Cain et al., 2011) and is better understood through the lens of motivated beliefs than traditional economic approaches (Moore et al., 2010). In one of the earlier studies on motivated beliefs, Thompson and Loewenstein (1992) show that biased patterns of information recall can slow down negotiation processes.

1.4.2 Implications at the Group Level

When motivated inaccurate beliefs intersect with social motives, they frequently lead to collective delusions (Bénabou, 2013). These delusions often originate at the individual level but grow and solidify within groups. As with self-enhancement driven by motivational beliefs, group delusions can sometimes have benefits. For example, a charity might perform better if its workers believe their work is meaningful, even if the outcomes or compensation do not directly support this view. However, Bénabou (2013) provides an extensive model showing how group delusions can become detrimental, creating conditions of ”blind persistence” that lead to negative outcomes. Group delusions tend to arise when individuals are mutually dependent, and when potential adverse outcomes are both rare and severe. Bénabou (2013) refers to this condition as ”Mutually Assured Delusion” (MAD). Hierarchical organizations are particularly vulnerable to this effect, as misbeliefs held by those in leadership often propagate downwards. The MAD effect also applies broadly, affecting various social contexts, from politics to religion. The alignment of individual beliefs with those around us contributes to ideological bubbles, creating feedback loops through which significant group-level biases emerge (Bishop and Cushing, 2009). For instance, during the COVID-19 pandemic, areas in the U.S. with a higher proportion of Republicans practiced social distancing less frequently (Allcott et al., 2020). Beyond a preference for aligning beliefs with one's community, this phenomenon may be partly explained by experimental evidence showing that individuals amplify their own biases by giving more weight to interactions that support their beliefs and discounting those that contradict them (Oprea and Yuksel, 2022). Individuals who willingly consume biased information are also less inclined to seek objective news (Chopra et al., 2019). Non-Bayesian, belief-motivated reasoning has been identified in assessments of news legitimacy, contributing to the spread of fake news (Thaler, 2024). These mechanisms foster ideologically extreme and overconfident individuals. Notably, overconfident individuals not only hold more extreme views but also vote at higher rates, significantly impacting collective decisions (Ortoleva and Snowberg, 2015). Importantly, Ortoleva and Snowberg (2015) find that ideological overconfidence is not mitigated by education but is aggravated by factors such as media exposure and age. Reference groups play a significant role in shaping biased perceptions about one's position in the income distribution (Cruces et al., 2013). Such biases may underlie phenomena like poorer populations supporting tax cuts. On a positive note, the study shows that individuals informed of their true income position tend to adjust their beliefs, with previously overconfident individuals increasing their support for redistribution. Similarly, Verma (2017) finds that among individuals with equal financial literacy, those overconfident in their financial skills are more likely to make harmful financial decisions. At an aggregate level, self-enhancement better predicts societal inequality than ideological beliefs in individualism. Countries where inflated self-image across personal traits is prevalent tend to have higher inequality levels (Loughnan et al., 2011). Such collective beliefs are influential in shaping policies; for example, national beliefs about the roles of luck and merit in economic outcomes, which are not necessarily grounded in reality, strongly correlate with levels of social spending (Alesina et al., 2001). Many of the side effects discussed here apply to organizational settings. Individual biases can accumulate to form mutually assured delusions, where small initial misbeliefs necessitate increasingly larger ones over time. Schrand and Zechman (2012) discuss how initial, non-fraudulent overconfident financial misstatements may require later support, potentially leading to intentional misstatements. In certain cases, motivated beliefs may better explain economic bubbles than traditional moral-hazard interpretations. For instance, Cheng et al. (2014) examine the subprime mortgage crisis and find that mid-level managers in securitized finance were often unaware of the risks in the housing market. Remarkably, these managers were sometimes more likely to buy a house at the peak of the bubble and sell later than the general population.

1.5 Strategies of Motivated Cognition

The strategies of self-deception fall into three general categories: strategic ignorance, reality denial, and self-signaling (Bénabou and Tirole, 2016). Strategic ignorance enables individuals to entirely avoid information, allowing them to remain in deliberate ignorance and avoid updating their beliefs. Consequently, useful information that could enhance decision-making is overlooked (Schwardmann, 2019). In reality denial, individuals cannot avoid information but instead asymmetrically update their beliefs, depending on the signal's direction (Eil and Rao, 2011a). For example, positive signals are more likely to be internalized than negative ones. Self-signaling involves interpreting decisions as signals about future outcomes that hold diagnostic utility. This strategy enhances the utility of decisions that promise positive future outcomes and diminishes it for negative ones, as individuals manipulate their self-perception to ensure a favorable belief update (Bodner and Prelec, 2003; Bénabou and Tirole, 2004, 2011; Mijović-Prelec and Prelec, 2010). Notably, the source of belief-based concerns does not necessarily align with a single type of self-deception strategy; for instance, a review of information avoidance behaviors reveals that a range of belief-based motivations can drive this behavior (Golman et al., 2017).

1.5.1 Strategic Ignorance

As with other forms of motivated beliefs, strategic ignorance can occasionally provide benefits, such as serving as a commitment tool (Carrillo and Mariotti, 2000). However, in many cases, its costs outweigh these advantages. Brunnermeier and Parker (2005) incorporate utility-driven biases into an optimal expectations model, revealing significant equilibrium differences compared to rational expectations. Notably, their model treats beliefs as choice objects. Using this model, Oster et al. (2013) examine testing behavior among individuals with a possible genetic predisposition to Huntington's disease. Findings indicate that individuals who receive strong signals encouraging them to test often avoid confirming a predisposition and continue their lives unchanged, while those who do confirm it tend to adjust their behaviors. Furthermore, untested individuals generally maintain overly optimistic beliefs regarding their likelihood of developing Huntington's disease. Schwardmann and Van der Weele (2019) also addresses how limited resources lead individuals to adopt optimistic beliefs about health risks, thereby reducing preventive care participation. Following Brunnermeier and Parker (2005) and Oster et al. (2013), this research suggests that enhancing prevention through

subsidies targeting those at the bottom could be beneficial in addressing motivated belief formation. Beyond post-feedback manipulations, Castagnetti and Schmacker (2022) find that individuals actively choose less informative signals. When offered feedback options, participants preferred structures where negative signals were noisy—a preference not observed in the absence of self-image concerns.

1.5.2 Reality Denial

Anticipatory anxiety about events, such as workplace accidents, may reduce safety measures, as utility decreases when one considers the possibility of an accident (Akerlof and Dickens, 1982). To avoid this utility reduction, individuals adopt an overly optimistic view of accident likelihood. Akerlof and Dickens (1982) show that such behavior aligns with rationality and perfect information. Individuals tend to update beliefs more accurately with belief-enhancing information and less accurately with negative information, a pattern associated with distinct areas of the prefrontal cortex (Sharot et al., 2011). This asymmetric updating often leads to confirmation bias. Eil and Rao (2011a) experimentally test belief updates regarding beauty and intelligence, revealing that individuals are less responsive to negative signals and that confirmation bias is not driven by affirming signals but by positive ones. Interestingly, in the experiments of Eil and Rao (2011a), Bayesian updating was closer to reality when faced with positive signals. These findings align with Benoit and Anderson (2012), who identify neurological mechanisms for voluntarily forgetting bad news. Various paths to self-deception include biased information seeking, biased signal weighting, and perceptual biases. Motivated memory, for example, contributes to self-deception. Established biases like sunk costs, overconfidence, and the endowment effect can be better understood by incorporating motivated memory into decision models (Gottlieb, 2014). Even when properly received, negative signals may temporarily affect beliefs but fade over time, whereas positive signals often have lasting impacts (Zimmermann, 2020). Experimental evidence of selective memory is provided by Saucet and Villeval (2019), who find that individuals remember altruistic choices better but do not consistently make self-serving decisions. More importantly, when participants were not responsible for their choices, selective memory disappeared. Similar findings are reported by Carlson et al. (2020), who note that individuals breaking personal norms of generosity are more likely to forget their past actions. In real-world settings, managers may become trapped in feedback loops of overconfident performance predictions and biased performance recall (Huffman et al., 2022), a similar effect to that found in investment decisions (Gödker et al., 2021). Although field research on motivated memory remains limited,

recent studies address its effects in areas such as fertility desires (Müller, 2022), school performance (Roy-Chowdhury, 2022), and gym attendance (Sial et al., 2023). These feedback cycles align with models like Köszegi (2006), describing how the pursuit of self-enhancing signals leads to overconfidence and ultimately to task choices beyond one's competence.

1.5.3 Self-Signaling

The self-signaling process is particularly useful for self-control efforts; however, an awareness of self-signaling can lead to excessive self-control, as each signal may be discounted.

A foundational example of self-signaling is demonstrated in the well-known experiment by Quattrone and Tversky (1984), a seminal study on self-deception. In this experiment, participants adjusted their perceived pain resistance based on whether their resistance implied a positive or negative health outcome in the future. Unknowingly, they altered the signal (pain resistance) to provide themselves with evidence supporting a favorable outcome (health). Fernbach et al. (2014) conducted two similar experiments, which provided further evidence of effort denial as a form of self-signaling. Additionally, Kajackaite (2015) offers valuable insights on how effort and information choice interact in self-signaling. In their experiment, higher effort levels were linked not only to increased personal payouts but also to increased payouts for a disliked charity. When participants had the option, they chose to remain ignorant about the extent to which their effort would benefit the disliked charity and consequently increased their effort. However, when ignorance was imposed rather than chosen, this effect on effort disappeared. These findings suggest that participants willingly selected ignorance in the first scenario to maintain a self-serving belief, thereby justifying their increased effort.

1.6 Self-Deception

Since the early years of self-deception research, scholars have debated its precise definition (Mele, 1997). In much of the experimental and behavioral economics literature, self-deception is often used as an umbrella term, sometimes interchangeably with various forms of motivated cognition. While this chapter does not seek to resolve these definitional debates, it adopts the relatively strict definition provided by Gur and Sackeim (1979). According to Gur and Sackeim (1979), self-deception meets four conditions: the individual holds two contradictory beliefs (a) simultaneously (b), one of which remains outside of conscious awareness (c), while the consciously held belief is motivated (d). The experiments presented in Chapters 2 and 4 are designed to meet these four conditions.

Many perceptual biases may not be belief-driven. For example, a driver's blind spot is always present, yet it can be mitigated by a simple turn of the shoulder. In contrast, when beliefs come into play, the same stimulus may be interpreted differently. Motivational influences significantly impact sensory perceptions, such as vision. Experimental evidence shows that individuals are inclined to perceive stimuli in alignment with their desires (Balcetis and Dunning, 2006) and to perceive desired objects as physically closer (Balcetis and Dunning, 2010). In a series of studies, Balcetis and Dunning (2006) demonstrate not only that stimuli are interpreted in a motivated manner but also that this process remains unconscious to the individual. In one study, even when a previously formed motivated belief was rendered disadvantageous, it persisted. This finding is significant as it implies that although the motivated process remains hidden from conscious awareness, the resulting belief is resistant to further updating.

1.6.1 Ambiguity

The four conditions outlined by Gur and Sackeim (1979) are not the only elements necessary for successful self-deception; ambiguity also plays a key role. All self-deception experiments involve some degree of ambiguity, which is essential for self-deception at both the initial belief formation stage (Balcetis and Dunning, 2006) and at later diagnostic stages (Sloman et al., 2010). Recent neuroscience research also underscores the importance of ambiguity in facilitating self-deception (Mei et al., 2023). Ambiguity is not only crucial in self-deception but also in unethical behavior more broadly. For instance, Pittarello et al. (2015) asked participants to report the value of a die shown on a screen, with their payout depending on the reported value. When higher die values were spatially closer to the correct response, participants were more likely to report higher values dishonestly. This effect disappeared when participants' payouts were tied to response accuracy.

1.6.2 Origins and Benefits

Among others, Mei et al. (2023) attempt to map the neurological underpinnings of self-deception, revealing significant differences between individuals who engage in self-deception and those who do not. These neurological differences are not unexpected, as researchers have long theorized about the characteristics and motivations distinguishing these individuals. For instance, Lynch and Trivers (2012) find that people scoring high on self-deception questionnaires tend to laugh less, suggesting a consistently lower public expression of personal preferences. Self-deception may have conferred evolutionary advantages by enabling humans to deceive others more effectively (Hippel and Trivers, 2011; Smith et al., 2017). This theory posits that by believing one's own lies, individuals reduce the cognitive load required to sustain deception. Supporting this theory, experimental evidence from behavioral economics shows that it is easier to mislead others about one's performance when individuals have inflated their self-perception through self-deception. For example, Schwardmann and Van der Weele (2019) demonstrate that individuals can more effectively deceive others about their performance when their perception of it has been positively biased through self-deception.

1.7 Conclusion

This chapter explores belief-based utility, the concept that beliefs provide direct utility to individuals, not solely as informational aids in decision-making. Beliefs may offer affective value by fulfilling emotional needs like security and self-esteem, as well as functional value by supporting goals like motivation and social signaling. Motivated cognition occurs when individuals process information with specific, often emotionally driven biases, leading to behaviors such as strategic ignorance, reality denial, and self-signaling. The interplay between individual and social beliefs illustrates how belief-based utility informs both personal well-being and broader societal dynamics. This framework provides insight into how individuals may act in ways that diverge from traditional economic rationality when motivated by self-image or social considerations. The following chapters build on these foundations through experimental research. Chapter 2 focuses on self-deception and altruism, and investigates the directionality of motivated cognition. It finds that self-deception does not always align monotonically with self-image concerns, as biases can shift direction depending on the relative intensity of conflicting motivations. Chapter 3 studies the influence of social image concerns on task choice in an organizational context, highlighting the potential inefficiencies due to sub-optimal task uptake. Chapter 4 explores how initial economic expectations shape preferences for redistribution, demonstrating that individuals who exceed their expectations often attribute their success to merit and oppose redistribution, while those who fall short tend to support it, attributing their outcomes to luck. Together, these chapters empirically investigate the mechanisms of beliefs, shedding light on how motivated reasoning shapes decisions, social interactions, and policy attitudes. The research of this thesis aims to provide a deeper understanding of human rationality and welfare.

2 Altruistic Self-Image and Self-Deception: A Non-Monotonic Relationship

Efforts to manage one's self-image frequently result in self-deception. The broader literature indicates a positive relationship between these two concepts. However, this research's theoretical framework suggests that this monotonic relationship does not persist in contexts where motivations conflict. To construct a conflicting setting, a novel experimental design is introduced, incorporating monetary incentives, altruistic preferences, and self-image concerns. This design allows for an approximation of self-image through a proxy, avoiding the bias associated with self-report methodologies. Participants received a stimulus indicating their participation in one of two distinct allocation tasks; one posed a moral dilemma between maximizing charitable contributions and personal gain, while the other did not. Misinterpreting the signal incurred a cost and did not change the assigned task. The findings reveal a non-monotonic relationship between self-image concerns and self-deception, with the direction of the bias depending on the dominant incentive; either personal gain or self-image concerns. This divergence suggests two distinct types of self-deception: profit-driven and image-driven. Profit-driven self-deceivers underestimated altruistic opportunities and exhibited low to moderate levels of altruism and self-image concerns. Conversely, image-driven self-deceivers overestimated scenarios necessitating ethical judgment and belonged to groups with heightened levels of altruism and self-image concerns. Given the contrasting tendencies of these self-deception types, the overall mean bias was null, highlighting the significance of integrating self-image considerations into self-deception research.

2.1 Introduction

Many people value being a good person, and this feeling is often tied to altruism. Imagine a group of people with six 1-Euro coins in their pockets walking down a street. On any given day, this street is occupied by five homeless individuals spread out along it. As they pass through the street, these people face a choice: donate one Euro, feel altruistic but become poorer; or withhold the donation, feel less altruistic but remain wealthier.

These passersby could follow one of three strategies. Walking down the street, some prioritize money, caring little about feeling altruistic, and reaching the end of the street with full pockets but no moral satisfaction. Others value their self-image as a good person but also like money. One way to deal with their problem strike a balance in their giving amount that allows them to keep part of their money and feel some moral satisfaction. Or, they could unconsciously adopt a third strategy: self-deception. They notice two homeless people, and give one Euro to each to avoid guilt, while retaining most of their money. This allows them to maintain a positive and high self-image, while keeping their pockets relatively full. A third group, very sensitive to their self-image, seeks to give as much as they can. Not only do they donate one Euro to each homeless person on the street, but they also mistakenly donate one Euro to a man lying on the floor, too inebriated to reach home after a night of heavy drinking. This group of people reaches the end of the street with empty pockets but with an unmatched sense of moral satisfaction.

In all three groups, self-image concerns play a role, either through their absence or their intensity. The same three groups emerge from the experimental results of this paper: people who are unconcerned, profit-driven self-deceivers, and image-driven self-deceivers. The latter group is of particular interest since it reveals that self-image concerns can lead to multi-directional biases, depending not on the structure of preferences, but on their balance. Self-deception is a subcategory of motivated beliefs, which in turn are part of a wider movement in economics: belief-based utility. Molnar and Loewenstein (2022).

The persistence of altruism, even in the absence of social motives, can be explained by a process of self-impression management Murnighan et al. (2001). In a study by Dana et al. (2007), participants decreased their giving when they could deny it to others, thus decreasing the social cost of their decision. They also decreased their rates when they could deny it to themselves, thus lowering the self-image cost of their decision. Similarly, participants in Dana et al. (2006) often decided to incur a 10% reduction in their maximum payout rather than reveal their selfish decision to their partners. In anonymous lying experiments, individuals can choose to be fully honest, fully dishonest, or lie partially, which supports the self-image management perspective of morality-related decision-making Fischbacher and Föllmi-Heusi (2013). Moral balancing is also identified in Ploner and Regner (2013), where participants who cheated to increase their funds, donated from those funds at higher rates. These findings indicate that prosocial behavior is the combined result of a variety of elements. Bénabou and Tirole (2006) model and subsequently locate four motivational sources: altruistic preferences, material self-interest, social image, and self-image concerns. In the current paper, only social-image concerns are excluded.

Having introduced the specific motivated beliefs strategy that will be examined in this paper, it is now time to look at the specific context. In dictator games, the standard economic view once held that dictator-giving should amount to zero. Nevertheless, decades of experimental data have shown that dictator-giving is often positive. Engel (2011) analyzed 616 dictator game treatments and found that the mean dictator giving is 28.35% of the total funds. Yet, as will be discussed in the following paragraphs, much of this altruistic behavior comes with strings attached. Social preferences are complex, and varied, and can be shaped by a variety of factors (for a review, see Fehr and Charness 2023). In the remainder of this section, I will focus on the different side-effects that motivated reasoning has on prosocial behavior.

In research exploring the relationship between self-deception and altruism, I must mention the influential study by Dana et al. (2007), which was successfully replicated in Larson and Capra (2009). In ”Exploiting Moral Wiggle Room”, the authors found that altruistic choices decrease when there is plausible deniability or uncertainty, with participants avoiding costless information that could resolve the uncertainty. Note that costless information avoidance contradicts the traditional economic view of information as a decision-enhancing input. Grossman and Van der Weele (2017) provide further experimental evidence of willful ignorance in social decisions, framing it in the context of self-signaling.

In our efforts to decrease our altruism, we may also distort our beliefs about others. Ambiguity surrounding the preexisting fund allocation of another individual can be interpreted in a self-serving manner, leading to reduced altruistic choices Haisley and Weber (2010), and Di Tella et al. (2015) show that our beliefs regarding the corrupt behavior of others increase when that allows us to act egotistically. However, both the experiment conducted by Ging-Jehli et al. (2020) and its replication, done by Di Tella et al. (2015), found no experimental evidence of absolute strategic cynicism. On the contrary, participants in the experiments appeared to hold excessively positive views regarding the behavior of others in absolute terms. Where Ging-Jehli et al. (2020) agrees with the previous research is the existence of relative strategic cynicism. Specifically, individuals who had a lot to gain by acting egoistically tended to convince themselves of others' bad behavior, and individuals who had little power to act egoistically tended to convince themselves that the other would behave altruistically.

This study contributes to the understanding of self-deception in decision-making, particularly in contexts involving altruism and self-image concerns. The participants in my experiment made allocation decisions between themselves and a charity in two different scenarios. In one scenario, they could maximize both their own payout and that of the charity at no cost (*win-win*). In the other scenario, maximizing the charity's payout necessitated a reduction in their own payout (*moral conflict*). Participants received a noisy signal to inform them of the scenario they were in. Identifying the signal is irrelevant to determining which is the active scenario, participants had a monetary incentive to correctly identify it.

The experiment is designed to meet all three conditions required to separate motivated cognition from heuristics (Bénabou and Tirole, 2016), and all four conditions necessary for this motivated cognition to be defined as self-deception (Gur and Sackeim, 1979). Nevertheless, I attempt to refine the first condition of Bénabou and Tirole (2016) regarding the directionality of motivated cognition. The bias that the participants in this research show maintains a directionality, but the direction of the bias differs based on the magnitude of their beliefs rather than their type. In contrast to Ging-Jehli et al. (2020), this difference in direction is not a result of a change in the belief at stake. Ging-Jehli et al. show that, in a relatively similar setting, individuals may reverse the direction of their beliefs about others depending on which belief about themselves is at stake. When the belief involved is fairness, and their actions do not align with this belief, individuals tend to convince themselves that their counterpart acted unfairly as well, therefore deserving the mistreatment. When forced to act virtuously, the belief at stake becomes whether they are being exploited. Consequently, they may convince themselves that the other acted virtuously and did not take advantage of the situation. In the present research, the belief at stake—altruism—remains the same; however, the direction of the bias changes depending on the magnitude of that belief. Moreover, the groups that follow these directions are identifiable. This represents a significant contribution of my experimental research, because it implies that people with similar beliefs and utility structures can display opposing biases in the same context. These opposing biases result in sub-optimal decisions and loss of welfare.

This study's experimental design enables the investigation of opposing beliefs by measuring proxies of beliefs in a self-signaling setting with conflicting motivations. The categorization of participants into distinct belief groups is a significant novelty of this research. In addition, the experimental design also allows for thorough testing of previous findings. It incorporates various levels of ambiguity, as well as assessments of reaction time.

The study of reaction time poses several challenges Spiliopoulos and Ortmann (2018). However, in this experiment, reaction time can serve as an indicator of self-deception. If less information is required to accept belief-confirming signals Ditto and Lopez (1992), we can expect participants to have shorter reaction times when identifying their preferred outcome. Since the favorable outcome varies across belief groups, reaction times should reflect these differences.

Thanks to the novel experimental design of this research, I can investigate self-deception from multiple perspectives. The most important contribution comes from the measurement of a belief proxy and the categorization of participants into distinct belief groups. This partitioning reveals opposing biases that arise from the same type of belief—a phenomenon that has not been previously documented in the literature.

2.2 Experimental Design

This experiment explores the relationship between self-deception and altruistic behavior in an ambiguous context. Participants are presented with two scenarios: *win-win* and *moral conflict*. In the win-win scenario, participants can maximize both personal and charity payouts, while in the "moral conflict" scenario, maximizing personal gain reduces charity payout. The charity chosen is "Give Directly", a charity that provides cash directly to impoverished individuals. The choice of Give Directly is a result of an effort to avoid any prior biases regarding the cause of a charity.[1]

The decision-making process involves choosing between actions A and B (Table 1) in one of the two scenarios. Barring any unforeseen combinations of preferences, the win-win scenario is the preferable one, since it maximizes everyone's payout. In the "moral conflict" scenario, more complex combinations of preferences play a role. For someone to give up 2 Euros by choosing action B over action A in the conflicting scenario, they must have some altruistic preferences. Of course, an individual may possess altruistic preferences that are

[1] **Click here to see this experiment's instructions.**

Table 2.1: Scenarios

Win-Win			Moral Conflict	
Give Directly	You		You	Give Directly
5	7	A	7	1.5
1.5	5	B	5	5

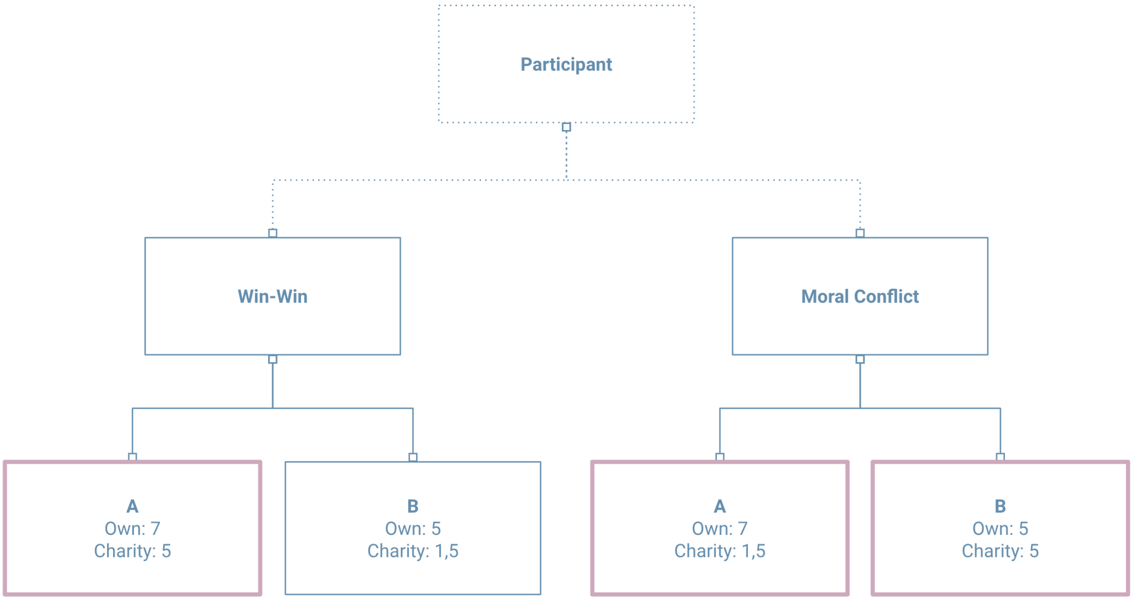


Figure 2.1: Experiment Decision Tree

not strong enough to justify losing 2 Euros. Finally, action B in the conflict scenario is designed to be the most efficient choice, so that the more selfish option (action A) is less appealing.

Ambiguity is of vital importance in the manifestation of self-deception, and the experimental setting introduces it with a stimulus signal that indicates the active scenario. Before each decision, participants are shown a screen displaying 84 dots for 250ms. The dots are unequally spread between the right and left sides of a diagonal line unequally. At the beginning of each trial block, participants are informed about the interpretation of the stimulus. During one block, seeing more dots on the left side of the diagonal line means that the active scenario is the win-win scenario, while during the other, it means that the active scenario is the morally conflicting one. The ambiguity level is determined by the difference in the number of dots on each side (Figure 2.2). For example, if more dots on the right represent the moral conflict scenario, stimuli marked in red refer to the moral conflict scenario and the one in green to the win-win scenario.

To ensure truthful stimulus identification reports, participants are incentivized to accurately report their perceived scenario with 1 Euro. Participants must identify the active scenario based on the stimulus, knowing that their response does not influence the actual scenario they are in. For example, if there are more dots on the right side of the diagonal

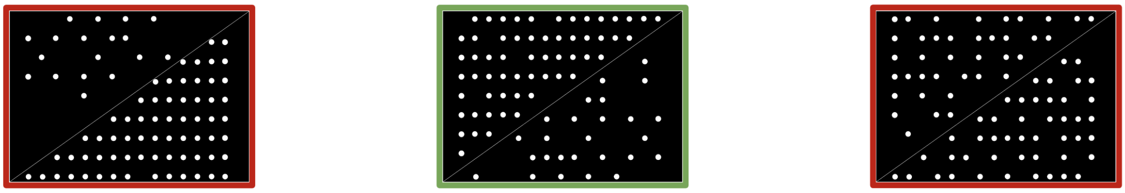


Figure 2.2: Dot Stimuli Examples

line, i.e. signifying the moral conflict scenario, choosing action A always gives 7 Euros and 1.5 Euros to the charity, regardless of which side they believe has more dots. Let's assume that the screen presented to a participant is the one in Figure 2.2 (a), with more dots on the left of the line; this would mean that the participants are making decisions in the win-win scenario. If they correctly report ”left” as the side with more dots, they receive 1 Euro as a bonus. If they report ”right” as the side with more dots, they do not receive the 1 Euro bonus. Note, that even if they incorrectly choose ”right” as the answer, the active scenario remains the same (win-win). The signal in Figure 2.2 (a) is stronger (44 more dots on the left) than in Figure 2.2 (b) (16 more dots on the right), therefore, there is more room for self-deception in the second case.

The subjects of the experiment complete two blocks of trials, with the only difference between them being the reversed interpretation of the stimuli. Each session includes 48 trials. The sequence of events for each trial is stimulus presentation, pattern identification, and allocation decision. At the end of the experiment, participants are incentivized to report the percentage of correct pattern identifications. This last assessment is used to gauge potential awareness of self-deception. Finally, some sociodemographic questions are collected.

In summary, most participants would prefer to see a stimulus that signifies a scenario where both parties maximize their payouts. However, their stimulus identification does not affect the active scenario, and a correct identification is monetarily incentivized. At the same time, as the stimulus signal grows stronger, unconsciously manipulating information becomes harder.

2.3 Theoretical Framework

2.3.1 Modeling Approach

The model presented below attempts to describe a morally conflicting decision; where the source of the moral conflict is an altruistic self-image. Similarly to Engelmann et al. (2024) and its model of anticipatory anxiety, I combine a few different approaches to the modeling of self-deception to describe the current setting. As explained in previous sections, the task that the participants undertake fulfills all four conditions of self-deception that were laid out in Gur and Sackeim (1979). The self-deception this experiment deals with is a result of an attempt at self-signaling Mijović-Prelec and Prelec (2010). Namely, individuals interpret a signal and, based on their interpretation, the self-image value of their later decision changes. So, in addition to the material payout, a self-image value is added to the expected utility function. Since individuals might also have genuine altruistic preferences, these preferences are also included in the expected utility function. Therefore, three out of the four elements of prosocial behavior outlined in Bénabou and Tirole (2006) are included (altruistic preferences, material self-interest, and self-image concerns), while only social-image concerns are left out. Ambiguity is important when it comes to self-deception; to account for it, I follow the approach of Bénabou and Tirole (2002) by including a cognitive cost of self-deception, which increases as the context becomes less ambiguous. Finally, in line with the optimal expectations model of Brunnermeier and Parker (2005), this model treats beliefs as objects of choice, though, in an indirect manner. Specifically, individuals do not choose their beliefs but the magnitude and direction of their bias. This final modification of optimal expectations does not significantly change the predictions of the model; however, it allows for a clearer and more interpretable equilibrium which highlights the directionality of self-deception.

2.3.2 The Model

R and L refer to the dot identification side (right/left). Therefore, the utility derived by correctly identifying "right" as the side with the most dots for someone who always chooses action A no matter which scenario they are in is:

$$U_R[(R,A),(L,A)|R]$$

Self-deception can present itself in individuals with heterogeneous characteristics and preferences, therefore, we must study the sets of strategies that encompass these differences. It is also of interest to study the cases where individuals choose B in one of the two scenarios, $U_R[(R,B),(L,A)|R]$. The cases where the left side has more dots are designed to be symmetrical, so only one side is included in the analysis. In the remainder of this section, I will first describe the model using the $[U_R[(R,A),(L,A)|R]$ strategy profile as an example and derive the self-deception equilibrium for this strategy. Then, I derive the self-deception equilibrium for those individuals who choose B in one of the two scenarios ($U_R[(R,B),(L,A)|R]$). The main assumption of this analysis is that individuals prefer action A (7 Euros for them and 5 for the charity) in the win-win scenario over action B (5 Euros for them and 1.5 for the charity) in the same scenario. Finally, given the previous assumption, I provide the condition that separates the strategies based on individual characteristics.

Always choosing A: $U_R[(R,A),(L,A)|R]$

In the present model, the object of choice is not the belief, but the bias (ω) of the individual. Bias skews the probability that the individual assigns to the scenario to maximize the expected payout. p denotes the initial perceived probability assigned to the scenario;

m denotes the correct stimulus identification reward. ΔY denotes the difference in charity payout between choosing A in the two different scenarios; ΔX denotes the difference in personal payout between choosing A in the two different scenarios; α denotes the individual's altruism and determines how much ΔY affects the expected utility. By design, $\Delta X = 0$ when moving from (R, A) to (L, A).

$$(p + \omega) \cdot (m + \alpha \cdot \Delta Y + \Delta X)$$

This first component represents the material self-interest and altruistic preferences of the individual. It includes the potential monetary reward (m) of the participant in case they identify the scenario correctly, the monetary reward of the participant according to the choice they made (ΔX) in the scenario they are in and the monetary reward of the charity (ΔY), which also depends on the choice made by the participant in the scenario and is weighted by a true altruism factor (α). The expected utility of the first component depends on the probability that the participant assigns on the (ω). So far, altruistic preferences and material self-interest have been included, the only prosocial behavior component missing is self-image concern. Nevertheless, before self-image concerns can be added, some limits to self-deception must be included.

As is evident from the above expression, an individual can always increase (ω) to increase their expected utility. Of course, that is unrealistic, since it implies that, if desired, no amount of unambiguous information would ever reach consciousness. To solve this problem, and to insert a role for ambiguity in the model, a cognitive cost of biased thinking $(\lambda(s))$ is added. The cognitive cost increases with the strength of the signal (s). When the signal (s) is low, ambiguity is high and the cognitive cost of self-deception is low. When ambiguity is low, the cost limits bias and incentivizes the subject to bring ω to 0.

$$-\lambda(s) \cdot \omega^2$$

In simple terms, the cognitive cost shows that when evidence is too strong, it is very difficult to ignore or misinterpret the evidence. In the context of the experiment, this term indicates that self-deceiving when there are 40 more dots on one side is harder than when the difference is only 8.

Subsequently, I am adding the third element of prosocial behavior, self-image concerns (σ), which capture the increase in utility derived by a positive self-image update related to the individual's choice between A and B, or the decrease of it by a negative update. c is equal to 1 when it signifies the moral conflict scenario and 0 when it signifies the win-win scenario.

$$-\sigma \cdot [c \cdot (p \cdot \omega) + (1 - c)(-(p \cdot \omega)$$

Some individuals are more sensitive to their self-image than others, this component includes a factor that represents that. When the scenario represented by a stimulus with more dots on the right is morally conflicting, positively biasing their belief in the scenario implies a larger reduction of utility, therefore, some individuals might add a negative bias to their belief to avoid the cost. Likewise, if the scenario is non-conflicting, the individual might increase their chosen belief through the bias term (ω). The size of the cost of changing these beliefs depends on each participant's self-image concerns, or in other words, on their self-image sensitivity to their actions. In the absence of self-image concerns, the cognitive cost is assumed to be strong enough to bring (ω) to zero.

Put together, the expected utility is:

$$U_R[(R,A),(L,A)|R] = (p+\omega)\cdot(m+\alpha\cdot\Delta Y+\Delta X)-\lambda(s)\cdot\omega^2-\sigma\cdot[c\cdot(p+\omega)+(1-c)(-(p+\omega)] \tag{2.1}$$

Maximizing the expected utility with respect to parameter ω, yields:

$$\omega^* = \frac{\alpha \cdot \Delta Y - 2 \cdot c \cdot \sigma + m + \sigma}{2 \cdot \lambda(s)}$$

The amount of self-deception can be measured by the difference in the belief that the answer is correct between the cases that the right side scenario is the win-win scenario ($c = 0$) and those that it is the moral conflict scenario ($c = 1$). So, self-deception in this case is defined as $\omega(0) - \omega(1)$.

In this model, the existence of the moral conflict scenario is implicit and related to the material payout. As a result, when $c = 1$, Y_R is negative ($\Delta \underline{Y}$) (which is what makes R the moral conflict scenario), and when $c = 0$, Y_R is positive ($\Delta \bar{Y}$).

In this experiment, $\Delta \bar{Y}$ = - $\Delta \underline{Y}$, therefore, the equilibrium amount of self-deception is:

$$SD_{AA} = \omega^*(0) - \omega^*(1) = \frac{\alpha \cdot \Delta Y + \sigma}{\lambda(s)} \tag{2.2}$$

Choosing B in one scenario: $U_R[(R,B),(L,A)|R]$

The expected utility now takes the following form:

$$U_R[(R,B),(L,A)|R] = (p+\omega)\cdot(m+\Delta X)-\lambda(s)+\omega^2+\sigma\cdot[c\cdot(p\cdot\omega)+(1-c)\cdot(-(p+\omega))]$$

The difference with the previous strategic profile is that now $\Delta X > 0$ and $\Delta Y = 0$ by design, and self-image concerns are not costly, but profitable (sign reversal), since choosing B when facing a moral conflict scenario is the altruist choice, providing a positive self-image.

Following the same steps as above we find:

$$\omega^* = \frac{\Delta X + 2 \cdot c \cdot \sigma + m - \sigma}{2 \cdot \lambda(s)}$$

And:

$$SD_{BA} = \omega^*(0) - \omega^*(1) = \frac{\Delta X - \sigma}{\lambda(s)} \tag{2.3}$$

Choice under a known moral conflict scenario

It is also interesting to study the conditions under which someone chooses B over A. In the win-win scenario, the utility derived by choosing A is always larger than choosing B since both the individual and the charity maximize their payout, therefore, in this section, the focus is the moral conflict scenario.

Under the assumption that the moral conflict scenario is known, $p = 1$; therefore, individuals choose B in the moral conflict scenario when:

$$U(A|Conflict) < U(B|Conflict)$$

$$=> X_A + Y_A \cdot \alpha - \sigma < X_B + Y_B \cdot \alpha + \sigma$$

$$<=> \frac{\Delta X - 2 \cdot \sigma}{\Delta Y} < \alpha \tag{2.4}$$

Rearranging the terms to reflect material interests on the left and altruistic preferences along with self-image concerns on the right provides a clearer picture:

$$<=> \Delta X < \alpha \cdot \Delta Y + 2 \cdot \sigma$$

A larger material payout decreases altruistic choices, while altruism and self-image sensitivity both increase altruistic choices. Given that the values of ΔX and ΔY are known, we can simplify further:

$$=> 1.75 \cdot \alpha + \sigma > 1$$

Note that - although the altruism factor is additionally weighted - a sufficiently high level of self-image concerns could also make an individual choose the altruistic outcome.

2.4 Model Analysis

2.4.1 Factors of self-deception

Self-Image of Altruism Sensitivity

Altruism self-image sensitivity (σ) increases self-deception when both altruism (α) and σ are low enough that inequality 2.4 is not satisfied. If both of these terms are sufficiently high—and the inequality is satisfied—σ decreases self-deception and can even make it negative when it becomes very large.

Individuals with characteristics that do not satisfy inequality 2.4 prefer to hurt the charity than to lose money in the morally conflicting scenario. If σ is low, these individuals' self-image does not get hurt by their selfish actions, therefore, they do not attempt to self-deceive. Nevertheless, if σ is high enough to be costly for these individuals, but not high enough to make them choose the altruistic outcome, they will attempt to self-deceive to protect their self-image.

On the other side, when inequality 2.4 is satisfied, individuals choose the altruistic outcome. However, not all of them with the same excitement. Those individuals who only barely satisfy the inequality have a low σ and would prefer to just be in the win-win scenario and make more money. As σ increases, the utility from acting altruistically also increases, making the morally conflicting scenario more appealing. In a more extreme scenario, a high σ could provide satisfaction from acting altruistically that is large enough to make the morally conflicting scenario preferable for these individuals. These individuals might attempt to self-deceive in the opposite direction, resulting in negative self-deception.

Altruism

Similarly to self-image concerns (σ), altruism (α) exhibits a complex relationship with self-deception. When inequality 2.4 is not satisfied, a higher level of altruism does increase self-deception. These individuals choose the non-altruistic outcome and will not self-deceive when their level of altruism is low. However, when their altruism is sufficiently high, the win-win scenario allows them to significantly increase their expected utility, so they might attempt to self-deceive towards that scenario. When both α and σ are large enough, inequality 2.4 is satisfied and individuals prefer the altruistic choice in the morally conflicting scenario. Altruism does not affect their self-deception levels anymore because of the payout structure of this experiment ($\Delta Y = 0$).

Monetary incentive and payouts

While existing theoretical models suggest that a higher monetary incentive reduces self-deception levels, the present model predicts that the size of the monetary incentive is not relevant. This result is in line with previous experimental findings (Engelmann and Pessoa, 2014; Engelmann et al., 2009).

By design, ΔY is only relevant for SD_{AA} (2.2) and ΔX is only relevant for SD_{BA} (2.3). In both cases, an increase in either ΔY or ΔX increases the level of self-deception towards the win-win scenario. More specifically, individuals who tend to choose the altruistic option in the morally conflicting scenario, are tempted to self-deceive towards the win-win scenario when ΔX increases. Individuals that tend to choose the non-altruistic option in the morally conflicting scenario already maximize their monetary payout, therefore, are only tempted to self-deceive towards the win-win scenario if ΔY increases, and as long as they have a positive α.

Cost of Self-Deception

The cost of self-deception ($\lambda(s)$) limits self-deception, as expected. The more ambiguous the setting, the lower the cost and the higher the absolute level of self-deception.

2.4.2 Measuring Altruism and Self-image

The main goal of this research is to disentangle the relationship between altruism and self-deception. Figure 2.3 shows the expected relationship that altruism and self-image sensitivity have with self-deception. Both characteristics increase (profit-driven) self-deception until inequality 2.4 is reached (strategy profile: Always A); after that point (strategy profile: A and B), true altruism ceases to play a role, and self-image sensitivity decreases profit driven self-deception. A strong enough self-image sensitivity could create self-deception in the opposite direction (image-driven).

Directly measuring self-image (α) is not feasible. To address this, a function $\Phi(\alpha, \sigma)$ is assumed, describing an individual's exhibited altruism in unambiguous scenarios without external pressure. Φ increases with both true altruism and self-image sensitivity. It quantifies altruistic choices where self-deception is not possible. Ambiguity is experimentally increased (increase in s), making the internal self-deception cost ($\lambda(s)$) prohibitively high. To achieve this, 8 of 48 trials included strong signals with little room for identification errors. Half of these trials signify a morally conflicting scenario. The share of altruistic choices in those 4 trials represents the Φ for each individual.

Although inequality 2.4 implies strict and constant values of α and σ for each individual, in practice individuals might slightly vary these values implicitly in a repeated trial setting, therefore, both of these factors should be seen as averages and individual trials might deviate from these averages. For instance, an individual for whom inequality 2.4 holds on average might choose the altruistic outcome 3 times out of the 4 unambiguous trials but deviate once by choosing the non-altruistic outcome. An individual with particularly low values of α and σ will have $\Phi = 0$ and individuals with very high values will have $\Phi = 1$. For individuals with intermediate values of α and σ, $\Phi \in [\frac{1}{4}, \frac{2}{4}, \frac{3}{4}]$. For simplicity, these groups will be referred to as follows:

- $\Phi = \frac{0}{4} \rightarrow$ Selfish
- $\Phi = \frac{1}{4} \rightarrow$ Partially selfish
- $\Phi = \frac{2}{4} \rightarrow$ Neutral
- $\Phi = \frac{3}{4} \rightarrow$ Partially altruist
- $\Phi = \frac{4}{4} \rightarrow$ Altruist

Although (α) and (σ) are not necessarily correlated, they are both positively correlated with Φ. The reversal of the relationship between (σ) and self-deception after inequality 2.4 provides a unique opportunity to observe whether self-image concerns dominate. Figure 2.4 shows the relationship between the measurable Φ and self-deception.

2.4.3 Econometric Model

A random-effects probit regression model will be used, with the probability of mistake as the dependent variable. The independent variables are the strength of the signal ($\lambda(s)$), a

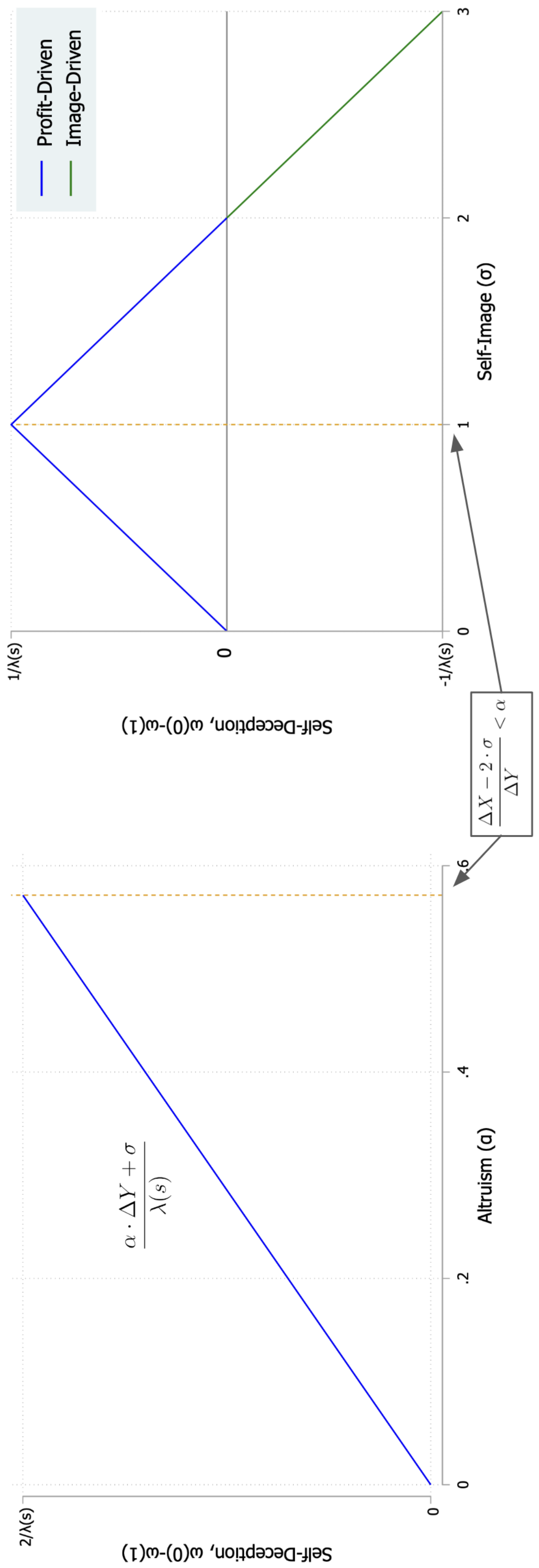


Figure 2.3: Self-deception and altruism α self-image σ

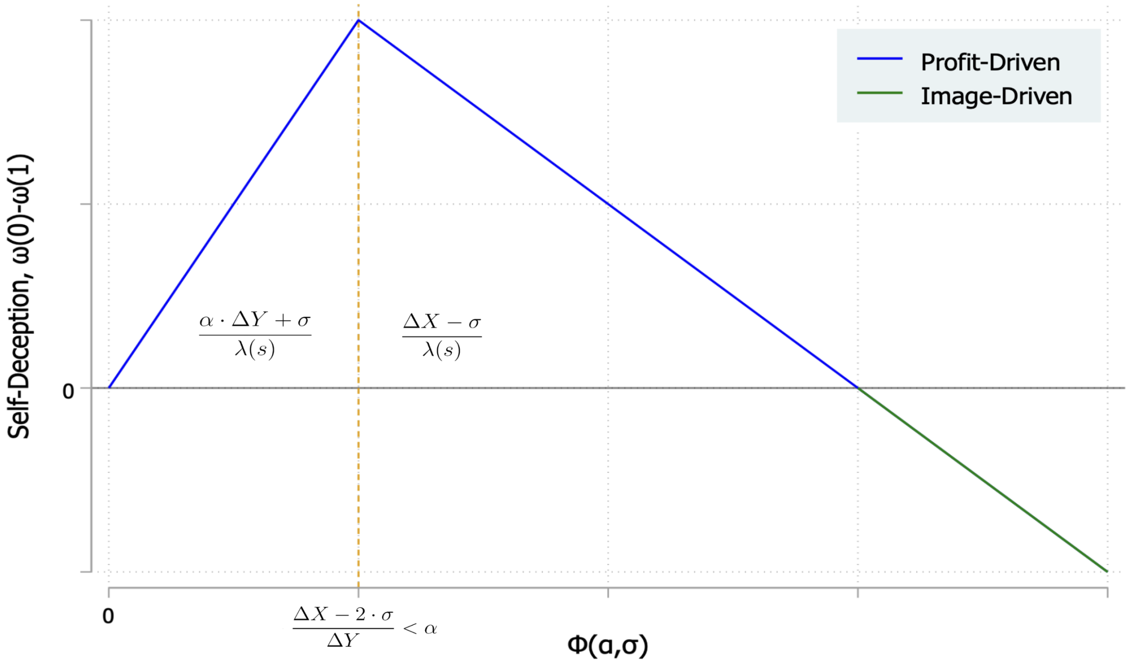


Figure 2.4: Self-deception and proxy of self-image and altruism Φ

binary variable indicating whether the trial refers to the moral conflict scenario (C), and the proxy of altruistic preferences, Φ. An interaction term between $\lambda(s)$ and C is added to capture the effect that ambiguity has specifically on biased mistakes. Finally, interaction terms between C and Φ are included. The econometric model takes the following form:

$$\begin{gathered} Y_i = \beta_0 + \beta_1 \cdot \lambda(s) + \beta_2 \cdot \ \mathrm{C} + \beta_3 \cdot \ \mathrm{C} \cdot \lambda(s) \\ + \beta_4 \cdot \Phi_{\frac{1}{4}} + ... + \beta_7 \cdot \Phi_{\frac{4}{4}} \\ + \beta_8 \cdot \mathrm{C} \cdot \Phi_{\frac{1}{4}} + ... + \beta_{11} \cdot \mathrm{C} \cdot \Phi_{\frac{4}{4}} + \epsilon_i \end{gathered}$$

2.4.4 Hypotheses

Hypothesis 1, There is self-deception among participants.

A self-deceiving mistake is a mistake that happens when $C = 1$, namely, when the scenario is morally conflicting.

H_0: $b_2 = 0$
H_1: $b_2 > 0$

Hypothesis 2, Ambiguity increases self-deception.

Mistakes are expected to generally increase with the ambiguity of the stimuli since the task becomes objectively harder. Our model also predicts that self-deception (W) increases with the ambiguity of the stimulus, s. In the econometric model, this is represented with the interaction term between C and $(\lambda(s))$.

H_0: $b_3 = 0$
H_1: $b_3 > 0$

Hypothesis 3, Self-deception peaks in neutrals ($\Phi(\alpha, \sigma) = \frac{2}{4}$).
Similarly to the second hypothesis, we are interested in the set of interaction terms between the moral conflict scenario, C, and the proxy of altruism, Φ. Φ will be positively correlated with self-deception up to $\Phi = \frac{2}{4}$ and negatively from $\Phi = \frac{3}{4}$ to $\Phi = \frac{4}{4}$. Neutrals are expected to exhibit the highest levels of self-deception, while perfectly selfish and perfectly altruist participants will exhibit the lowest amounts of self-deception.

H_0: $b_8 = ... = b_{11} = 0$
H_{1_A}: $b_8 > 0$
H_{1_B}: $b_9 > 0$
H_{1_C}: $b_{10} > 0$

2.5 Results

Two days of sessions were conducted at Pompeu Fabra University's Behavioral Experimental Sciences Laboratory (BESLab) and three days of sessions were conducted at the Behavioral Experimental Science Laboratory of the University of Barcelona. One of the session days had to be excluded from the analysis due to excessive rational choice violations (see Appendix A2). This section presents the experimental results from 190 participants (106 female, 80 male, 4 who did not identify with the former two). The average age of the participants was 21.16 years and the majority of them (81.59%) were undergoing their undergraduate studies. Finally, 6 Φ groups were identified and included 44 Selfish, 12 Partially Selfish, 15 Neutrals, 20 Partially Altruist, 90 Altruist, and 9 Virtue Consumers. The last group refers to participants who even made a mistake in the easiest 8 trials, resulting in choosing the altruistic outcome one additional time compared to the expected maximum. Instead of adding this group with the altruists, the group has been isolated due to its outlier behavior.

2.5.1 Descriptive Results

The stimulus task produced a total error rate of 6.18%, with the majority of errors, 84.9%, committed in the hardest 3 levels of the 12 levels of difficulty of the task. The easiest two levels of the task are excluded from the mistake analysis to avoid endogeneity since those two levels are used to build the proxy of altruism and self-image, Φ. Figure 2.5 shows the average mistake rates per difficulty level. The stronger a signal is, the easier the trial. As can be seen, the error rates remain close to zero for signals larger than 20. Table 2.2 compares the number of mistakes between the hardest 20 and the easiest 20 trials. There were almost 9 times more mistakes in the harder trials (12.66%) than in the easier trials (1.42%).

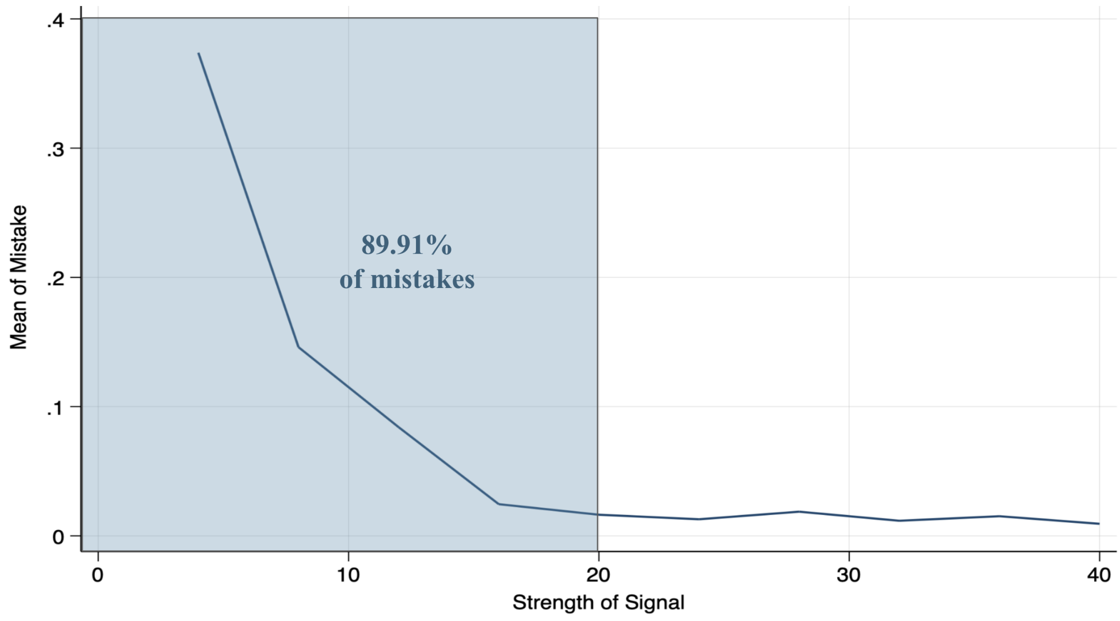


Figure 2.5: Mistakes By Difficulty

Due to this large inequality, I am proceeding with this analysis by including tables, graphs, and regressions at both the 40-trial level and the hardest 20-trial level. Self-deception requires ambiguity (Balcetis and Dunning, 2006; Sloman et al., 2010; Mei et al., 2023); therefore, an analysis of the hardest 20-trial level can tell us more about motivated cognition than the overall 40-trial level. Additionally, given the very low error rate of the easiest 20 trials, a comparatively larger share of mistakes could have been a result of circumstances such as accidental button presses or distraction.

	Hardest 20	**Easiest 20**	**Total**
Correct	3,319 (46.98%)	3,746 (53.02%)	7,065 (100.00%)
Mistake	**481 (89.91%)**	**54 (10.09%)**	535 (100.00%)
Total	3,800 (50.00%)	3,800 (50.00%)	7,600 (100.00%)

Table 2.2: Mistakes by signal group difficulty

Table 2.3 summarizes the mistake rates of each group. The mistake rates of participants when faced with a win-win scenario do not vary substantially. However, when faced with the moral conflict scenario, participants displayed a wide range of error rates. The moral conflict scenario error rate of virtue consumers was 5.56% while that of the partially selfish was more that double that rate, at 11.67%. Net bias refers to the absolute difference between the error rates in moral conflict and win-win scenarios for each participant. As can be seen on the fourth column of Table 2.3, net bias peaks among the partially selfish and is at its minimum among the altruists, with rates of 3.75% and -1.28% respectively. When we focus on the hardest 20 trials, the picture becomes clearer. The net bias of all groups becomes larger towards directions in line with the existence of self-deception. In the face of increased ambiguity, the selfish group exhibited a small increase of 0.22 percentage points, implying that its bias might be circumstantial. While other groups such as the partially selfish, displayed an increase of 4.58 percentage points, a 222% increase compared to the 40% level net bias. Finally, note that the overall net bias across groups is almost zero, with a value of -0.08% at the 40-trial level and -0.05% at the hardest 20-trial level. Studying the overall net bias of the participants would lead to the conclusion that there are no biases in mistake rates. Among the - self-deception facilitating - 20 hardest trial, the net bias is actually slightly lower, which would imply that there is no self-deception present. However, the individual mistake rates of the groups tell a different story. This result highlights the importance of including separation measures based on beliefs and preferences when we research biases and self-deception.

		Moral			**(20 Trials)**
	Mistakes	**Conflict**	**Win-Win**	**Net Bias**	**Net Bias**
Φ					
Selfish	0.0761	0.0818	0.0705	0.0114	0.0136
Partially Selfish	0.0979	0.1167	0.0792	0.0375	0.0833
Neutrals	0.0833	0.0900	0.0767	0.0133	0.0267
Partially Altruist	0.0638	0.0625	0.0650	-0.0025	0.0050
Altruist	0.0642	0.0578	0.0706	-0.0128	-0.0222
Virtue Consumers	0.0611	0.0556	0.0667	-0.0111	-0.0222
Total	0.0704	0.0700	0.0708	-0.0008	-0.0005

Table 2.3: Summary of Mistakes per Group

Figure 2.6 graphically represents the mistake rates of the groups by type of scenario. As we saw in Table 2.3, win-win mistake rates do not exhibit a pattern, while moral conflict mistake rates are higher in the first three groups and lower in the last three. Of course, overall error rates are higher for all groups when we zoom in on the hardest 20 trials. Nevertheless, this increase in the error rate does not necessitate a directional increase in the net bias.

The net biases are amplified when we study the higher ambiguity trials, which facilitate self-deception. Figure 2.7 exhibits this pattern. Additionally, Figure 2.8 isolates the changes in the net bias of each group between the 40-trial level and the hardest 20-trial level. It must be highlighted that these are expected patterns according to the theoretical framework of this research. The changes are not uniform and are larger and directional only for the partially selfish, and neutrals, who exhibit patterns of profit-driven self-deception; and for the altruist and virtue consumers, who exhibit patterns of image-driven self-deception.

The fact that the above-mentioned pattern is observed not only implies that there might be self-deception, but also that the use of Φ as a proxy of altruistic preferences and self-image is effective. Figure 2.16 includes a breakdown of bias direction per altruistic preference group in the 20 hardest trials (see tables in Appendix A3 for figure including

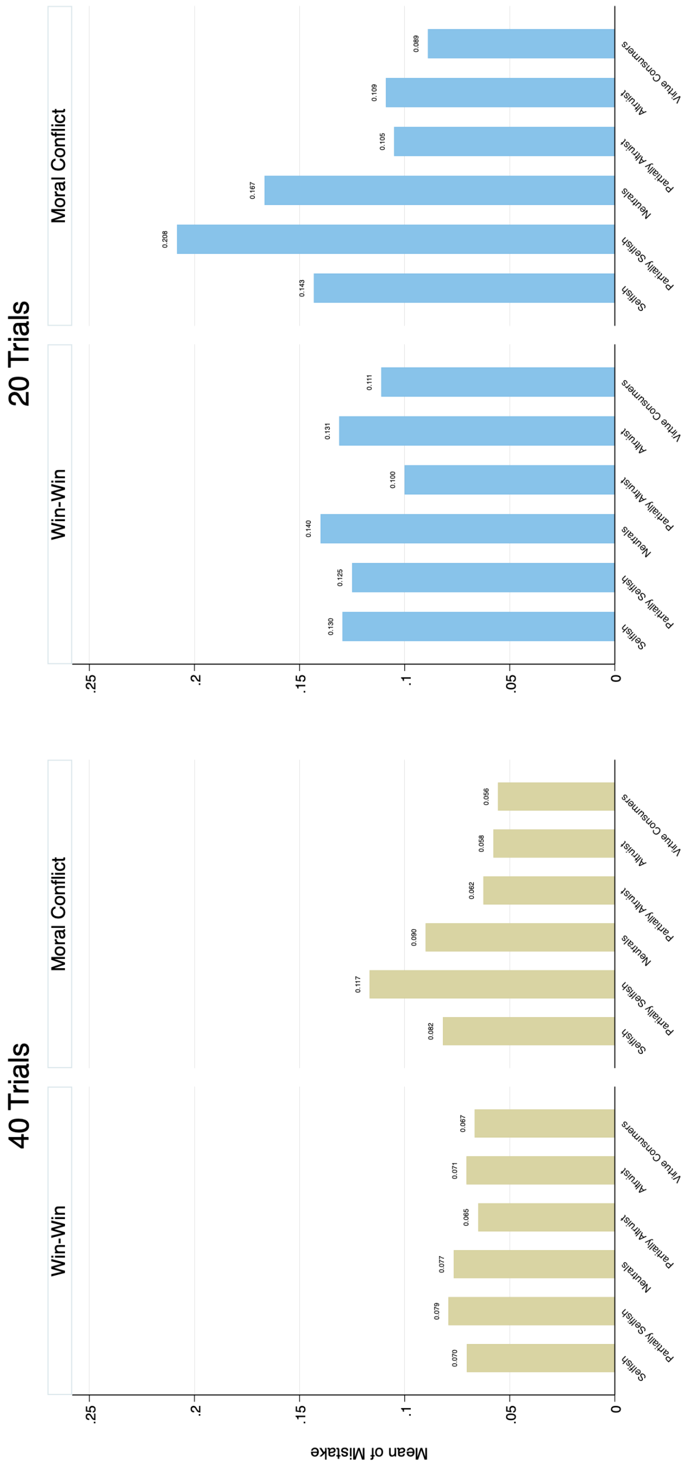


Figure 2.6: Type of Mistake by Group, and 40 and 20 trials

all 40 trials). A negative net bias is characterized as image-driven self-deception and a positive one as profit-driven. The middle bars refer to participants who did not exhibit a bias in their mistake direction. A comparison of the image and profit columns of the figure reveals that the bias direction of participants matched the model's predictions. The

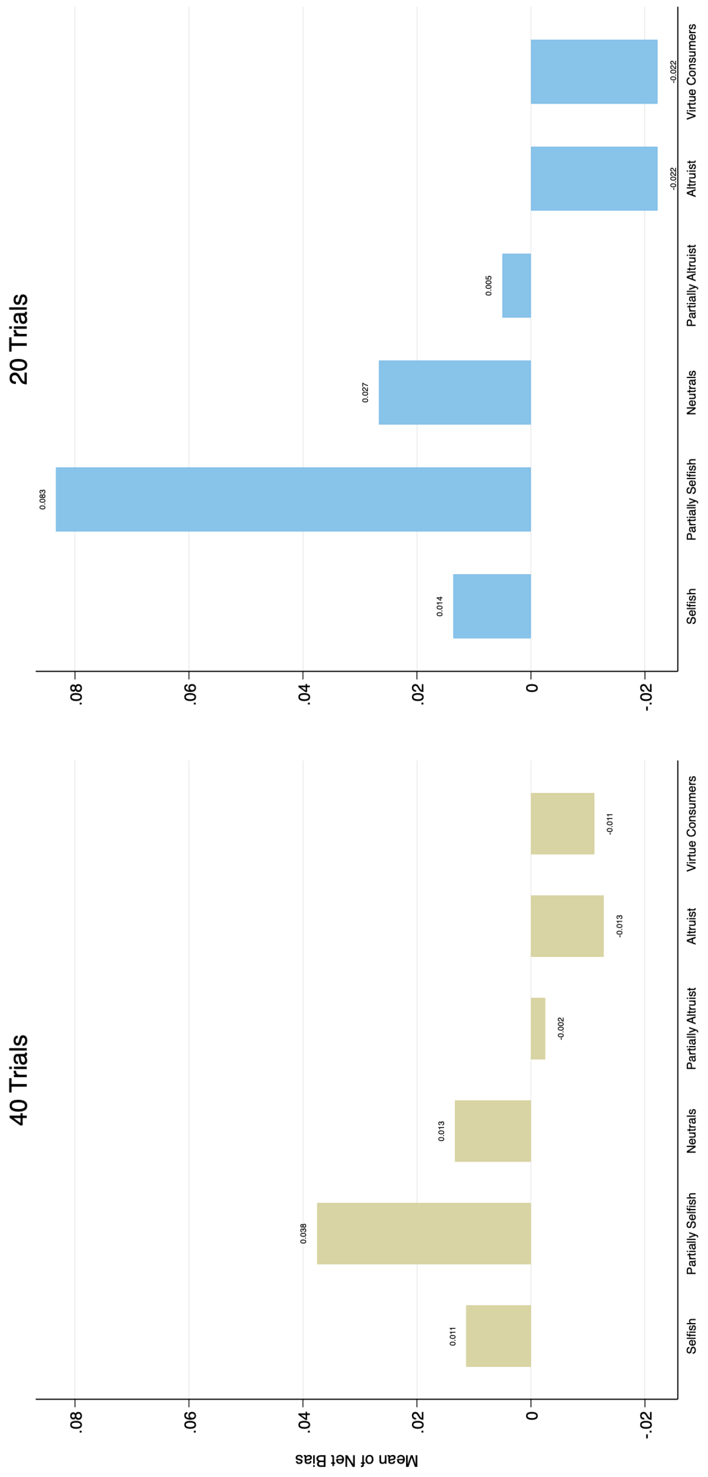


Figure 2.7: Net Bias by Group, 40 and 20 trials

imbalance between image and profit bias, in favor of the latter, begins already among the selfish and continues with the partially selfish and neutrals. Starting with the partially altruist, the imbalance is reversed and a bias towards the morally conflicting scenario is favored. The bias implying image-driven self-deception continues to be favored among the

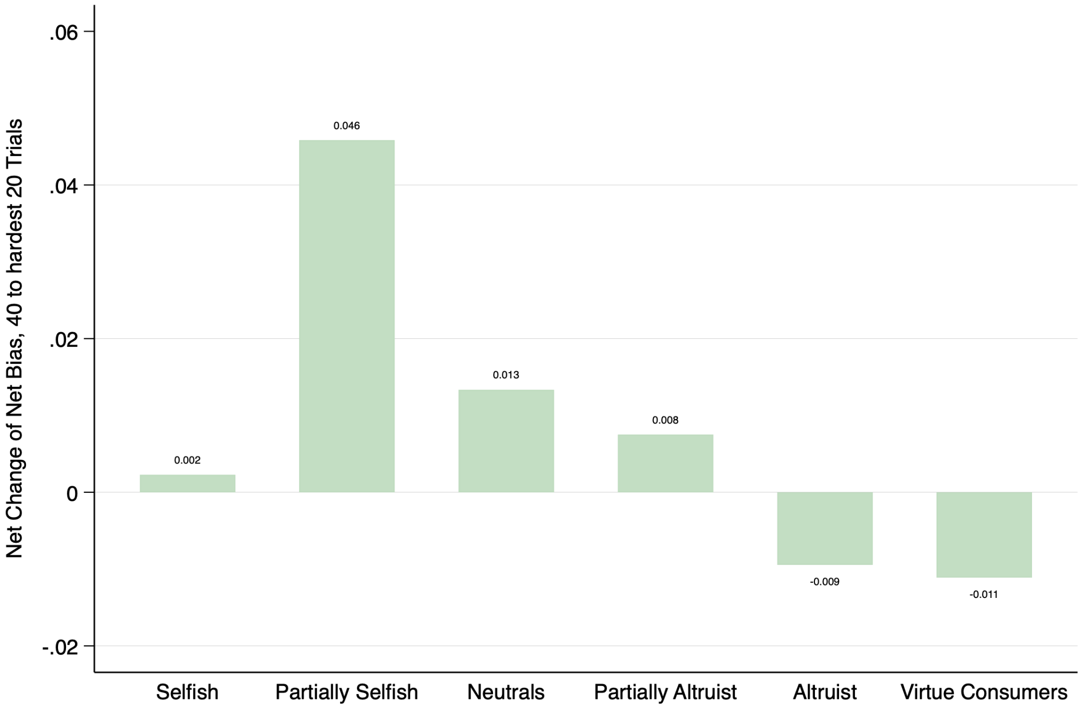


Figure 2.8: Net Change of Bias by Group, 40 to 20 trials

altruists and its dominance peaks among the virtue consumers. The largest difference is found in the partially selfish. Half of the partially selfish exhibited a bias towards the win-win scenario, while only 8.33% of them displayed a bias towards the morally conflicting scenario.

Reaction time can be a complicated topic to study when it comes to self-deception. Although there is evidence that less information is needed to reach a favorable conclusion Ditto and Lopez (1992), in the case of this experiment, the amount of information received was predetermined and presented to the participants for a predetermined amount of time. Figure 2.10 shows the net speed of making a correct or wrong scenario identification. Making a mistake generally takes longer, since the trials with the most mistakes are also the harder trials. In the figure, when the value of a bar is positive it signifies that the group it refers to took more time to identify, correctly or incorrectly, the win-win scenario. When the value is negative, the group of participants spent more time to identify the moral conflict scenario compared to the win-win scenario. The reaction time pattern matches the net bias pattern both for correct and incorrect identifications. More specifically, the partially selfish group took almost 7 seconds longer to mistakenly identify win-win as the correct scenario, while the virtue consumers spent almost 5 extra seconds on average when identifying the moral conflict scenario. These patterns imply a possible self-convincing in the decision-making process. Figure 2.11 includes only the net speed of the incorrect identifications of each group, showing that the reaction time differences of the partially selfish and virtue consumers are non-zero at their respective 95% confidence intervals.

The above results do not provide definite proof of self-deception. Nevertheless, the

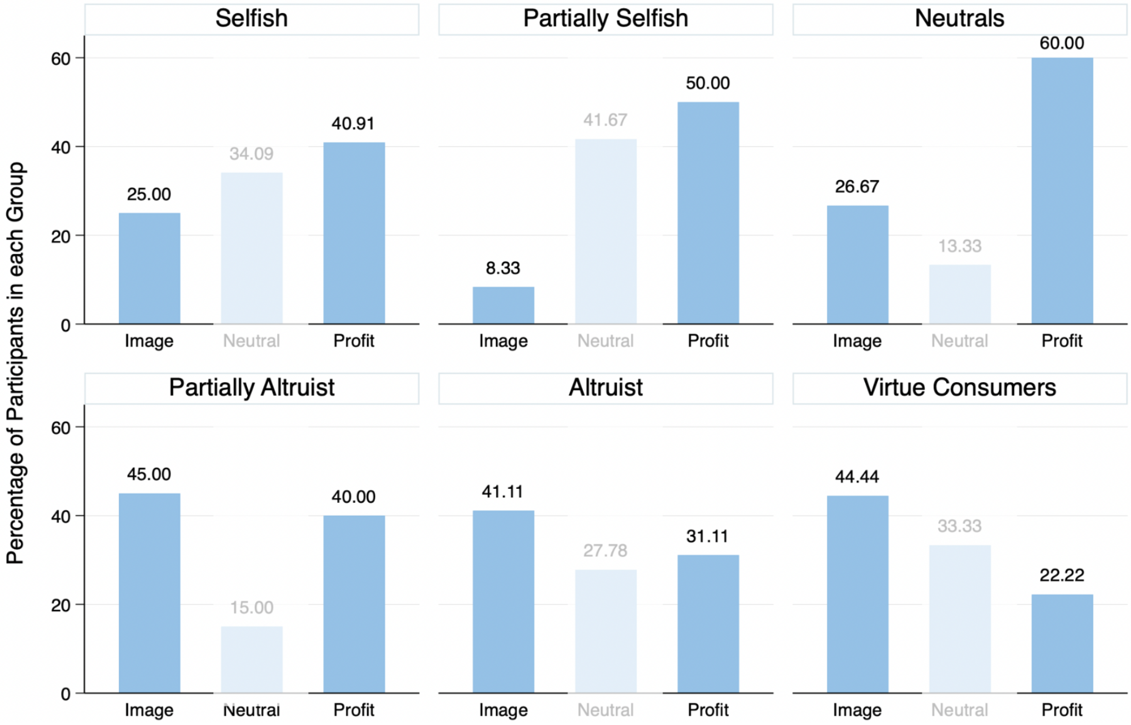


Figure 2.9: Share of Participants in each Category of Self-Deception - 20 Hardest Trials

coexistence of these independent patterns provides significant evidence of the existence of a non-monotonic self-deception.

2.5.2 Regressions and Margins

Before I continue with the main econometric model of this paper, it is useful to study the aggregate patterns. Table 2.4 includes four different aggregate OLS models with robust standard errors and with net bias as the dependent variable. The interpretation of the table is a bit complicated, since the coefficients of each Φ group express the difference with the mean net bias of the selfish group. The altruists in the 40-trial model without gender (model 1) show on average a lower net bias than the selfish group at $p < 0.1$. However, this limited result disappears in the rest of the models. In the 20-trial models, the partially selfish exhibit a higher net bias than the selfish, also at $p < 0.1$.

To get a better understanding of the bias size and direction of each group, we need to examine the predictive margins of the model. The 40-trial model of Table 2.5 shows that the partially selfish group has on average a net bias of 3.9% at a statistical significance level of just below 95%. However, the 20-trial model shows that in the harder portion of the experiment, partially selfish participants displayed an average net bias of 8.6% at p=0.013. Once again, the partially selfish group is the group most consistent with profit-driven self-deceptive behavior. The graphical representation of the the predictive margins can be seen on Figure 2.12.

Four random effects probit regressions were conducted to study self-deception at the individual trial level (Table 2.7). Models 1 and 2 include all 40 trials and models 3 and

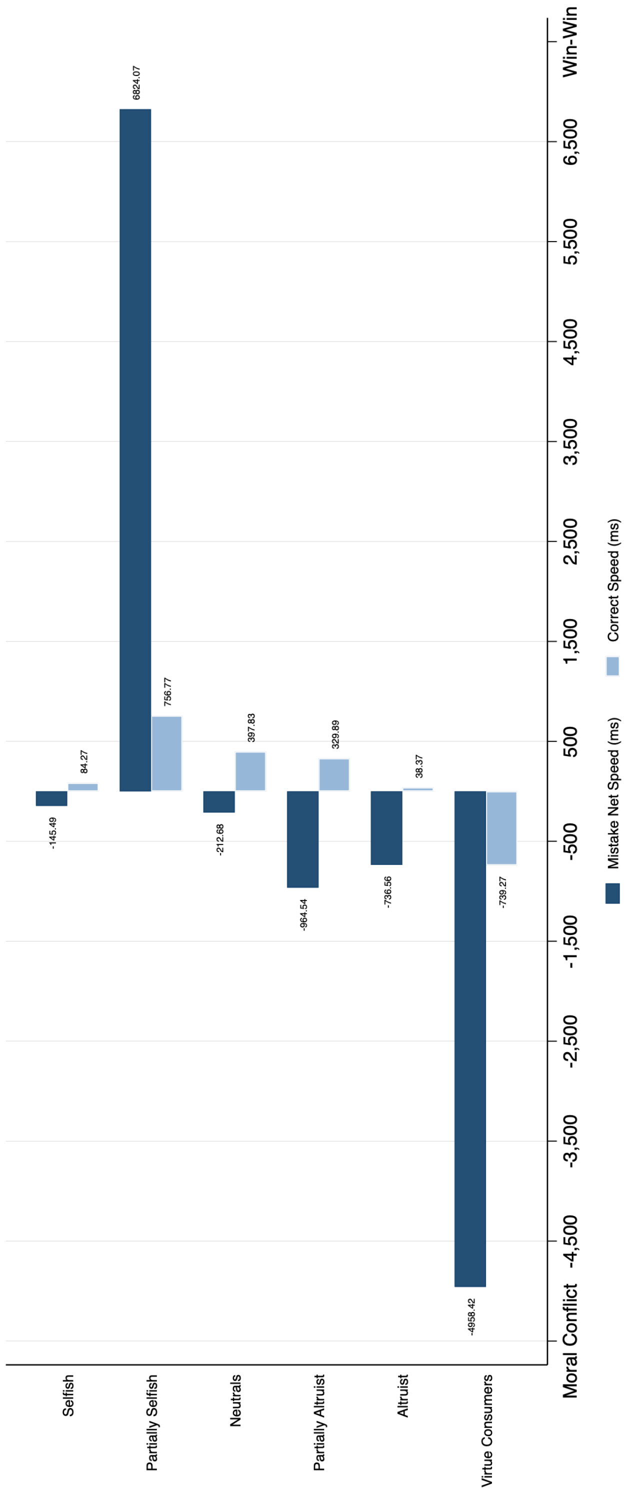


Figure 2.10: Net Speed 20 Trials

4 include only the 20 hardest trials. Models 2 and 4 also include gender and timing as control variables. Most gender coefficients are statistically significant and negative due to the higher error rates of the females in the win-win scenario, which is the baseline of the gender category. The higher bias of female participants towards the morally conflicting scenario reduces the coefficients and increases the standard errors of the altruists and

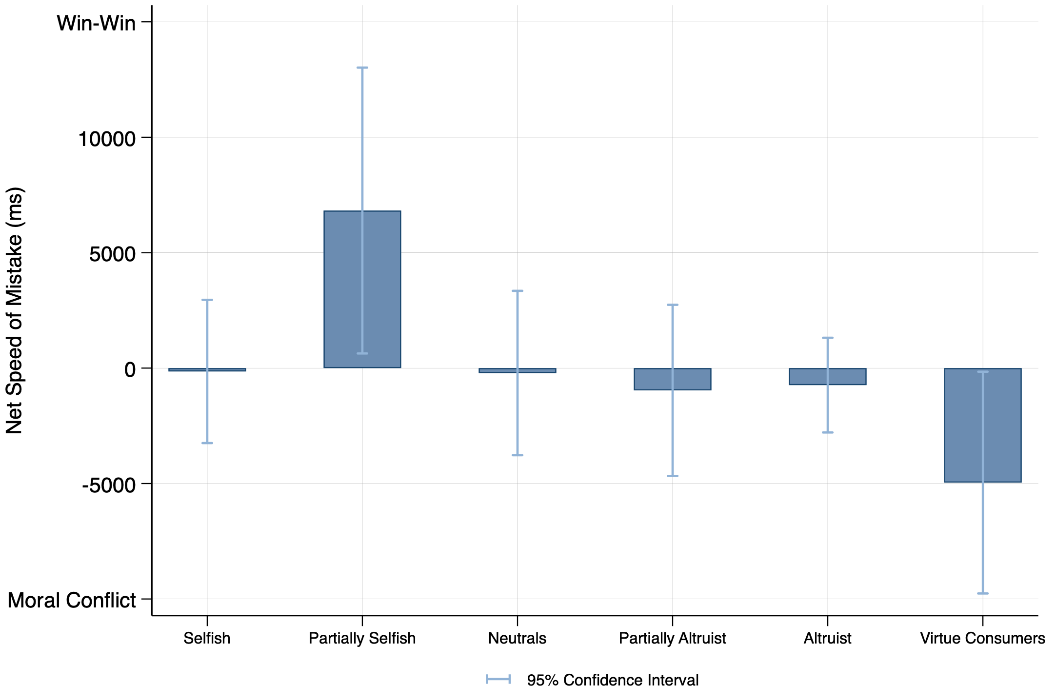


Figure 2.11: Net Speed Mistake (ms) with 95% CI

virtue consumers, implying that their biases towards the moral conflict scenario might be explained by gender and not their self-image concerns. When compared to the baseline error rate of the selfish group in the win-win scenario, the partially selfish consistently show increased mistake rates when faced with the morally conflicting scenario. Their strongest increase is observed in the 20-trial model with a coefficient of 0.694 and a $p < 0.01$. The neutral group also has increased error rates when faced with a moral conflict scenario in models 2 and 3 ($p < 0.1$) and model 4 ($p < 0.05$). Finally, the selfish groups show higher error rates when faced with a moral conflict scenario compared to when they are faced with a win-win scenario at $p < 0.1$)

To better understand the results of the random effects probit model, we need to once again consult the predictive margins. The contrasts of the predictive margins for models 2 and 4 are displayed on Table 2.8. The contrasts refer to the difference in the probability of mistake that each group exhibits between facing a morally conflicting scenario and facing a win-win scenario. Once again, the partially selfish group is the only group with a bias that is significantly different than 0. At the 40-trial model, the bias towards the win-win scenario is of minor statistical significance ($p < 0.1$), while in the 20-trial model the statistical significance increases and passes the 95% threshold. The marginal bias in these two cases almost mirrors the levels observed at the aggregate model's margins in Tables 2.5 and 2.6.

Finally, the linear regression results of Table 2.9 show that the partially selfish group took almost 7 seconds longer than the selfish group to make a mistake when faced with a moral conflict scenario ($p < 0.05$) and, compared to the same baseline, the virtue consumers

Table 2.4: Aggregate Regression Results (OLS)

	(1) 40 Trials	(2) 40 Trials	(3) 20 Trials	(4) 20 Trials
Partially Selfish	0.026	0.030	**0.070***	**0.076***
	(0.02)	(0.02)	(0.04)	(0.04)
Neutrals	0.002	0.005	0.013	0.018
	(0.02)	(0.02)	(0.04)	(0.04)
Partially Altruist	-0.014	-0.013	-0.009	-0.007
	(0.02)	(0.02)	(0.04)	(0.05)
Altruist	**-0.024***	-0.021	-0.036	-0.030
	(0.01)	(0.01)	(0.03)	(0.03)
Virtue Consumers	-0.022	-0.015	-0.036	-0.023
	(0.02)	(0.02)	(0.03)	(0.03)
Gender		✓		✓
Constant	0.011	0.003	0.014	-0.002
	(0.01)	(0.01)	(0.02)	(0.03)
R^2	0.04	0.05	0.04	0.05
N	190	190	190	190

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 2.5: Predictive Margins - Aggregate OLS Model - 40 Trials

	Margin	**S.E.**	**t**	**P-value**	**[95% C.I.]**
Selfish	0.009	0.011	0.790	0.430	-0.014, 0.032
Partially Selfish	**0.039**	**0.020**	**1.900**	**0.059**	**-0.002, 0.079**
Neutrals	0.014	0.014	0.980	0.328	-0.014, 0.042
Partially Altruist	-0.004	0.022	-0.180	0.858	-0.046, 0.039
Altruist	-0.012	0.008	-1.500	0.134	-0.028, 0.004
Virtue Consumers	-0.006	0.017	-0.370	0.711	-0.040, 0.027

took almost 5 seconds more to make a mistake when faced with a win-win scenario. The margins on Table 2.10 confirm the pattern observed in the regression results. On average, the partially selfish took 6.82 seconds longer to make a mistake towards the win-win scenario and the virtue consumers took 4.95 seconds longer to make a mistake towards the morally conflicting scenario. The predictive margins are visualized in Figure 2.14.

2.6 Discussion

The results of the experimental findings of this research are in line with its theoretical framework; nevertheless, the statistical strength of the evidence is weak. There is non-monotonic self-deception, with biases moving in different directions depending on the category of altruistic preferences and self-image concerns that individuals belong to. Despite the need for further experimental research, the behavior of participants in this experiment in terms of bias sign, bias magnitude, and reaction time, points to the existence of a mechanism that triggers two distinct directions of self-deception for participants with different self-image concerns.

Table 2.6: Predictive Margins - Aggregate OLS Model - 20 Trials

	Margin	**S.E.**	**t**	**P-value**	**[95% C.I.]**
Selfish	0.009	0.021	0.450	0.652	-0.032, 0.051
Partially Selfish	**0.086**	**0.034**	**2.510**	**0.013**	**0.018**, **0.153**
Neutrals	0.028	0.032	0.870	0.386	-0.035, 0.091
Partially Altruist	0.002	0.041	0.060	0.952	-0.079, 0.084
Altruist	-0.021	0.016	-1.330	0.187	-0.052, 0.010
Virtue Consumers	-0.014	0.026	-0.510	0.608	-0.065, 0.038

Beyond the aggregate Much of the literature on self-deception attempts to shed light at individual-level biases. Nevertheless, to achieve this goal effectively, we need to find ways to include a measure of self-image in our models and our experimental designs. This experiment was an attempt to do that. The bias of the more selfish groups toward the win-win scenario is of roughly equal magnitude with the bias toward the morally conflicting scenario that the more altruistic groups exhibited. The average net bias across all participants was approximately zero. As a result, the omission of Φ would lead us to believe that there is no pattern of behavior.

The non-monotonic relationship between self-deception and the self-image of altruism will not always result in a near-zero average bias. In different settings, the varying self-image group dynamics could drive the net bias in either direction. Studying these patterns at the aggregate level and without the inclusion of self-image measurements will not be possible.

Profit and Image-Driven Self-Deception When the setting becomes more complex and motivations collide, we can have more than one type of self-deception. In this case, two types of self-deception are identified: profit-driven and image-driven self-deception. The conflict between the two motivations - profit and self-image - exists in both types, but one of the two dominates in each case. Note that the structures of the utility functions remain the same, the magnitude of self-image concerns and altruistic preferences are what cause the change of bias direction.

Moreover, the fact that profit-driven self-deception peaks among the partially selfish and shows a positive trend even among the selfish implies that self-image concerns might begin to play a role at very low levels of altruism. A less statistically consistent result is the image-driven self-deception of the more altruistic groups and implies that some individuals are overestimating the extent of a problem - along with the appropriate personal cost - to increase their self-image. More importantly, in this experiment, the self-deception of those individuals not only decreases their payout but also the payout of the charity. Each time the image self-deceivers mistakenly saw a conflict scenario, they chose the option that gives more to the charity under the conflict scenario; however, in the correct, win-win, scenario the same choice results in a lower payout to the charity.

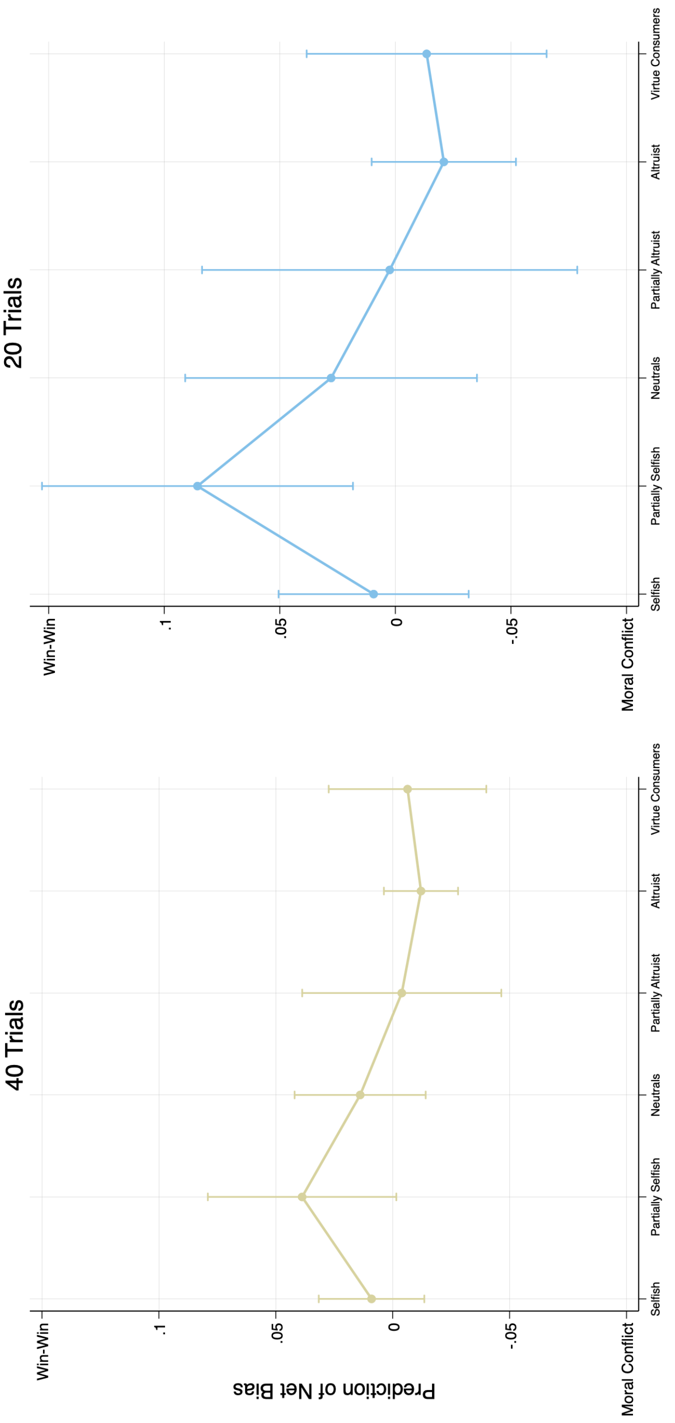


Figure 2.12: Predictive Margins OLS, 40 and 20 Trials

The two directions of self-deception Once we establish the existence of the two self-deception directions, we can see it in a variety of contexts. In this experiment, people who know that they might betray their self-image avoided the conflicting scenario by deploying profit-driven self-deception. However, when altruistic preferences start taking over, bias lowers. After a point, the self-image value of feeling like an altruist makes the diagnostic utility of acting altruistically in the conflict scenario too attractive, resulting

Table 2.7: Random Effects Probit Model, Probability of Mistake

	(1)	(2)	(3)	(4)
	40 Trials	40 Trials	20 Trials	20 Trials
Signal Strength	-0.076***	-0.076***	-0.138***	-0.138***
	(0.01)	(0.01)	(0.01)	(0.01)
Win-Win				
#Partially Selfish	0.108	0.078	0.016	-0.013
	(0.23)	(0.24)	(0.23)	(0.24)
#Neutrals	0.079	0.031	0.062	0.001
	(0.16)	(0.16)	(0.17)	(0.17)
#Partially Altruist	0.025	0.022	-0.152	-0.153
	(0.15)	(0.15)	(0.18)	(0.18)
#Altruist	0.019	-0.019	0.034	-0.008
	(0.11)	(0.11)	(0.12)	(0.12)
#Virtue Consumers	-0.027	-0.090	-0.117	-0.182
	(0.21)	(0.23)	(0.20)	(0.21)
Moral Conflict				
#Selfish	0.129	0.302	0.103	**0.376***
	(0.10)	(0.18)	(0.12)	(0.21)
#Partially Selfish	**0.369***	**0.542****	**0.419****	**0.694*****
	(0.19)	(0.26)	(0.20)	(0.27)
#Neutrals	0.215	**0.373***	**0.268***	**0.517****
	(0.14)	(0.19)	(0.16)	(0.21)
#Partially Altruist	-0.016	0.161	-0.088	0.191
	(0.21)	(0.26)	(0.21)	(0.26)
#Altruist	-0.121	0.050	-0.123	0.144
	(0.11)	(0.18)	(0.12)	(0.20)
#Virtue Consumers	-0.154	0.010	-0.292	-0.023
	(0.32)	(0.38)	(0.27)	(0.33)
Timing		**0.006*****		0.002
		(0.00)		(0.00)
Gender		✓		✓
Constant	**-0.395*****	**-0.450*****	0.148	0.204
	(0.11)	(0.14)	(0.12)	(0.13)
$\ln(\sigma_u^2)$	-1.954***	-1.987***	-2.525***	-2.608***
	(0.19)	(0.20)	(0.32)	(0.36)
N	7600	7600	3800	3800
Groups	190	190	190	190

Standard errors in parentheses * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

in image-driven self-deception among the more altruistic groups. My experimental results also fit the theoretical work performed by Nyborg (2011), which separates consumers into duty-oriented and warm glow consequentialist. The former group avoids information, even at a cost, if they believe it will increase their moral obligations. The latter group seeks information, driven by a desire to enhance their self-image.

Despite being driven by the same type of motivations individuals might behave very differently. For example, profit-driven self-deceivers might ignore the dangers of climate

Moral Conflict vs Win-Win	**P-value**	**Contrast**	**Std. Err.**
40 trials			
Selfish	0.3189	0.011	0.011
Partially Selfish	**0.0966**	**0.034**	**0.020**
Neutrals	0.2155	0.017	0.013
Partially Altruist	0.7856	-0.006	0.022
Altruist	0.1138	-0.012	0.008
Virtue Consumers	0.5794	-0.009	0.016
Joint	0.2307		
20 trials			
Selfish	0.5282	0.013	0.021
Partially Selfish	**0.0198**	**0.079**	**0.034**
Neutrals	0.2186	0.040	0.032
Partially Altruist	0.8696	0.007	0.041
Altruist	0.1391	-0.023	0.015
Virtue Consumers	0.4060	-0.019	0.022
Joint	0.1106		

Table 2.8: Contrasts of Predictive Margins, 40 and 20 Trials

Table 2.9: Net Reaction Time of Mistake, OLS

	Mistake Net Reaction Time
Partially Selfish	**6969.559****
	(3105.182)
Neutrals	-67.191
	(2233.031)
Partially Altruist	-819.054
	(2329.755)
Altruist	-591.074
	(1860.614)
Virtue Consumers	**-4812.933****
	(2472.348)
Constant	-145.488
	(1540.964)
R-squared	0.0536
N	174

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

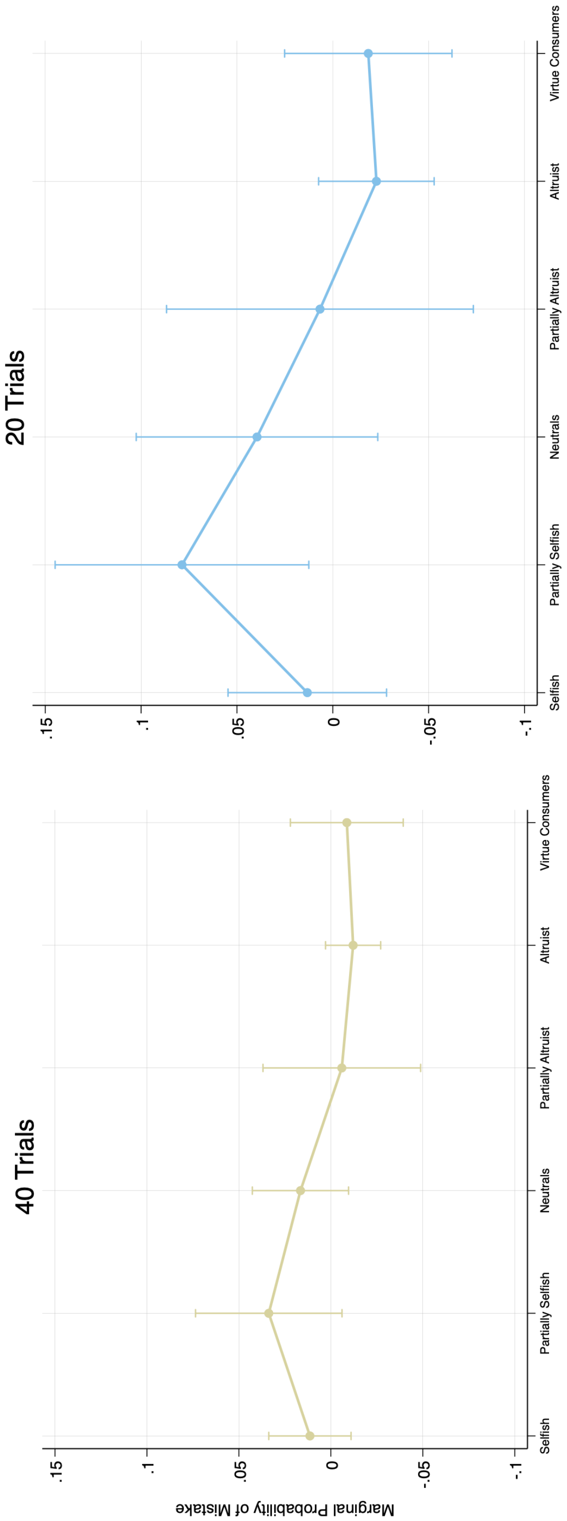


Figure 2.13: Categorical Contrasts 40-20

change, while image-driven self-deceivers might be putting their hopes on individually costly solutions that do not address the problem sufficiently. As employees, profit-driven self-deceivers might ignore their company's environmentally harmful actions and the second group might fall victim to company greenwashing. Both of these problems are caused

Table 2.10: Predictive Margins - Net Mistake Speed - OLS - 20 Trials

	Margin	S.E.	t	P-value	[95% C.I.]
Selfish	-145	1540	-0.090	0.925	-3187, 2896
Partially Selfish	**6824**	**2695**	**2.530**	**0.012**	**1501, 12146**
Neutrals	-212	1616	-0.130	0.895	-3403, 2977
Partially Altruist	-964	1747	-0.550	0.582	-4414, 2485
Altruist	-736	1042	-0.710	0.481	-2795, 1322
Virtue Consumers	**-4958**	**1933**	**-2.560**	**0.011**	**-8775, -1141**

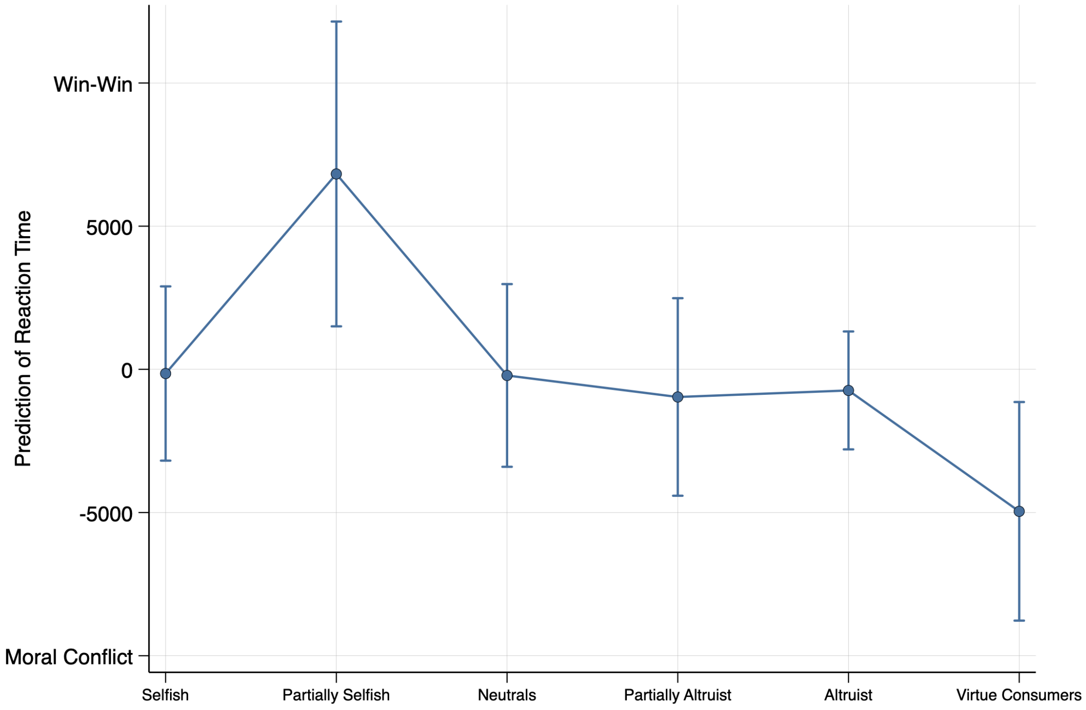


Figure 2.14: Net Speed of Mistake, Margins

by an effort to manipulate one's self-image. Ambiguity and a lack of transparency, in the first case, do not allow us to benefit from the potential - albeit limited - contribution of the profit-driven self-deceivers. In the second case, a group willing to contribute significantly to the cause of climate change is inefficiently utilized. On the bright side, identifying the kind of self-deception individuals deploy can also create opportunities. Profit-driven self-deceivers can be seen as a group that can be convinced with more robust information, while the willingness and resources of the image-driven self-deceivers can be stirred towards more productive solutions.

Political correctness is a topic where the profit and image-driven analysis of self-deception can help us understand some behaviors that limit our ability to communicate and progress. Political correctness is not by itself a negative term. Nobody desires to live in a world of constant offense. The profit-driven self-deception equivalent here is the person who ignores clear opportunities to speak up on important offenses. This person might allow individuals around her (or herself) to be hurt due to their self-deception. This person, of course, does not need to only exist in the sphere of political correctness. For example, a man who stands only to lose socially and professionally, but cares about equality to an extent, might self-deceive away clear injustices against female coworkers in the workspace. On the other hand, many people hold a disdain for political correctness because of extreme behaviors and phenomena such as cancel culture. Overly sensitive political correctness is the definition of image-driven self-deception. The person who gets offended on behalf of some other person in response to a comment that most would deem innocent is an image-driven self-deceiver.

Limitations and future research The inclusion of $\Phi(\alpha, \sigma)$ as a proxy for the self-image of altruism proved far from perfect, but was a good first step toward the measurement of self-image. The overall patterns and the fact that the directions of net bias coincided with the model predictions are encouraging. Nevertheless, future research should come up with approaches to solve the problem of self-image measurement. Clarity in the measurement of self-image should be a priority.

The experimental design of this research had a few drawbacks. Many of the signal levels were too easy, which significantly decreased the statistical power of the results. In the future, a more efficient approach would be to skip signal levels and repeat the high error levels instead. Ambiguity is very important when it comes to self-deception and this experimental design did not include enough of it. If, similarly to the present research, an implicit measurement of self-image is needed, the stronger signal trials can be included but the trials in between need not be.

2.7 Conclusion

This research provides novel insights into the complex relationship between self-deception, altruism, and self-image concerns. The experimental findings highlight the existence of two distinct types of self-deception: profit-driven and image-driven. Profit-driven self-deception is characterized by individuals underestimating altruistic opportunities to protect their material self-interest, while image-driven self-deception involves overestimating scenarios that require ethical judgment to enhance one's self-image.

This paper was an attempt to quantify self-image, but the measurement used is only a proxy; therefore, future research on the topic must improve on the measurement of self-image. Exploring more complex and motivational conflicting settings and testing its findings on real-world applications should be at the forefront of self-deception research.

The non-monotonic relationship between self-deception and self-image of altruism suggests that individuals with different levels of altruistic preferences and self-image concerns exhibit opposing biases in their decision-making processes. This divergence underscores the importance of considering individual differences in self-image when studying self-deception.

One significant implication of these findings is that self-deception is not an effortless process. The reaction time patterns observed in the experiment indicate that participants engaged in a self-convincing process, which required time and cognitive effort. If self-deception requires effort, it might be easier to find tools to combat it.

Future research should focus on refining the measurement of self-image and exploring more complex, motivationally conflicting settings. Understanding the nuances of self-deception can have practical applications in various domains, including health behaviors, conflict of interest scenarios, and prosocial behavior. By identifying the types of self-deception individuals are prone to, interventions can be tailored to address specific biases, leading to more effective strategies for promoting honesty, ethical behavior, and better decision-making.

In conclusion, this study advances our understanding of self-deception by highlighting the role of self-image concerns and the non-monotonic nature of the relationship between self-deception and altruism. The experimental design and findings contribute to the broader literature on belief-based utility and motivated cognition, offering a novel framework for future research.

Appendix A

A.1 Rejection due to Utility Violations

Between Session 2 and Session 3, the experiment moved from the experimental laboratory at Universitat Pompeu Fabra (BESLAB) to the one at the University of Barcelona. Due to an omission in the recruitment process, a few participants with poor English language skills were included in Session 3. As a result, the utility violations were significantly higher than the rest of the groups (Table 2.11). A utility violation is defined as an action that reduced both the participant's own payoff and the one of the charity with the knowledge of the participant.

Session No.	**1**	**2**	**3**	**4**
2	-0.004104	.	.	.
	(1.000)	.	.	.
3	0.031213	0.035317	.	.
	(0.002)	**(0.003)**	.	.
4	0.0055	0.009604	-0.025713	.
	(1.000)	(1.000)	**(0.073)**	.
5	-0.006226	-0.002122	-0.03744	-0.011726
	(1.000)	(1.000)	**(0.005)**	(1.000)

P-values in parentheses

Table 2.11: Comparison of Violations by Session

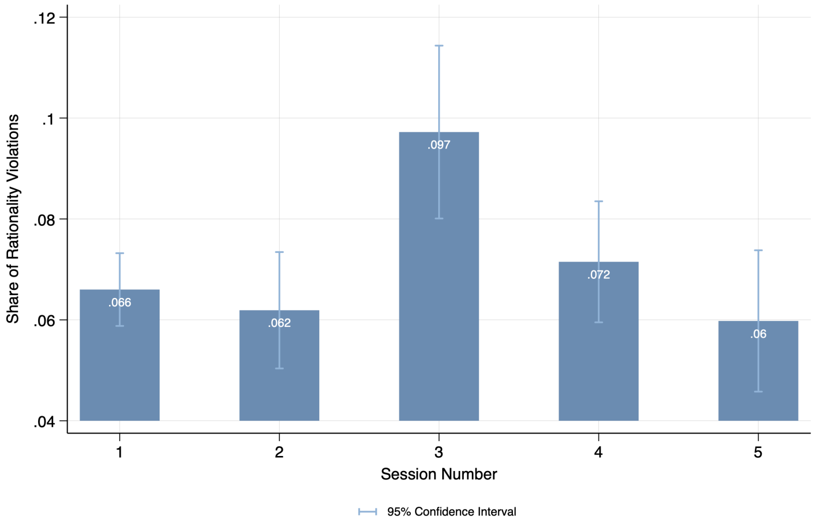


Figure 2.15: Violations By Session

A.2 Tables

Table 2.12: Mistakes by Gender, number and percentage

	Correct	**Mistake**	**Total**
Female	4,760 (93.55%)	328 (6.45%)	5,088 (100.00%)
Male	3,621 (94.30%)	219 (5.70%)	3,840 (100.00%)
Other	175 (91.15%)	17 (8.85%)	192 (100.00%)
Total	8,556 (93.82%)	564 (6.18%)	9,120 (100.00%)

Table 2.13: Actions by Gender, number and percentage

	A	**B**	**Total**
Female	3,258 (64.03%)	1,830 (35.97%)	5,088 (100.00%)
Male	2,738 (71.30%)	1,102 (28.70%)	3,840 (100.00%)
Other	136 (70.83%)	56 (29.17%)	192 (100.00%)
Total	6,132 (67.24%)	2,988 (32.76%)	9,120 (100.00%)

Table 2.14: Field of Study by Gender

	Female	Male	Other	Total
Communication	1			1
Economics & Business	63	58		121
Engineering	3	8	1	12
History	1			1
Natural Sciences	15	6		21
Other	16	4	1	21
Social Sciences	7	4	2	13
Total	106	80	4	190

A.3 Figures

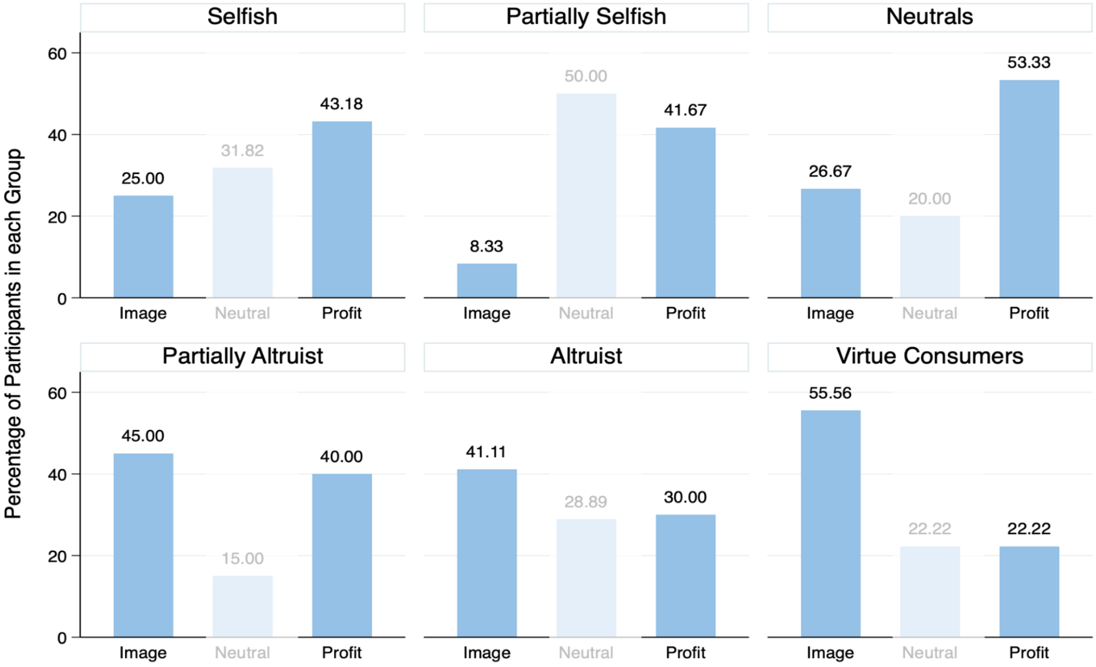


Figure 2.16: Share of Participants in each Category of Self-Deception

3 Project Choice and Social Image Concerns

With Claudia Cerrone[1], *Alessandro De Chiara*[2], *and Ester Manna*[3][4]

We propose and experimentally test the idea that social-image concerns may create conflicts of interest in a principal-agent relationship. We design an experiment where a principal may delegate the choice of a risky project to an agent. It would be desirable if the agent refrained from choosing a project when uninformed. The agent knows which project should be undertaken only if his/her relative performance in an IQ test was high. Yet, they can blame luck if the project does not pan out. Social-image concerns arise if the principal only observes the project chosen by the agent and not also their relative performance in the test. We find that as long as the interaction is anonymous social-image concerns are inconsequential. By contrast, they do matter when the relationship is not anonymous as the principal observes the agent's picture. In that case, uninformed agents are significantly more likely to choose a risky project and we highlight how this effect is driven by male participants.

3.1 Introduction

There now exists abundant experimental evidence that individuals care about their social image, namely, what other people think of them, and this affects their behavior (e.g., see the influential work by Andreoni and Bernheim, 2009). Such evidence mostly focuses on pro-social behavior, that is, individuals care about being viewed as altruistic. Recent papers have also shown that individuals are reluctant to be perceived as needy or low skilled with potentially broad economic and social implications (see Friedrichsen et al., 2018, and Valdez Gonzalez et al., 2023).

However, it remains an important understudied question how the individual's willingness to be perceived as "better" then they actually are affect their behavior in relevant economic interactions. For instance, within organizations when individuals interact with their bosses or colleagues, or between organizations when they interact with customers or clients. To provide an example, consider a worker's decision of accepting tasks or responsibilities

[1] City, University of London, Northampton Square, London EC1V 0HB, United Kingdom. E-mail: Claudia.Cerrone@city.ac.uk.

[2] Department of Economics, Universitat de Barcelona and Barcelona Economic Analysis Team (BEAT), Avinguda Diagonal 696, 08034, Barcelona, Spain. E-mail: aledechiara@ub.edu.

[3] Professora Agregada Serra Húnter, Department of Economics, Universitat de Barcelona and Barcelona Economic Analysis Team (BEAT), Avinguda Diagonal 696, 08034, Barcelona, Spain. E-mail: estermanna@ub.edu.

[4] We would like to thank Christiane Schwieren and Natalia Valdez Gonzalez for helpful comments.

even when they are unsure as to their ability to deliver on what agreed upon.[5] Of course, this may happen because employees are unable to say no to a hierarchical superior or because they want to try their luck if successful completion of a task generates current or future monetary benefits (e.g., improve the worker's reputation and therefore may lead to promotions or bonuses). Another reason may be that workers accept to undertake difficult projects to boost their social image, that is, to improve the opinion their superiors or colleagues have of them.

The research question we ask in this paper is indeed whether social image, in itself, can create conflicts of interests even though, monetarily, there are none. We do so by devising a novel laboratory experiment that mimics the principal-agent interaction arising when it comes to delegating the choice of a project. The general idea is that accepting to undertake some specific risky project may allow a worker to signal their superior ability, skills, or talent. However, there might be negative economic repercussions if low-ability individuals undertake more difficult projects for mere signaling reasons. Anticipating this, a principal may even refrain to delegate the choice of more complicated projects. The experiment is designed to shut down channels other than social image that may account for the selection of more difficult tasks, like reputation-building or the inability to say no to a superior.

Our paper belongs to the economic analysis of organizations that look beyond the analysis of mere incentive schemes (e.g. see Section 7 of Gibbons et al., 2013), which, among other things, has emphasized the role of an organization's mission (e.g., see Besley and Ghatak, 2005) and of the employee's identity as an organization insider (Akerlof and Kranton, 2005) for performance. We extend this strand of the literature by theoretically positing that worker's social image concerns may affect their behavior and this may ultimately impact the success of an organization.[6] We also design a novel laboratory experiment to test the model's prediction.

We first develop a theoretical framework that provides the basis for our empirical predictions. In the model, a principal needs to decide whether to delegate the choice of a project to a potentially better informed agent. There is no material conflict of interests between the principal and the agent. However, as the choice of the project may reveal information about the agent's type, the weight an agent attaches to his social-image may affect the agent's behavior and the benefits of the economic interaction. The model suggests that uninformed agents may choose projects even though it would be better if they refrained from doing so in order to maximize their own and social material payoff. The

[5] Even CEOs may accept task when they feel unprepared and seek little help on how to do the job. In a survey of 402 CEOs from 11 countries, the leadership advisory firm Egon Zehnder found that more than two thirds of CEOs acknowledged that they were not fully prepared for the job and only half of them turned to senior management for advice. See *"Survey: 68% of CEOs Admit They Weren't Fully Prepared for the Job"*, published online on Harvard Business Review on July 20, 2018.

[6] In this respect, a noteworthy departure from Bénabou and Tirole (2006) is our focus on an agent's concern for what a principal may think of their skills and talent rather than their prosocial motivation and how this may affect the principal's delegation choice.

model also predicts that this choice distortion is magnified when the signal on the agent's type is more salient.

We test these predictions in a laboratory experiment where 300 participants randomly play the role of either principals or agents. Like in the model, principals must decide whether to delegate the choice of a project (one of three boxes in the experiment) to the agent or select an outside option that gives both parties a small sure payoff. Without knowing his principal's delegation decision, an agent can choose one of the boxes or select the same outside option. One box contains some money for both the principal and the agent with some positive probability, whereas the other boxes are always empty. The payoffs are chosen in such a way that a risk-neutral agent who does not have social-image concerns will always choose a box if he knows which one may contain the money and always select the outside option, otherwise. However, an agent is privately provided this information only if he scored in the top 50% among all agents in the experimental session in a previously conducted IQ test. We have conducted three treatments. In the control treatment, the principal knows whether her agent's performance is in the top 50%, whereas in the baseline treatment she does not know this information. In the photo treatment, the principal does not know the agent's relative performance and is shown their picture. Thus, social-image concerns are not relevant in the control treatment, since the project choice is not needed to convey information about the agent's performance. Social-image concerns may matter in the baseline treatment and, even more, in the photo treatment as we increase the salience of the signal by removing the anonymity of the interaction.

We find that the fraction of uninformed agents who choose the project is essentially identical in the control and baseline treatment (47% and 48%, respectively), whereas it is higher in the photo treatment (64%). The difference is not only statistically significant, but also economically meaningful. Notably, the increase can be entirely ascribed to the different behavior of male participants in the non-anonymous and the anonymous treatments.[7] Thus, our contribution is threefold. First, we theoretically and experimentally highlight how social-image concerns can create conflicts of interest in meaningful economic interactions, entailing substantial economic losses. Second, we show how social-image concerns are decisively more relevant in non-anonymous interactions. Third, we document remarkable gender differences in the concern for one's social image.

The rest of the paper proceeds as follows. In the next section, we review the related literature. In Section 3.3, we develop the theoretical framework. In Section 3.4, we describe the experimental design. In Section 3.5, we present the results of our analysis, which we discuss in Section 3.6. We provide concluding remarks in Section 3.7.

[7] When we only consider male participants, the fraction of uninformed agents who choose the project goes from 32% and 40% in the control and baseline treatment, respectively, to 71% in the photo treatment.

3.2 Related Literature

Our paper belongs to the literature studying the implication of social-image concerns for economic agents' behavior. In this regard, Friedrichsen and Engelmann (2013) conducted an experiment to investigate the interaction between intrinsic motivation and social approval in ethical consumption. They found that participants were more willing to pay a premium for Fairtrade chocolate when their choice could be observed by others. This behavior was driven by social approval seeking, as individuals wanted to signal their ethical consumption to others. The effect was more pronounced among participants who were not intrinsically motivated to buy Fairtrade products, highlighting the influence of social image concerns on ethical behavior.

Our paper is more closely related to those papers highlighting the negative effects of social image. In particular, Friedrichsen et al. (2018) used a laboratory experiment to show that social image concerns reduce the take-up of welfare benefits. Participants avoided claiming benefits to evade the stigma associated with being perceived as low-skilled or dependent on others. This avoidance highlights how social image concerns can lead to economically sub-optimal decisions, as individuals prioritize maintaining a positive social image over receiving beneficial transfers. Differently from previous papers, our work is the first to point out possibly negative consequences of social image in a principal-agent setting, with notable organizational implications.

Our paper is closely related to Katok and Siemsen (2011) who conducted an experiment to study the impact of reputation concerns on task choice. They found that with reputation concerns, both highly capable and less capable agents chose more difficult tasks, leading to reduced success rates (83% for highly capable, 44% for less capable). When reputation concerns were removed, success rates improved significantly (89% for highly capable agents and 58% for less capable agents). However, the experiment did not only eliminate reputation concerns but also social image concerns altogether. In addition, the treatment that eliminated reputation concerns also eliminated any incentive for either the highly or the less capable agents to choose a hard task. As a result, the increase in success rates can not be attributed exclusively to a lack of reputation or social image concerns, since other factors, including risk preferences and social image concerns, were also removed.

Our findings also contribute to our understanding of gender differences pertaining to social-image concerns and their implications. Murad et al. (2019) explore the impact of different incentives (competitive, social-value, and social image) on task performance and examine gender differences in response to these incentives. They found that competitive and social image incentives significantly improve performance, with competitive incentives having a more pronounced effect on males, who performed better than females under competitive conditions. No significant gender difference was observed in the social image incentive group, indicating equal benefit from public recognition for both genders. Females outperformed males in the social-value incentive group, highlighting their responsiveness to altruistic motivations. These findings suggest that visibility and competition are powerful motivators, with competitive incentives being particularly effective for males, while social image incentives work well across genders.[8]

Lastly, although they do not directly refer to social image, Burks et al. (2013) and Schwardmann and Van der Weele (2019) explain the usefulness of biased belief formation in persuading others. In evolutionary terms, social image concerns might have risen as a way to force the individual to employ these strategies and convince others of their skills or value. Burks et al. (2013) conducted experiments to examine overconfidence as a social signaling bias. Participants performed cognitive tasks and estimated their relative performance, knowing that these estimates would be used to persuade others for monetary rewards. The study found that participants inflated their performance estimates to appear more competent, driven by the potential financial gain from successful persuasion. This strategic behavior underscores the role of social signaling in workplace overconfidence. The study rejected self-image concerns and Bayesian updating as primary drivers, emphasizing that overconfidence was mainly induced by the desire to send positive signals to others. Schwardmann and Van der Weele (2019) tested the strategic self-deception hypothesis, finding that people self-deceive into higher confidence to be more persuasive. Their experiments showed that overconfidence increased when individuals could profit from persuading others, supporting the theory that overconfidence serves an adaptive social function. This behavior was driven by the opportunity to gain financially from successful persuasion.

3.3 Theoretical Model

In this section we develop the theoretical model that will provide the foundations for the hypotheses on the effect of social image concerns that we will test in the laboratory experiment. We consider an environment in which a principal (she) must decide whether to delegate the choice of a risky project to an agent (he) or select a safe outside option. The probability that the project is successful depends on the agent's ability. The agent privately observes his ability and decides whether to choose a project or select the outside option. We enrich this standard delegation model by allowing for social image concerns. Specifically, we assume that the agent may care about being viewed as of high ability by the principal.[9] Real-world applications abound: from investors who may ask their financial advisors for investment strategies to CEOs who may delegate the choice of revamping a moderately successful product line to a company division. We assume that the players are both risk neutral and self-interested, but the agent is concerned about what the principal thinks of him.

The agent can be of two types, $\theta \in \{L, H\}$, and we denote by $\gamma \in (0,1)$ the probability that the agent is of type H. The agent knows his type and chooses an action $x \in \{a, \emptyset\}$.

[8] We provide a more in-depth discussion of gender differences in social-image concerns in Section 3.6.

[9] We defer a discussion of the relation with other models that include social-image concerns or deal with delegations of authority till the end of this section.

Action a is risky in that it stochastically affects an outcome $y \in \{S, F\}$ and it holds that

$$Pr[y = S|\theta = H] = p > q = Pr[y = S|\theta = L],$$

with $p < 1$ and $q > 0$. Action $\emptyset$ is safe in that it generates a type-independent outcome $y = R$ for sure.

Let $m(x, \theta)$ be the expected monetary payoff to a type-θ agent who chooses action x. It holds that:

$$m(a, H) > m(\emptyset, H) = m(\emptyset, L) > m(a, L).$$

In particular, we assume that if $y = S$ (respectively, $y = R$) the principal receives s_P (r_P) and the agent receives s_A (r_A). If $y = F$ both the principal and the agent receive a payoff normalized to 0. As a result, the above assumption on the agent's expected monetary payoffs implies that

$$ps_A > r_A > qs_A > 0. \tag{3.1}$$

For the principal we assume the following inequalities:

$$ps_P > r_P > qs_P > 0. \tag{3.2}$$

The principal does not know the agent's type, but she observes which outcome has realized. She can form a belief about the agent being a high type, which we denote $\tau(y)$ as it depends on the outcome $y \in \{S, F, R\}$. As mentioned above, in our model, the agent cares about being perceived as a high-type by the principal. This component of the agent's utility, which we will at times refer to as the *agent's reputational payoff*, is increasing in the likelihood that the principal thinks he is a high type $\tau(y)$ and the sensitivity to such social image concerns is measured by the parameter η.[10] The expected utility of agent θ with sensitivity η can be written as:

$$U_{\theta,\eta}(x) = m(x(\theta, \eta), \theta) + \eta \cdot \tau(y(x(\theta, \eta))).$$

We assume that η is known to be distributed according to a continuous distribution function $G(\cdot)$ with density $g(\cdot)$ on $[0, \infty)$.

Timing of the model. The sequence of events is as follows.

1. Nature draws the agent's type.

2. The principal decides whether to delegate the choice of the project to the agent ($d = 1$) or select the outside option ($d = 0$).

3. The agent observes his type and, if $d = 1$, decides which action x to take.

[10] For simplicity, we assume that being perceived as a low-type does not carry any disutility. What ultimately matters is the difference between being perceived as a high ability agent as compared to a low ability one. Therefore, to simplify computations, we normalize the value attached to being a low-type to 0.

3.3.1 Equilibrium Analysis

We look for the Bayesian Equilibria of the game, which are characterized by (i) the agent's project choice x as a function of his ability (θ) and social-image concerns (η); (ii) the principal's belief about the agent's type, which is a function of the outcome, $\tau(y)$; (iii) the principal's delegation choice d, which is a function of the expected agent's ability and social-image concerns. Being rational, the agent will correctly anticipate the principal's belief about his type, which is formed according to Bayes' rule.

We begin by briefly considering the scenario with no social-image concerns.

Remark 1. *If $\eta = 0$, the agent's action is $x(H) = a$, $x(L) = \emptyset$, and the principal's belief is $\tau(S) = \tau(F) = 1$ and $\tau(R) = 0$. The principal always delegates, i.e., $d = 1$.*

Proof. See that $a = \arg\max_x U_{H,0}(x)$ and $\emptyset = \arg\max_x U_{L,0}(x)$ because of Assumption (3.1). The action and its resulting outcome y fully separate types and the principal prefers to delegate because of Assumption (3.2) as $\gamma p s_P + (1-\gamma) r_P > r_P$. □

If the agent does not attach any value to be viewed as a high type by the principal, he will choose $x(H) = a$ and $x(L) = \emptyset$. Therefore, $\tau(S) = \tau(F) = 1$ and $\tau(R) = 0$. In words, choosing a project would always reveal that the agent is a high type, irrespective of the outcome, and choosing the outside option would always reveal that the agent is a low type. Thus, looking only at the material payoffs, there is no conflict of interests between the principal and the agent. Hence, when there are no social-image concerns, the principal always delegates because she can trust the agent will always take the optimal action.

We now study the more general case where η can be positive. To make the problem interesting we assume that if all L-type agents choose $x = a$, the principal would rather select the outside option than delegate this choice to the agent:

$$\gamma p s_P + (1-\gamma) q s_P < r_P. \tag{3.3}$$

The first result we can prove is that high-type agents never want to opt for the outside option: by selecting the risky action they can expect to get a higher monetary reward and may better signal their ability.

Lemma 1. *H-type agents always choose $x = a$.*

Proof. Suppose by contradiction that there exist an H-type agent with $\eta' > 0$ who is indifferent between $x = a$ and $x = \emptyset$. For such agent, the following must hold:

$$r_A + \eta' \cdot \tau(R) = p s_A + \eta' \cdot [p\tau(S) + (1-p)\tau(F)].$$

Since $p s_A > r_A$ by Assumption (3.1), it must be that $\eta' \cdot [p\tau(S) + (1-p)\tau(F)] < \eta' \cdot \tau(R)$, which implies that all H-type agents with $\eta > \eta'$ will have a strict preference for $a = \emptyset$ and only $G(\eta')$ H-type agents will choose $x = a$. However, because $y = S$ (respectively, $y = F$) is more likely when $\theta = H$ (resp., when $\theta = L$), the above equation also implies that $\tau(R) > q\tau(S) + (1-q)\tau(F)]$ and since $r_A > q s_A$ because of Assumption (3.1), all L-type agents would choose $x = \emptyset$. But then, if an H-type agent chooses $x = a$, he would reveal that is of high-type, i.e., $\tau(S) = \tau(F) = 1$ and the above equation cannot hold. □

As a result, social-image concerns do not alter the behavior of H-type agents who continue to select the risky task. However, there cannot be a fully-separating equilibrium because a low-type agent who cares enough about his social image would rather deviate to be viewed as a high-type.

Lemma 2. *There cannot be a fully-separating equilibrium where all agents with $\theta = H$ choose $x = a$ and all agents with $\theta = L$ choose $x = \emptyset$.*

Proof. Suppose by contradiction that there exists such a fully-separating equilibrium. If so, $Pr[\theta = H|y \in \{S, F\}] = 1$. But then low-type agents who are sufficiently concerned about their social image, that is, for whom $\eta > r_A - qs_A$ would prefer to deviate and choose $x = a$. Then $Pr[\theta = H|y \in \{S, F\}] < 1$. □

In the next proposition, we characterize the cutoff value of η above which an L-type agent chooses the risky action.

Proposition 1. *There exists a threshold value $\hat{\eta} > 0$, such that all L-type agents with $\eta > \hat{\eta}$ choose $x = a$ and all L-type agents with $\eta < \hat{\eta}$ choose $x = \emptyset$.*

Proof. First notice that choosing $a = \emptyset$ reveals that $\theta = L$ because of Lemma 1. Then, the cutoff value of η, denoted $\hat{\eta}$, for which an L-type agent is indifferent between $x = a$ and $x = \emptyset$ satisfies:

$$r_A = qs_A + \hat{\eta} \cdot [q\tau(S) + (1 - q)\tau(F)],$$

where

$$\begin{aligned}\tau(S) =& Pr[\theta = H|y = S] = \frac{Pr[y = S|\theta = H]Pr[\theta = H]}{Pr[y = S|\theta = H]Pr[\theta = H] + Pr[y = S|\theta = L]Pr[\theta = L]}\\ =& \frac{p\gamma}{p\gamma + q[1 - G(\hat{\eta})](1 - \gamma)};\end{aligned}$$

and

$$\begin{aligned}\tau(F) =& Pr[\theta = H|y = F] = \frac{Pr[y = F|\theta = H]Pr[\theta = H]}{Pr[y = F|\theta = H]Pr[\theta = H] + Pr[y = F|\theta = L]Pr[\theta = L]}\\ =& \frac{(1 - p)\gamma}{(1 - p)\gamma + (1 - q)[1 - G(\hat{\eta})](1 - \gamma)}.\end{aligned}$$

The indifference condition can be rewritten as:

$$r_A - qs_A = \hat{\eta}[q\tau(S) + (1 - q)\tau(F)].$$

Note that, the LHS is independent of η, whereas the RHS is strictly increasing in η, which means that there exists at most one admissible value of $\hat{\eta}$ that satisfies the above equality. That is, all L-type agents with $\eta \in [0, \hat{\eta})$ strictly prefer $x = \emptyset$ to $x = a$ and all L-type agents with $\eta > \hat{\eta}$ strictly prefer $x = a$ to $x = \emptyset$.

It is easy to see that $\hat{\eta}$ is increasing with r_A and decreasing with s_A. It is also decreasing with γ. The intuition is that the observation of $y = S$ or $y = F$ is more likely to come from an H-type agent. To see this, let

$$H := r_A - qs_A - \hat{\eta}[q\tau(S) + (1 - q)\tau(F)] = 0,$$

and, by applying the implicit function theorem,

$$\frac{\partial \hat{\eta}}{\partial \gamma} = -\frac{\frac{\partial H}{\partial \gamma}}{\frac{\partial H}{\partial \hat{\eta}}} = -\frac{-\hat{\eta}\left[q\frac{\partial \tau(S)}{\partial \gamma} + (1-q)\frac{\partial \tau(F)}{\partial \gamma}\right]}{-[q\tau(S) + (1-q)\tau(F)] - \hat{\eta}\left[q\frac{\partial \tau(S)}{\partial \hat{\eta}} + (1-q)\frac{\partial \tau(F)}{\partial \hat{\eta}}\right]} < 0$$

because $\frac{\partial \tau(S)}{\partial \hat{\eta}} > 0$, $\frac{\partial \tau(F)}{\partial \hat{\eta}} > 0$, $\frac{\partial \tau(S)}{\partial \gamma} > 0$, $\frac{\partial \tau(F)}{\partial \gamma} > 0$.

The effect of q and p is instead ambiguous. □

L-type agents with strong enough social-image concerns will be willing to suboptimally choose a risky action.

We now proceed to investigate the principal's choice. As, in equilibrium, she correctly predicts the behavior of the different types of agents, the principal is more likely to delegate authority over the choice of the project to the agent when there are relatively more H than L types (i.e., γ is higher), when the proportion of low types who choose the safe option is larger (i.e., $\hat{\eta}$ is higher), when her personal gain from a successful implementation of the task, $s_P - r_P$ is greater. The following proposition formalizes the principal's delegation decision under the assumption that, if indifferent, she prefers to delegate authority.

Proposition 2. *The principal chooses $d = 1$ if $\hat{\eta} \geq \bar{\eta}$ and $d = 0$ otherwise, where $\bar{\eta}$ is determined from the following equation:*

$$G(\bar{\eta}) = 1 - \left(\frac{\gamma}{1-\gamma}\right)\left(\frac{ps_P - r}{r - qs_P}\right). \tag{3.4}$$

Proof. Suppose that the principal believes that all H-type agents and a proportion $1-G(\bar{\eta})$ of the L-type agents choose $x = a$. Then, she would be indifferent between choosing $d = 1$ and $d = 0$ if

$$\gamma ps_P + (1-\gamma)\Big[[1-G(\bar{\eta})]qs_P + G(\bar{\eta})r_P\Big] = r_P.$$

The above can be rewritten as:

$$\gamma(ps_P - r_P) = (1-\gamma)\Big[r_P - G(\bar{\eta})r_P - [1-G(\bar{\eta})]qs_P\Big],$$

and rearranging we obtain (3.4). As the principal rationally anticipates that $G(\hat{\eta})$ L-type agents choose $x = \emptyset$, the statement of the proposition follows. Lastly, see that $\bar{\eta}$ decreases when γ increases. □

Plausibly, an agent's concern for social image depends on the inferred characteristic or the observer's identity. For example, agents may be more invested in managing perceptions of their cognitive abilities when observed by superiors compared to manual skills. By the same token, agents may care more about the judgment of people with whom they have a close relationship than strangers they may never see again. To take this into account, let us modify the agent's utility by assuming that the agent's reputational payoff depends on a parameter $V \geq 0$, which stands for visibility. That is, an agent's expected utility is:

$$U_{\theta,\eta}(x) = m(x(\theta,\eta),\theta) + \eta \cdot \tau(y(x(\theta,\eta)))V.$$

The next proposition studies the effect of visibility on the agents' and the principal's choice.

Proposition 3. *An increase in visibility V induces more L-type agents to choose $x = a$ and, accordingly, increases the likelihood that a principal chooses $d = 0$*

Proof. First, observe that the choice of H-type agents is unaffected by the parameter V. It is easy to see that $\frac{\partial \hat{\eta}}{\partial V} < 0$ and, therefore, it becomes more likely that $\hat{\eta} < \bar{\eta}$ when V increases. □

We predict that when the reputational payoff is more important, because the project choice more saliently reveals an ability the agent cares about, we should observe more risky actions undertaken by the agents and less delegation.

Relation to the Literature. Our model of delegation of authority is partly inspired by Aghion and Tirole's influential paper (Aghion and Tirole, 1997). There are some noteworthy differences, though, as the principal cannot implement the task, and the agent's ability (e.g., the agent's knowledge of the projects that generate positive payoffs in Aghion and Tirole, 1997) is exogenously given. More prominently, absent social-image concerns, the preferences of the principal and the agent are aligned. In our setting, social image concerns may create a conflict of interests between the players as the project choice may convey information about some agent's characteristic he may deeply care about. Akin to many standard delegation models, we also assume that the principal cannot commit to contingent transfers (e.g., see Alonso and Matouschek, 2008).

Image concerns have also been widely studied in the economics literature (e.g., see Bénabou and Tirole, 2006, and Ellingsen and Johannesson, 2008). In this respect, our model is closer to Khalmetski and Sliwka (2019) where agents do not want to be viewed as liars. However, unlike us, they study individual decision makers who also incur a lying cost and their choice set is not necessarily binary.

3.4 Experimental Design

The experiment uses a between-subject design and has three parts. Participants are provided with specific instructions for each part sequentially.[11]

Part 1. In the first part of the experiment, each participant is randomly assigned to the role of agent or to the role of principal, neutrally termed as "role A" and "role B", respectively. Participants are randomly matched in groups of five members: one principal and four agents.[12] This was illustrated to participants using Figure 3.1. Every agent is given up to 10 minutes to complete a 30-question IQ test. Principals do not take the test but can see it.

[11] **Click here to see this experiment's instructions.**

[12] We made this design choice as we are primarily interested in the behavior of a subgroup of the agents.

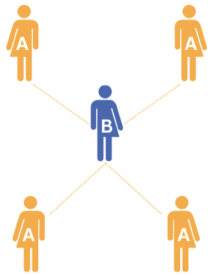


Figure 3.1: Group

Part 2. In the second and core part of the experiment, every group is faced with three boxes: one blue, one green, and one red. One of these boxes contains some money: €12 for the agent and €10 for the principal. The other two boxes are empty. When an empty box is chosen, both the principal and the agent earn €0. When the box containing the money is chosen, with a 75% probability the box can be opened and will pay €12 to the agent and €10 to the principal, and with a 25% the box cannot be opened and will leave agents and principals with €0. If no box is chosen (outside option), the principal earns €6 and each of the four agents earns €4. Which box contains the money may vary across agents in a group. A principal's choice affects all agents she is linked to (henceforth referred to as "her agents"), whereas an agent's choice may only affect their principal.

Initially, no participant knows which of the three boxes contains the money. They only know that each box has the same probability of containing the money. Later in the experiment, depending on their relative performance in the test, the agents might find out which box contains the money, as further explained below. Principals never find out which box contains the money.

After receiving instructions and answering some review questions, participants go through the three following stages.

Stage 1: Each principal must decide between choosing the outside option and allowing her agents to make a decision. If a principal delegates the decision, one of her four agents will be selected at random and their decision will determine the principal's payoff.

Stage 2: Each agent learns whether they scored in the top 50% in the test among the participants in the session. If they did, they immediately find out which box contains the money. If they did not, they do not find out. Potential ties are broken by using the agents' completion time in the test.

Stage 3: Without knowing the decision made by their principal in Stage 1, each agent must decide whether they want to choose the outside option or a box. Note that an agent's decision will only be implemented if their principal delegated decision making authority to them in Stage 1. If their principal choose the outside option, their agents' decisions will have no consequences.

Part 3. At the end of the experiment, we ask agents whether they want to know their percentage of correct answers in the test and their ranking relative to other participants.

We also ask all participants some final questions to elicit their attitudes towards risk (Charness and Gneezy, 2010), their social preferences (Bartling et al., 2009), and their demographics.

Treatments. In our first treatment ("Baseline"), the principal does not find out whether her agents scored in the top 50% in the test. We use this as a baseline to determine whether some agents choose the box even if uninformed, which could be driven by their desire to signal high ability to the principal, i.e., by their social image concerns.

Hypothesis 1. *A positive fraction of uninformed agents will choose a box.*

In order to understand whether uninformed agents choose a box because of social image concerns, we use two different approaches/treatments. In our second treatment ("Principal learns"), we inform the principals about their agents' relative performance in the test. This effectively removes the possibility that uninformed agents choose a project to signal their ability to the principal. That is, social image concerns should no longer affect the agents' decisions. Our hypothesis is the following. If uninformed agents choose a box because of social image concerns, then the fraction of uninformed agents choosing a box should be lower when the principal is informed about her agents' performance than in the baseline where the principal remains uninformed.[13]

Hypothesis 2. *The fraction of uninformed agents who choose a project will be significantly lower when the principal is informed about their agents' test score than when the principal remains uninformed.*

In our third treatment ("Photo"), the principal remains uninformed as in the baseline, and we remove the agents' anonymity by showing each agent's photo to their principal. Our hypothesis is that the removal of anonymity will increase the agents' social image concerns, and thus the fraction of uninformed agents choosing a box. This is in line with the result of Proposition 3 that an increase in visibility induces more type-L agents to select the risky action a.

Hypothesis 3. *The fraction of uninformed agents who choose a project will be significantly higher when the agents' photos are shown to their principal than when interactions are fully anonymous.*

Table 3.1 summarises our treatments.

Implementation. The experimental sessions were programmed in o-Tree and run at the lab of the University of Barcelona in 2024. In total, 300 subjects participated in the experiment: 240 were assigned to the role of agent ("role A") and 60 to the role of principal ("role B"). The average duration was about 35 minutes. The average earnings were €11, including a show-up payment of €5. The experiment was preregistered on AsPredicted.

[13] If principals anticipate this effect on the agents, they might delegate decision making authority more when they are informed about their agents' performance than when they are not.

Table 3.1: Treatments

Treatment	Principal learns agents' performance	Agents' identity is anonymous	# subjects
Baseline	No	Yes	75
Principal learns	Yes	Yes	100
Photo	No	No	125

3.5 Experimental Results

In this section we report our experiment's results. Our first goal is to verify whether some of the agents who scored in the bottom 50% in the test (and thus remained uninformed about which box contained the money) choose a box nevertheless. Our second, and main goal is to determine whether this behaviour is at least partly driven by social image concerns.

Table 3.2: Frequency of agents choosing a box by treatment

Treatment	Agent is uninformed	Agent is informed	Total
Baseline	.48	1	.74
	(40)	(40)	(80)
Principal learns	.47	.93	.7
	(30)	(30)	(60)
Photo	.64	.9	.77
	(50)	(50)	(100)

Number of observations in parentheses.

Table 3.2 illustrates the agents' frequency of choosing a box over the outside option by treatment. In the "Baseline", where the principal does not learn whether the agents scored above the 50% in the test, nearly half of the uninformed agents (48%) choose a box anyways. This supports Hypothesis 1. All the informed agents choose a box, which reassures us that participants understood the task and responded to incentives properly.

In the "Principal learns" treatment, where uninformed agents can no longer signal their ability to the principal by choosing a box, we still observe that nearly half of them (47%) choose a box. The fraction of uninformed agents choosing a box is nearly identical, regardless of whether the principal learns about their performance in the test. This suggests that the uninformed agents' decision to choose a box observed in the "Baseline" is not driven by social image concerns. Hence, we find no evidence supporting Hypothesis 2. A potential reason why social image concerns do not seem to play a role in these two treatments is that the agents' identities remained fully anonymous during the experiment.

In the treatment "Photo", anonymity is lifted: each agent's photo is shown to the principal they are linked to. The fraction of uninformed agents choosing a box rises to 64%, which is significantly higher than the fraction of uninformed agents choosing a box in the non-anonymous treatments (one-tailed t-test, $p = 0.034$). This supports Hypothesis 3.

Table 3.3: Frequency of uninformed agents choosing a box by treatment and gender

Treatment	Male	Female	Total
Baseline	.32	.67	.48
	(22)	(18)	(40)
Principal learns	.4	.6	.47
	(20)	(10)	(30)
Photo	.71	.58	.64
	(24)	(26)	(50)

Number of observations in parentheses.

Next, we examine whether uninformed agents' project choice behaviour varies by gender. Table 3.3 illustrates uninformed agents' frequency of choosing the box for males and females.[14] We observe strong gender differences. The fraction of uninformed males who choose a box in the "Photo" treatment (71%) is significantly higher than the fraction of uninformed males who choose a box in the anonymous treatments (one-tailed t-test, $p = 0.0028$). In contrast, there is no significant treatment effect for females. We also observe that, while more uninformed males than females choose a box in the "Photo" treatment, the opposite is true in the "Baseline".

To further investigate the comparison between the "Baseline" and the "Principal learns" treatment, as well as the comparison between the "Photo" treatment and the anonymous treatments, we use logit regressions. Table 3.4 reports the marginal effects of informing the principal about their agents' performance in the test on uninformed agents' probability of choosing a project. As in this table we compare the first two treatments, the total number of uninformed agents is 70. A description of the variables follows. The variable *Agent chooses box* takes value 1 if the uninformed agent chooses a box and value 0 if the agent picks the outside option. The treatment variable *Principal learns* takes value 1 if the principal learns whether each of their agents scored above 50% in the test and 0 if the principal does not learn. Given the gender effects shown in Table 3.3, we report the marginal effects of informing the principal for female and male agents separately. The variable *Prosocial* takes value 1 if a participant is both weakly and strongly prosocial. A participant is weakly prosocial if, when faced with two allocations where they earn the same amount, they prefer the allocation where their partner earns more. A participant is strongly prosocial if they prefer to equally split a pie with their partner than to have a bigger share of the pie than the other. The variable *Investment* measures the amount that participants choose to invest in a risky venture. Participants who are not risk averse should invest all their endowment. The variable *# People known* denotes the number of people present in the room that a participant reports they know. Finally, the variable *See*

[14] In the male category, there is also one uninformed agent who classified themselves as "Other". Our results do not change if we include this observation in the female category or if we drop it.

score takes value 1 if an agent decides to learn their score in the test at the end of the experiment and 0 otherwise.

All our specifications show the marginal effects of informing the principal on uninformed agents' likelihood of choosing a project instead of the outside option. Specification (1) controls for the treatment variable, *female*, and their interaction. Specification (2) additionally controls for *Investment*. Specification (3) additionally controls for *Prosocial*. Specification (4) also controls for *# People known* and *See score*. As in the tests, we do not observe a treatment effect. Informing the principal about their agents' performance in the test does not affect uninformed agents' project choice.

Table 3.4: Impact of informing the principal on uninformed agents' project choice

DV: Agent chooses box				
	(1)	(2)	(3)	(4)
Principal learns				
if female=0	0.082	0.090	0.073	0.039
	(0.149)	(0.150)	(0.147)	(0.155)
if female=1	-0.067	-0.081	-0.056	-0.113
	(0.192)	(0.190)	(0.195)	(0.191)
Investment		Yes	Yes	Yes
Prosocial			Yes	Yes
# People known				Yes
See score				Yes
N_I	70	70	70	70

*** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$
Logit regression. The table reports the marginal effects of informing the principal about their agents' performance in the test on uninformed agents' probability of choosing a box. Standard errors are in parentheses. N_I denotes the number of uninformed agents when we compare between the "Baseline" and the "Principal learns" treatment.

Table 3.5 reports the marginal effects of showing each agent's photo to their principal on uninformed agents' probability of choosing a project. The treatment variable *Photo* takes value 1 if the principal can see the photo of each of their agents ("Photo" treatment) and 0 if no photo is shown and thus there is full anonymity ("Baseline" and "Principal learns" treatments). Note that, while there is a difference between the two anonymous treatments (whether the principal learns about their agents' performance or not), it is reasonable to pool their data into a joint "Anonymous" treatment as we observed no difference in behaviour between the two.[15] The total number of uninformed agents is 120. The other

[15] Recall that, as shown in Table 3.2, the fraction of uninformed agents choosing a box is nearly identical

variables are the same as in Table 3.4. We find that the showing the agents' photos to their principal significantly increases the fraction of uninformed male agents choosing a box. For females, there is no significant change. The result is robust to different specifications.[16]

Table 3.5: Impact of showing agents' photo on uninformed agents' project choice

DV: Agent chooses box	(1)	(2)	(3)	(4)
Photo				
if female=0	0.351***	0.372***	0.371***	0.376***
	(0.119)	(0.115)	(0.117)	(0.119)
if female=1	-0.066	-0.064	-0.070	-0.039
	(0.133)	(0.125)	(0.120)	(0.119)
Investment		Yes	Yes	Yes
Prosocial			Yes	Yes
# People known				Yes
See score				Yes
N_I	120	120	120	120

*** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$

Logit regression. The table reports the marginal effects of showing each agent's photo to their principal on uninformed agents' probability of choosing a box. Standard errors are in parentheses. N_I denotes the number of uninformed agents when we compare between the "Photo" treatment and the anonymous treatments.

3.6 Discussion

In this section, we discuss the findings of the paper and we provide some further information about the results of the experiment.

Choose the box to feel like a winner. In the control and baseline treatment, nearly half of the uninformed agents choose a box. In addition to risk-loving attitudes, a possible explanation for why a large fraction of participants opens the box without information is provided by Köszegi (2006) who presents a model where individuals derive "ego utility" from maintaining a positive self-image. This leads to overconfidence and misaligned task choices. In his model, individuals might prefer more challenging tasks to uphold their ego, even when these tasks are not materially beneficial. His paper discusses how overconfidence

in the two anonymous treatments.

[16] In Table 3.7 in the Appendix, we report the marginal effects of showing each agent's photo to their principal on uninformed agents' probability of choosing a project without distinguishing between males and females. The result is still statically significant.

can be detrimental to employee motivation and performance, as workers might choose tasks that enhance their self-image rather than those that match their actual skills. In a similar vein, in our model a subordinate may choose a suboptimal project to improve their social image. However, if simply opening the correct box makes you feel like a winner, it can add another layer as to why someone might open a box without having private information.

Over-confidence, competition, and gender. Our result that male participants tend to care more about being perceived as high-skilled than their female counterparts, with relevant payoff consequences, may be related to their over-confidence. Other papers have found that men are more influenced by social-image considerations. For instance, Ewers and Zimmermann (2015) explore how image utility affects the reporting of private information about skills and abilities. They develop a model and test it through a controlled lab experiment. Participants answered quiz questions and reported their performance either privately or publicly. The results showed that public reports were significantly inflated compared to private ones, suggesting that social approval seeking leads to overconfident self-assessments. Notably, men were more likely to inflate their performance in the presence of an audience than women, indicating a stronger response to social image concerns. Also our results suggest that male participants display social-image concerns in a non-anonymous setting. However, there are mixed results when it comes to gender on the agent' side. Murad et al. (2019) find increased performance for both genders due to social image concerns and recently Haeckl (2022) explored gender differences in self-assessments influenced by social image concerns. In a laboratory experiment, women increased their self-assessments when these were made public, but only if the actual performance remained private. This suggests that women are particularly sensitive to social image in public self-assessments. However, this increased confidence did not translate into increased effort or performance, indicating that social-image concerns impact self-perception but not necessarily task performance.

We can also contemplate a positive relationship between social-image concerns and competition to explain why men are more concerned about social image than women. Several papers have suggested that men are more responsive to competition than women (see among others Gneezy et al., 2003, Gneezy and Rustichini, 2004, and Backus et al., 2023). These papers have also shown that, being more competitive, men perform better than women in different settings, even when there are no differences in their abilities. In our paper, uninformed men undertake more often the risky project to signal their talent to the principal, and this decision entails substantial economic losses.

Delegation decision. While our paper's focus is on the agent's project choice, it is interesting to also look at the principal's delegation decision. Table 3.6 reports the frequency of delegation by treatment and gender. We observe that female principals delegate less than male principals in the first two treatments. The difference across treatments remains non significant. Interestingly, female principals delegate significantly more in the "Photo" treatment than in the anonymous treatments. This could be because some female principals may display social image concerns: they may care about being perceived as willing to trust the agents, i.e., trustful. If so, their utility might also include a reputational payoff and they might be willing to delegate authority even if they expect that it is likely that low-ability agents will betray their trust and select $x = a$. The underlying idea being that they prefer to be seen as trustful, or even gullible, than cynical. This willingness to exhibit trust may be especially strong when image concerns are more prominent, i.e., V is higher in our "Photo" treatment, and may trigger reciprocity considerations.[17] For male principals there is no significant difference across treatments, even we observe a reduction in the delegation choice in the "Photo" treatment with respect to the anonymous ones: some of them anticipates that uninformed agents choose the box more often in the "Photo" treatment.

Table 3.6: Frequency of delegation by treatment and gender

Treatment	Male	Female	Total
Baseline	.818	.556	.7
	(11)	(9)	(20)
Principal learns	.857	.625	.733
	(7)	(8)	(15)
Photo	.75	1	.92
	(7)	(17)	(25)

Number of observations in parentheses.

Finally, we informally discuss the effect of players' attitudes towards risk and altruism on the principals' delegation decision.

The impact of risk-aversion on the principal's delegation choice is ambiguous. The intuition is the following. If agents are risk averse, they are less willing to take the risky action. The presence of social image concerns, especially when there is more visibility, may induce some risk-averse H-type agents to choose the risky task, even though their expected material payoff would be maximized by $x = \emptyset$. Risk-averse L-type agents would be less willing to choose $x = a$ for two reasons: first, they dislike the material risk this task entails; second, the choice of $x = \emptyset$ does not unequivocally reveal that they are low-ability types. Risk-averse principals would see smaller benefits from delegating but anticipating that the agents would also be more cautious in their choices might actually delegate more. On the contrary, risk-loving attitudes would induce a greater proportion of L-type agents to choose $x = a$ and would not affect the fact that all H-type agents choose the risky task. A risk-loving attitude may encourage principals to delegate more, even though they may also anticipate that low-ability risk-loving agents will be inclined to choose $x = a$ more often. The correlation coefficient between delegation and investment is positive, but

[17] Note that, if anything, reciprocity considerations may encourage agents to choose the task that maximizes the material payoff. This could also explain why women who act as agents do not change their behavior across treatments.

pretty small.

As for the principal's altruism, if the principal cares about the material payoff of the agent, she may actually delegate less often: the principal anticipates that many low-ability agents will end up mimicking the high-ability ones choosing an action that leads to a lower material payoff. The correlation coefficient between delegation and our variable of altruism (prosocial) is equal to -0.2638.[18]

3.7 Conclusions

Social-image concerns may affect individual choices. Our paper theoretically and experimentally shows that economic agents may be willing to take sub-optimal actions in order to boost their social image even if this may hurt other individuals, with whom they are linked. This result is obtained when the interaction is non-anonymous, which makes the agent's choice more salient for his social-image. Importantly, we provide evidence that male individuals are those whose behavior is more affected by social-image concerns.

Future research may attempt to dig deeper into the novel findings of our research. It may be interesting to see if stigma drives the uninformed agents' strengthened inclination to choose a project in the photo treatment. Studying the interplay between over-confidence and social-image concerns can also be a valuable research venue to better understand why male agents have shown this predisposition to potentially forego money in order to improve their social image.

[18] For the agent, if he also cares about the principal's material payoff, he will be less willing to choose $x = a$ when $\theta = L$. Yet, the change appears to be only quantitative. The correlation coefficient between choosing the box when uninformed and prosocial is equal to -0.1658.

Appendix B

B.1 Tables

Table 3.7: Impact of showing agents' photo on uninformed agents' project choice

DV: Agent chooses box	(1)	(2)	(3)	(4)
Photo	0.169*	0.178**	0.173**	0.175**
	(0.091)	(0.088)	(0.087)	(0.087)
Investment		Yes	Yes	Yes
Prosocial			Yes	Yes
# People known				Yes
See score				Yes
N_I	120	120	120	120

*** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$

Logit regression. The table reports the marginal effects of showing each agent's photo to their principal on uninformed agents' probability of choosing a box. Standard errors are in parentheses. N_I denotes the number of uninformed agents when we compare between the "Photo" treatment and the anonymous treatments.

4 High Hopes or Hard Truths?

Expectations, Merit Beliefs, and Redistribution Preferences

Individuals tend to accept inequality they view as meritocratic but resist it when they attribute it to luck. Winners in a distribution process often overemphasize the role of merit, a behavior frequently attributed to self-serving bias. This experimental research attempts to understand the persistence of the winner's bias through self-image maintenance. To quantify the role of beliefs in this process, I include expectations about one's future income. Expectations are an indicator of one's prior beliefs, therefore, the discrepancy between expected and achieved income can reveal the intensity of someone's self-image gains or losses. The experimental results indicate that, when income is a noisy signal of one's ability, participants whose income expectations are not met may protect their self-image by attributing outcomes to luck, which in turn leads them to advocate for greater redistribution. Additionally, participants redistributed along ideological lines when their expectations were met, but the effect of ideology disappeared when they were not. Lastly, beliefs about personal merit tended to overshadow beliefs about group merit in preferences for redistribution.[1]

4.1 Introduction

Beliefs about whether income is driven by merit or luck consistently shape redistribution preferences. Societies that attribute income to effort generally favor lower redistribution (Alesina and Ferrara, 2005) and these beliefs strongly affect redistribution support even when accounting for self-interest (Fong, 2001). Of course, not all luck is equal—losing a home to a hurricane differs fundamentally from losing it in a poker game. Experimental findings illustrate that when inequalities stem from externally imposed luck, support for redistribution increases (Cappelen et al., 2013; Akbaş et al., 2019; Mollerstrom et al., 2015). Thus, when poverty's situational causes are highlighted, support for egalitarian policies rises (Koo et al., 2023). In a nutshell, meritocratic individuals view inequality more favorably when it is the result of merit and less favorably when it is the outcome of imposed luck.

The position individuals find themselves in on the income distribution significantly affects their assessments of the process' meritocracy. Those finding themselves higher on the income distribution tend to attribute their success to merit (Di Tella et al., 2007;

[1] I am deeply grateful to Alessandro De Chiara, Ester Manna, Alexandrina Stoyanova and Joël van der Weele for their invaluable comments and suggestions. Their insights significantly contributed to the development of this paper.

Deffains et al., 2016), while those at the bottom blame luck (Valero, 2022), and other factors, such as the selfishness of the rich (Almås et al., 2022). This pattern is reflected on the lower support for redistribution found among those with higher incomes (Cohn et al., 2022; Konow, 2000) and the evidence that when individuals are informed that their income position is lower than previously thought, they tend to increase their support for redistribution (Cruces et al., 2013). In summary, meritocratic individuals accept inequalities due to merit and individuals become more meritocratic when they belong to higher income groups, resulting in the winner's bias, which is expressed through a lower support for redistribution.

Recent experimental evidence has shed light on the winner's bias. Deffains et al. (2016) exogenously assigns participants, who are blind to their condition, to either a hard or an easy income-generating task designed to create a successful and an unsuccessful group. They find that even in the absence of monetary incentives, successful participants chose to redistribute less in a subsequent round. While successful participants credited their effort and focus for their success, unsuccessful participants blamed factors such as task clarity for their failure. Interestingly, successful participants rated the tasks difficulty higher than unsuccessful ones, despite being assigned to the easy task. This dynamic affected individuals across the ideological spectrum, low-earners on its right end decreased their meritocratic beliefs and increased their support for redistribution.

Other similar binary success experimental designs provide complementary evidence. Fehr and Vollmann (2022) find that high earners overestimate effort's role in their success, leading to lower support for redistributive taxes. This pattern appears across political lines, with both liberal and conservative high earners linking their success to merit. When given the opportunity to pay a small amount to learn whether they were assigned to the hard or the easy task, half of the participants decided not to take it. Amasino et al. (2023) account for personal and social norms, finding that beliefs regarding the merit of one's own performance remain a significant factor affecting preferences for redistribution, but concluding that a shift towards libertarian and meritocratic norms—among those with a higher income—might be the most important component. Finally, both Valero (2022) and Dorin et al. (2021) provide additional experimental evidence that high earners often attribute income to effort and low income to luck despite a lack of monetary incentives for themselves.

My experimental research attempts to advance the literature on redistribution preferences by clarifying the dynamics behind the winner's bias and differs from previous experimental attempts to understand the topic in several key areas.

Binary Splits. My experimental design does not rely on an exogenously imposed binary split of success and failure (Deffains et al., 2016; Fehr and Vollmann, 2022; Valero, 2022). Instead, it includes a noisy signal that applies equally to everyone. Thanks to this element, the setting resembles societal dynamics more closely, since no participant's income is impervious to noise. Namely, the signals that participants receive in this experiment are noisy but remain informative. This differentiation also allows me to study the winner's bias in a continuous, instead of binary, setting.

Expectations. Unlike Amasino et al. (2023), I am not using moral and social norms as mediating factors of the winner's bias. Instead, I focus on the potential intensity of self-image concerns by incorporating the income expectations of participants. Moreover, the divergence between expectations and achieved income allows me to differentiate between the biased winners (or losers) specifically in terms of self-image concerns.

Imposed Confidence. In contrast to previous research (Fehr and Vollmann, 2022; Hansson and Sund, 2023), beliefs such as confidence are not exogenously imposed. In my design, the measured confidence is a more salient belief of the individual, not a positive or negative shock imposed on them by the experimenter. This distinction is crucial when dealing with self-image issues, as deeply held beliefs are stronger than imposed ones.

Merit Ratings. The design incentivizes participants to report their merit beliefs indirectly, without explicitly asking them to do so. This is a significant deviation from existing approaches, since, to my best knowledge, merit beliefs are generally measured through direct participant reports.

Two Levels of Merit Beliefs Another goal of this research is to understand whether there is a difference between image-maintaining beliefs that may arise when individuals think about the group compared to when they think about themselves. For example, someone might believe that a task is meritocratic due to overwhelming evidence or ideology, yet still feel personally unlucky and believe they deserved better. The design explicitly distinguishes between the individual's beliefs about merit at the group and personal levels.

Bayesianism. Finally, the winner's bias is ultimately a matter of belief updating. This experiment accounts for the extent of each participant's Bayesianism, that is, their Bayesian reasoning skills in the absence of other motivations. When someone is generally bad at the internal statistics of Bayesian updating, they might behave differently than someone more proficient. Cognitive biases like the endowment effect or loss aversion typically decrease among more sophisticated individuals (Frederick, 2005). However, ideologically motivated biases are often more pronounced in individuals with higher intelligence (Kahan, 2013; Kahan et al., 2017). On the other hand, motivated thinking requires a directionality in bias (Bénabou and Tirole, 2016), therefore, I will also be distinguishing between negative and positive deviations from the perfect Bayesian behavior. To analyze these relationships, I propose and use a novel weighting procedure that—to the best of my knowledge—has not been utilized before. This addition allows me to further account for mediating factors concerning the winner's bias produced by self-image maintenance.

The experimental results suggest that when income serves as a relatively unreliable indicator of one's ability, participants whose income falls short of their expectations may protect their self-image by attributing the outcome to luck, which subsequently increases their support for redistribution.

The remainder of the paper is organized as follows: Section 4.2 describes the experimental design; Section 4.3 includes the theoretical specification of the experimental setup;

Section 4.4 provides the hypotheses and explains their logic; Section 4.5 presents the results; Section 4.6 discusses the findings and provides context; and Section 4.7 summarizes and concludes the paper.

4.2 Design

4.2.1 Treatments

This experiment consists of a control and two treatments. The only difference between the two treatments is the type and amount of noise added in the test-scoring process. More specifically, in the *Profitable Noise* treatment, the noise added is slightly higher but gives an advantage to the participants by assigning them higher incomes. In the *Neutral Noise* treatment, the noise in the signal is slightly lower and the resulting participant incomes are similar to what they would be in the absence of noise, as is the case in the control.[2]

Part 1 - Prior Beliefs Participants are informed that they will need to complete a 15-question Raven's progressive matrices test in 4 minutes. They are shown a 5-question sample to become familiar with the test. They are incentivized to report how many questions they believe they will answer correctly. Their future performance estimation becomes the maximum amount they can receive in a bonus payment they might collect at the end of the experiment. If they reach or exceed their prediction, their bonus payment equals their prediction multiplied by 20 Euro cents. If they do not reach or exceed their prediction, the bonus payment becomes a random number between 0€ and 1€. Therefore, if a participant predicts they will answer 12 answers correctly and they manage to reach or exceed that score, they receive 2.4€ if they are selected for the bonus payment. If they do not answer at least 12 correct answers, they will receive a random amount between 0€ and 1€.

Afterwards, participants are informed of the test's grading system. A correct answer gives 1 point with a 70% probability and 0 points with a 30% probability. In the Profitable Noise treatment, an incorrect answer gives 1 point with a 35% probability and 0 points with a 65% probability. The sole difference between Profitable and Neutral noise treatments is that in the latter an incorrect answer returns a point with a 15% probability instead of 35%. This change reduces the total noise of income as a signal, but disadvantages participants with more incorrect answers.

Following the explanation of the probability system, participants are given the opportunity to provide another prediction based on the probabilistic point system. One participant and one of the two predictions are randomly chosen for payment at the end of each session of the experiment. The initial prediction serves as our baseline prior belief; however, the second prediction is also informative. The difference between the two predictions provides evidence of the strength of the prior beliefs and the extent to which the participant is a Bayesian. For example, a participant who expected to answer all questions correctly

[2] **Click here to see this experiment's instructions.**

initially should now expect to only receive 70% of the points; a participant who expected to answer no questions correctly should now expect to receive 35% of the points. Failure to properly update between the two predictions implies that the individual exhibits flawed updating even in the absence of distracting motivations.

Finally, before proceeding to take the test, participants are informed that their test payment will be proportional to their score, with a maximum payment of 2.50€.

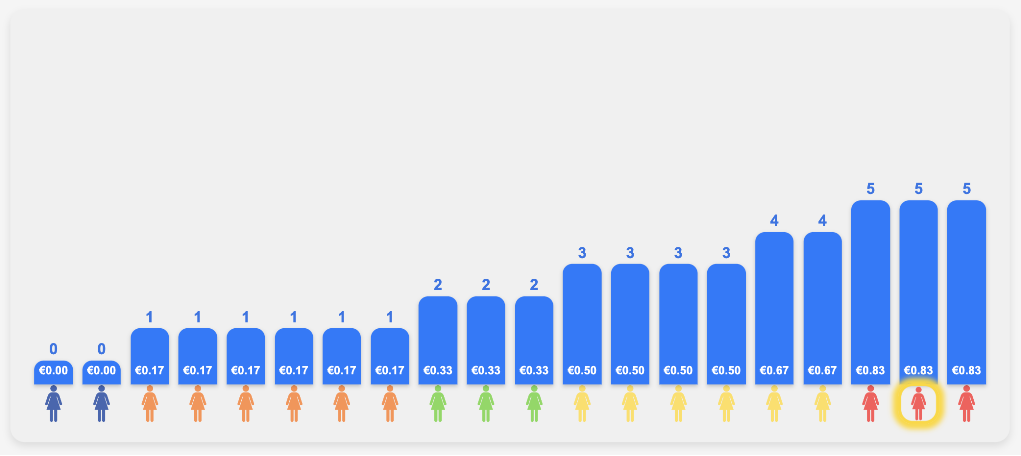
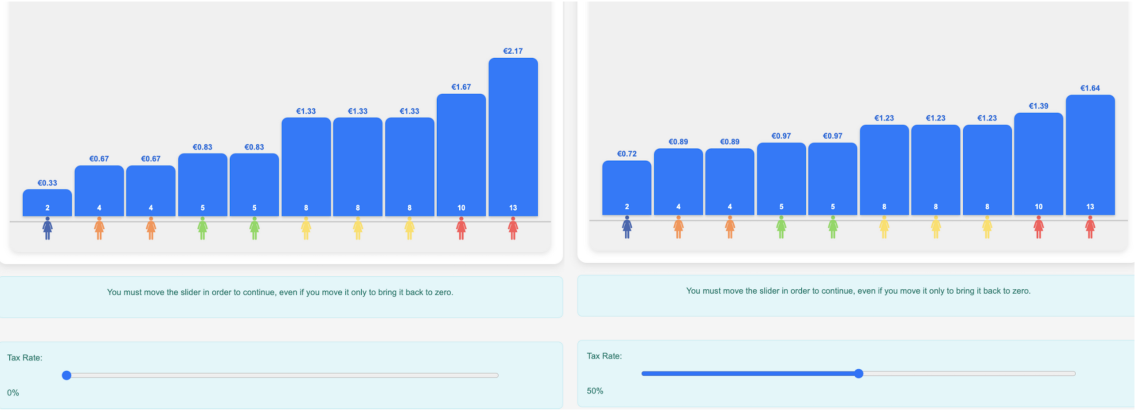


Figure 4.1: Overall Session Distribution (Example)

Part 2 - Redistribution and Merit Assessments Participants are first informed of their percentage score, their ranking among all participants, and their corresponding payment. They are then shown a payment distribution graph (Figure 4.1) displaying their income distribution and their position in it. On the next screen, participants are randomly split into Group A and Group B. Using a slider, the participants of Group A can vote on a redistribution rate for Group B, and vice versa (Figure 4.2). It is made clear to participants that their redistribution vote will not affect their payment or the payment of anyone in their group.



You must move the slider in order to continue, even if you move it only to bring it back to zero.

Tax Rate:

0%

You must move the slider in order to continue, even if you move it only to bring it back to zero.

Tax Rate:

50%

Figure 4.2: Opposite Group Redistribution - r=0, r=50%

Participants are then incentivized to guess how many test questions they answered correctly. This response is used to record their belief update or post-test confidence.

Participants are also asked to make an investment decision to assess their risk preferences.

As the last incentivized response, participants are shown the payment and score distribution that the grading system produced for the participants of the other group. They are asked to draw their estimation of what the distribution would look like if each correct answer gave 1 point and each incorrect one gave 0 points (Figure 4.3). Participants can draw any distribution they see fit by adjusting three sliders. This response aims to elicit the participants' estimations of merit at the task level. Finally, participants answer a series of demographic and ideology-related questions.

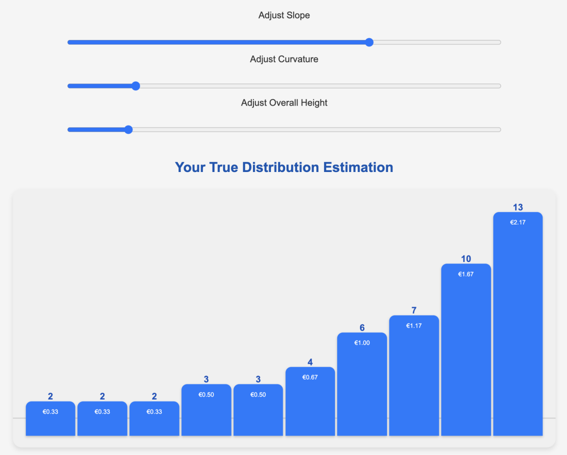


Figure 4.3: True Distribution Estimation Task Example

4.2.2 Control

The control of this experiment differs from the treatments in one major way: there is no probability mechanism in determining point scores. Therefore, all final scores, payments, and distributions seen by participants are based on the total correct answers they gave on the test. Due to this difference, participants in the control only make one score prediction, are not asked to draw a distribution, and do not make post-test estimations of their performance. On Figure 4.4 see a timeline of the decisions that participants make during the experiment. Decisions that are only made in the treatments are highlighted in red.

4.3 Theoretical Specification

The model of this experiment incorporates prior and posterior beliefs, showing how efforts to influence belief updates impact redistribution preferences through merit over- or underestimation in task outcomes. The approach draws on Bénabou and Tirole (2002), which explores self-confidence's effect on motivation and economic choices, integrating consumption utility with self-perception of ability. Their model examines belief updates in response to new information and the psychological costs of sustaining self-confidence.

While similar to Bénabou and Tirole's model, this model diverges by focusing on strategic information manipulation to maintain or adjust self-beliefs, offering a complementary

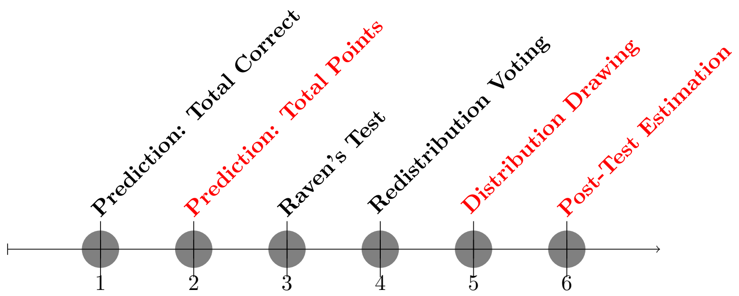


Figure 4.4: Timeline of Experiment

perspective on belief updating. A distinctive feature of this model is that individuals directly control the weighting of prior beliefs and new information, thereby managing the perceived strength of signals. The model also accounts for cognitive costs in adjusting these weights. Crucially, it includes a mechanism linking self-image concerns to merit and luck perceptions, allowing redistribution preferences to be modeled in terms of self-image related to performance.

Given that this is a primarily experimental research, the theoretical specification presented below is limited. If interested in a more detailed version, please consult the supplemental material (Appendix C.3). The total utility function $U(\omega, r)$ combines *Belief-Dependent Utility*, *Cognitive Cost*, and *Fairness Concerns*:

$$U(\omega, r) = \underbrace{\omega \cdot \chi \cdot \pi + (1 - \omega \cdot \chi) \cdot S}_{\text{Belief-Dependent Utility}} - \underbrace{k(n)(1-\omega)^2}_{\text{Cognitive Cost}} - \underbrace{f((1-m-r)^2)}_{\text{Fairness Concerns}} \tag{4.1}$$

The *Belief-Dependent Utility* represents the utility derived from updating one's belief about one's ability. The belief update is a weighted combination of the prior belief (measured by the performance prediction) π and the signal S (measured by actual income achieved), where χ is the Bayesian weight produced by the relative variances of the prior belief and the signal, and ω is the manipulation factor chosen by the individual. By manipulating ω, the participant can choose whether to put more weight on their prior belief or the signal, depending on what is more beneficial to them.

The *Cognitive Cost* limits the ability of individuals to manipulate their beliefs, increasing as ω deviates from 1 and or with a higher level of noise n ($k'(n) < 0$), reflecting the effort required to manipulate the belief weights.

Fairness Concerns represent the cost of non-meritocratic inequality, where f captures the degree of aversion. Individuals increase their chosen redistribution rate r when their perception of meritocracy m decreases. m is a complement of luck and can be written as $m = 1 - \frac{\chi \cdot \omega}{\Lambda}$, with $\Lambda(\beta, \sigma_0^2) = 1 + \beta \cdot \sigma_0^2$, representing doubt in one's abilities.

Equilibrium. Optimizing in periods 2 (redistribution rate) and 1 (ω), the two equilibrium values are:

$$r^* = \frac{\chi \cdot \omega}{\Lambda} \tag{4.2}$$

$$\omega^* = 1 + \frac{\chi(\pi - S)}{2k} \tag{4.3}$$

Equation 4.5 shows that the redistribution rate r will be a mediated but increasing function of how much weight one puts on their prior belief. When someone completely discounts the signal they receive, they act as if this signal should also be discounted for others. As a result, the individual attempts to fix the resulting inequalities through an increase in the redistribution rate. The impact of χ and ω on the chosen redistribution rate is mediated by the strength of prior beliefs, as determined by the degree of self-doubt Λ. Namely, given a low doubt about one's ability, a person might perceive the overall process as fair but still believe that their performance does not accurately reflect their true abilities.

Note that in other contexts the redistribution rate could be a cost since being taxed reduces one's income. However, in the design of this experiment redistribution rates do not hurt the individual who sets them. In other words, redistribution rates solely represent preferences for redistribution.

Equation 4.6 demonstrates that - mediated by k - ω increases with the difference between expectations and performance $\pi - S$; the lower one's received signal is compared to their prior belief, the more weight they will add to their prior beliefs when updating. When the signal is higher than their prior belief, individuals might attempt to read too much into that signal by decreasing their ω.

Combining the two equations, I find that individuals who perform below their initial expectations are likely to favor higher levels of redistribution, while those who exceed their expectations will tend to support lower levels of redistribution. A high cognitive cost (k) due to lower noise or high doubt in one's ability limits the positive effect of the differential between expectations and performance on demand for redistribution. In this experiment, (k) is the lowest in the control and highest in the Profitable Noise treatment. However, the Profitable Noise treatment is designed to produce high signals (S) which reduce the incentive for belief manipulation; therefore, manipulations of ω are expected to be higher in the Neutral Noise treatment, where noise is higher than the control and incentives are strong by design.

4.4 Hypotheses

General Intuition The experiment measures initial expectations for correct answers and points, redistribution rate votes for other groups, beliefs about task meritocracy (based on perceived income distribution), and post-test beliefs on correct answers. An individual attributing results to luck is expected to favor higher redistribution rates and draw distributions with GINI coefficients that differ from the observed distribution. Although the model does not predict the exact direction of these coefficients, high-ranking individuals who attribute their position to luck are likely to draw higher GINI coefficients if they believe they should have scored better. This study aims to identify individuals who over- or underestimate the role of luck or merit. Using participants' point predictions as indicators of initial beliefs, those falling short of expectations may manage self-image by overemphasizing luck, while those meeting or exceeding expectations may view outcomes as meritocratic. This process is mediated by individuals' capacity to differentiate between personal and group-level beliefs and their degree of Bayesian reasoning, measured by the accuracy of updates between predictions.

Personal versus Group-level Beliefs Some participants may attribute personal outcomes to luck while viewing the overall test as merit-based. Such individuals might replicate observed GINI coefficients in their distributions but overestimate their correct answers post-test, suggesting a separation of personal and group merit beliefs. This bias thus operates at two levels: personal and group beliefs, with self-image influencing personal beliefs. Increased luck beliefs at the group level are likely to coincide with increased personal luck beliefs. However, increased luck beliefs at the personal level do not require increased luck beliefs at the group level.

Bayesianism Initial predictions of correct answers and points provide a measure of participants' Bayesianism. A fully Bayesian participant expecting 15 correct answers would predict 11.25 points, while a non-Bayesian might expect the full 15 points. Participants' capacity for accurate updates influences later belief adjustments. Unlike perfect Bayesians, non-Bayesian individuals rely more heavily on initial beliefs, exhibit confirmation bias, and may show overconfidence. This measure indirectly captures the cognitive capacity relevant to motivated reasoning, as - in contrast to bounded rationality or heuristics - more skilled reasoning tools often aid in its deployment.

Overvaluing Initial Beliefs Participants scoring below their expectations are likely to maintain initial beliefs, attributing outcomes to luck. They may favor higher redistribution rates, overestimate post-test performance, and draw lower GINI coefficients when lower in distribution and higher GINI coefficients when higher. Additionally, participants are expected to update beliefs accurately at higher rates when signals are positive (Eil and Rao, 2011b), while negative signals could prompt non-Bayesian responses. Consequently, when contrasted with the lack of Bayesian updating when faced with negative signals the Bayesian behavior of those receiving positive signals appears as a bias. This bias causes participants scoring above expectations to overvalue merit, leading to lower redistribution rate votes and GINI coefficients near observed values.

Difference Between the Two Treatments The Profitable Noise treatment introduces conditions in which noise increases incomes, minimizing the role of initial beliefs. In the Neutral Noise treatment, incomes are comparable to the control group, allowing motivated thinking under conditions of perceived randomness. Namely, the noise of the treatment's signal and its average values create better conditions for the treatment's participants to employ motivated thinking

Control Key measures—total correct answer predictions, test scores, income, and redistribution voting—are consistent across conditions. With no changing incentives, initial predictions should not differ significantly between treatment and control. However, the prediction-test score difference is expected to have a smaller effect on redistribution preferences in the control group, as bias often rises with noise. When predictions exceed actual incomes, control group participants are still likely to support higher redistribution rates, though redistribution demands are generally expected to be lower due to fairer grading perceptions and less bias potential. Table 4.1 includes a summary of the main predictions.

Table 4.1: Summary of Main Predictions

H1	The higher the initial overconfidence, the higher the redistribution rate votes. The role of expectations will be stronger in the Neutral Noise group.
H2	The higher the redistribution rate voted: (a) The higher the post-test confidence. (b) The lower the estimation of true inequality.
H3	The higher the initial overconfidence, the higher the post-test confidence.
H4	The higher the initial overconfidence, the lower the participants' estimations of true inequality.

4.5 Results

4.5.1 Descriptive results

The experiment was pre-registered on AsPredicted.org prior to data collection, programmed using oTree, and conducted on Prolific.com. A total of 507 participants completed the three experiment treatments (Control: 163, Profitable Noise: 155, Neutral Noise: 189), comprising 242 females, 254 males, and 11 who either did not respond or selected ”other”. The average payment was €2.37. Average scores were highest in the Profitable Noise treatment and slightly lower in the Neutral Noise treatment (Table 4.2, column 2). Overconfidence was most pronounced in the control group, followed closely by the Neutral Noise group, with substantially lower levels in the Profitable Noise group (Table 4.2, column 3). Differences in voted redistribution rates were minor, with the Neutral Noise group voting for the highest average rates and the Profitable Noise group the lowest (Table 4.2, column 4).

Table 4.2: Summary of Score, Overconfidence, and Tax Vote by Treatment

Treatment	Score	Overconfidence	Tax Vote
Control	5.9387	4.1779	31.9632
Profitable Noise	7.3677	1.9935	28.7226
Neutral Noise	5.7249	3.6402	33.5132

Average subgroup inequality - measured with the GINI coefficient - was the highest in

Table 4.3: Summary of Original and After Tax GINI by Treatment

Treatment	Original GINI	After Tax GINI
Control	0.2121	0.1460
Profitable Noise	0.1328	0.0946
Neutral Noise	0.1646	0.1083

the control group and lowest in the Profitable Noise group (Table 4.3, column 2). Given that the score probability mechanism in the Neutral Noise group was designed to hurt those at the bottom of the distribution, the resulting inequality of the group is between the levels of the other two treatments.

Figure 4.5 shows the average redistribution rates voted by each income quartile, separated by treatment and overconfidence. The top three graphs, labeled "Modest," represent participants without initial overconfidence, while the bottom three display the redistribution rates voted by overconfident participants. Only the overconfident participants display a discernible negative correlation between income and voted redistribution rates. The more pronounced pattern appears among the overconfident participants of the Neutral Noise group.

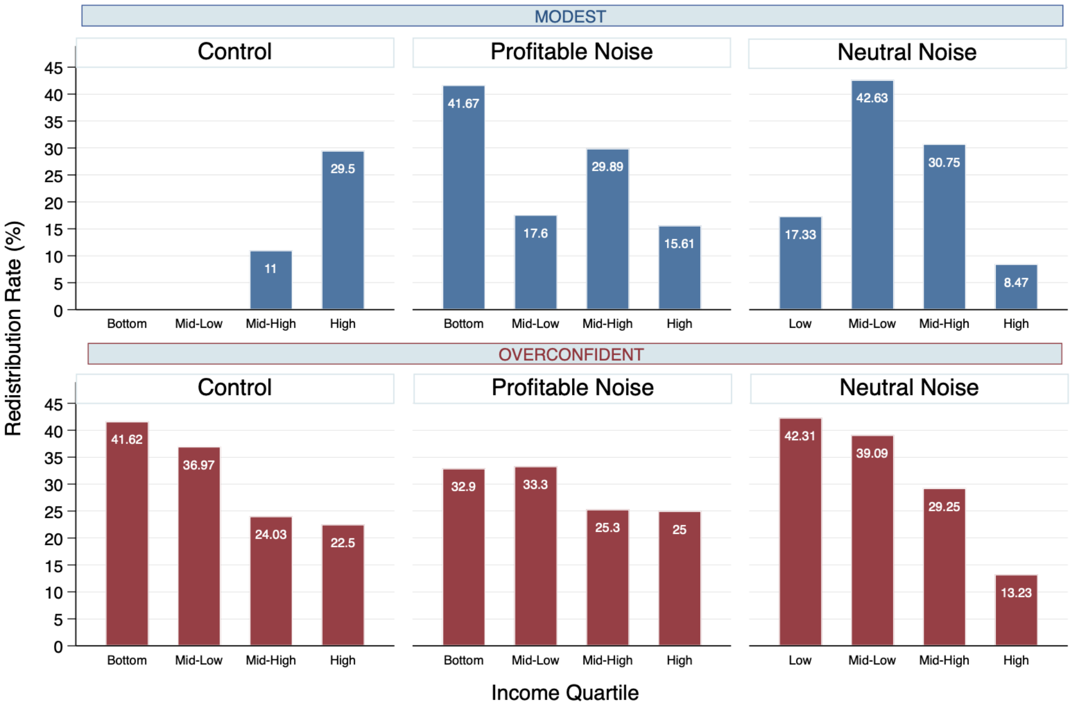


Figure 4.5: Redistribution Rate Votes over Income Quartiles, by Treatment, Overconfidence

As shown in Figure 4.6, which displays the redistribution rates for each of the five ideology groups and separates the data between modest and overconfident participants, a clear pattern emerges. Modest participants tended to vote for redistribution rates that aligned with their ideological beliefs. Specifically, the highest redistribution rates were observed among participants on the left end of the ideological spectrum, while the lowest redistribution rates were recorded among those on the right end. In contrast, overconfident participants did not exhibit the same ideological consistency. Their voting behavior for redistribution rates deviated from the ideological pattern seen in modest participants.

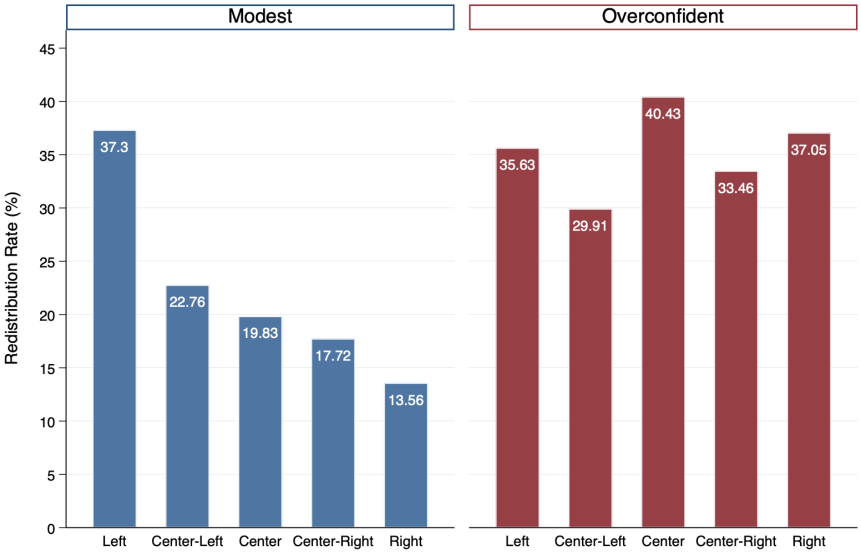


Figure 4.6: Redistribution Rate Votes over Ideology Groups, by Overconfidence

In the Profitable and Neutral Noise treatments, participants provided two additional responses. First, post-test confidence was measured (column 2, Table 4.4). The Neutral Noise group retained some of its initial overconfidence, while post-test overconfidence in the Profitable Noise group was near zero. Second, participants guessed the income distribution of the other group without the probability mechanism (column 4, Table 4.4). Table 4.4 shows that the true GINI coefficients for the Profitable Noise group are higher than those produced by the probability mechanism, leading to lower observed inequality in this treatment. In the Neutral Noise treatment, noise minimally impacts inequality, showing a small difference between true and noise-generated inequality.

Both groups guessed more unequal income distributions than those displayed or the true distributions, with the Profitable Noise group estimating far more inequality despite receiving a lower inequality signal compared to the Neutral Noise group.

Table 4.4: Summary of Post-Confidence and GINI Guesses by Treatment

Treatment	Post-Confidence	Original Gini	Guessed GINI
Profitable Noise	-0.0789	0.1321	0.1921
Neutral Noise	1.0168	0.1649	0.1833

Discussion of Descriptive Results Across Treatments

Table 4.2 confirms that treatment conditions were successfully implemented. In the Neutral Noise treatment, added noise yielded incomes that closely resembled those in the control group, while the Profitable Noise treatment resulted in significantly higher incomes. This pattern is reflected in overconfidence levels: both the control and Neutral Noise groups exhibit similar overconfidence, while the Profitable Noise group displays substantially reduced overconfidence.

Table 4.3 further reveals how these relationships extend to within-group inequality. Although the Neutral Noise group's inequality falls between the control and Profitable Noise groups, it is less aligned with the control than the average scores were. Despite this, the Neutral Noise group favored higher redistribution rates, suggesting that initial overconfidence may play a role in these preferences.

Figure 4.5 offers additional insights, consistent with the literature, indicating that participants with low or no overconfidence tend to adjust beliefs more in line with Bayesian reasoning. Overconfident individuals, however, struggle to maintain their self-image following negative feedback. In the Profitable Noise group, the relationship between income and redistribution rate vote is less pronounced, likely due to lower overconfidence resulting from higher income. In contrast, this negative relationship is stronger in the Neutral Noise group, where lower income and additional noise may allow for the rationalization of lower outcomes compared to the control.

The influence of overconfidence is further illustrated in Figure 4.6, where overconfident participants show no correlation between redistribution rate and ideology, in contrast to modest participants, who did follow an ideological pattern. This suggests that overconfidence impacted their decision-making in a way that overrode the typical ideological influences, disrupting the expected alignment between ideological beliefs and redistribution preferences. It should be noted that being in the top 50% of a session's income distribution did not exhibit the same pattern (Appendix, Figure 4.7).

Table 4.4 further confirms that treatment conditions were met. The GINI coefficient in the Neutral Noise group reflects inequality levels similar to a noise-free scenario, whereas inequality in the Profitable Noise group was reduced, aligning with the observed income increases. Post-test overconfidence was absent in the Profitable Noise group, while it remained positive in the Neutral Noise group, indicating that participants in the Profitable Noise group adjusted beliefs based on post-test signals, while those in the Neutral Noise group maintained beliefs of deserving better outcomes. Interestingly, participants in the Profitable Noise group estimated higher inequality (true GINI) within their group, despite experiencing lower inequality. This discrepancy could be stemming from the group's higher confidence caused by the higher signals.

Table 4.5: Regression Results - Tax Vote, All Treatments

	(1)	(2)	(3)	(4)
INCOME	-2.570**			
	(0.831)			
#Control		-2.727**	-2.718**	-2.545**
		(0.885)	(0.908)	(0.927)
#Profitable Noise		-2.295**	-2.416**	-2.420*
		(0.879)	(0.893)	(0.944)
#Neutral Noise		-3.100***	-3.312***	-2.785**
		(0.923)	(0.936)	(0.990)
OVERCONFIDENCE	0.878			
	(0.540)			
#Control		0.702	0.736	1.353+
		(0.725)	(0.744)	(0.748)
#Profitable Noise		0.336	0.375	0.680
		(0.737)	(0.752)	(0.796)
#Neutral Noise		1.588*	1.579*	1.744*
		(0.753)	(0.753)	(0.802)
Male			4.461	6.809*
			(2.992)	(3.177)
IDEOLOGY				
#Modest				-2.155**
				(0.803)
#Overconfident				0.035
				(0.416)
Constant	44.823***	45.240***	43.988***	40.024***
	(6.896)	(7.047)	(7.203)	(7.624)
Observations	507	507	496	446
Adjusted R^2	0.057	0.055	0.060	0.080

Standard errors in parentheses. + p<0.1, * p<0.05, ** p<0.01, *** p<0.001.

4.5.2 Regression results

Redistribution, Income and Overconfidence

Table 4.5 presents the results of four OLS regressions, with redistribution rate votes as the dependent variable. In the first model, which includes all participants without distinguishing by treatment, participants' incomes exhibit a statistically significant negative correlation with redistribution rate votes. Specifically, a higher income correlates with lower redistribution rate preferences, suggesting that participants who perform better on the task are less inclined to support higher redistribution rates. This inverse relationship between income and redistribution rate vote is both statistically significant and robust.

When the analysis is disaggregated by treatment, the negative effect of income on redistribution rate votes persists across all three treatments, maintaining statistical significance. Notably, in the Neutral Noise treatment, overconfidence exhibits a statistically significant positive effect on redistribution rate votes. This indicates that in this particular treatment, participants who displayed higher levels of overconfidence were more likely to favor increased redistribution rates.

The addition of gender as a control variable does not alter the core relationships observed in terms of direction or magnitude (Column 3). However, when ideology is added to the model (Column 4), and the analysis is split between modest and overconfident participants based on the previously observed patterns, I find that ideology plays a significant role among modest participants. Specifically, there is a significant negative correlation between ideology and redistribution rate voting for modest participants, but this relationship is absent among overconfident participants, aligning with our earlier observations.

The inclusion of ideology also results in a slight decrease in the coefficient size of income in both the control and neutral noise groups and a reduction in the statistical significance of income effects in the profitable noise and neutral noise treatments. Additionally, I observe a modest increase in the coefficient for overconfidence in the neutral noise group. Notably, in the control group, there is a substantial increase in the coefficient of overconfidence, along with a slight gain in statistical significance. Finally, adding ideology to the model also brings statistical significance to the gender variable. Specifically, male participants show a tendency to vote for higher redistribution rates on average.

Regression Results, Weighted

I now focus specifically on the two treatment groups: Profitable Noise and Neutral Noise. In Table 4.6, the first column replicates the model from Table 4.5 (column 4) but is restricted to these two treatments. The results are broadly similar to those in the previous analysis, although some statistical significance is lost. For instance, the effect of income on redistribution rate votes in the Profitable Noise treatment is no longer as statistically significant as previously observed.

The rest of the columns display the same regression but with observations weighted according to the weighting procedure detailed in (Appendix Section C.4). The model of the second column is weighted to emphasize responses from more naive participants, giving more weight to participants who deviated from the Bayesian perfect prediction in absolute terms. With this approach, the coefficients and their statistical significance remain largely similar to those in the general model. However, two notable changes are that in the naive-weighted model, the coefficient for the Profitable Noise treatment becomes statistically significant at the 5% level and the one of ideology among modest participants loses some of its statistical significance.

In the third column, the weights are adjusted to emphasize more sophisticated participants—those who closely approximated perfect Bayesian updating between predictions. Here, the results remain consistent with the general model, maintaining similar statistical significance and coefficient sizes.

The fourth and fifth columns introduce weights which account for numerical deviations from the perfect Bayesian prediction, distinguishing between positive and negative deviations. Namely, the weights used in the last two columns are directional. The optimism weight emphasizes positive deviations from the perfect Bayesian prediction and the pessimism weight highlights negative deviations. In the fourth column, which includes weights for optimism, the role of overconfidence diminishes overall, displaying no statistical significance in either the Neutral or Profitable Noise treatments. Compared to the unweighted model, the income coefficients increase and the one of the Profitable Noise gains statistical significance. The coefficient of gender decreases and displays a decrease in statistical significance.

In the fifth column, which includes weights for pessimism, the statistical significance of income for both groups is reduced, accompanied by a reduction in the sizes of the coefficients, nonetheless, a slight statistical significance is retained in the coefficient of the Neutral Noise group. However, the coefficient for overconfidence in the neutral noise group gains statistical significance and increases in size. Additionally, gender plays a more prominent role, displaying a larger and more statistically significant correlation with the redistribution rate. There is a 9 percentage-point disparity in baseline redistribution voting between the optimist and pessimist models. The optimism-weighted model includes a constant of 43.8%, while the pessimist-weighted model starts from a constant of 34.9%.

Regression Results - Post-Confidence and Merit Perceptions

To evaluate the validity of post-confidence levels and participants' GINI estimates as indicators of merit beliefs, I conducted several regressions analyzing the relationships between post-confidence, GINI estimates, and redistribution votes. In the first column of Table 4.7, post-confidence serves as the dependent variable, with participants categorized not by treatment but by their initial confidence level—either modest or overconfident. This distinction recognizes that overconfident individuals are generally more prone to belief manipulation. The results indicate a positive correlation between redistribution and post-confidence among overconfident participants, suggesting that higher redistribution votes are associated with increased post-test confidence within this group.

In the second column, the dependent variable is the change in GINI estimates, defined as the difference between participants' GINI estimates during the experiment and their initial observations. A positive relationship emerges between post-confidence and GINI estimate change for initially overconfident participants, implying that higher post-confidence correlates with a perception of lower inequality within the group. This suggests that participants perceiving their performance as undeservedly low are inclined to view the group as exhibiting greater inequality than observed.

The third column explores the relationship between GINI change and redistribution vote. Here, the relationship is weak, achieving statistical significance only at the 10% level. This finding indicates that while a significant relationship exists between redistribution vote and post-confidence, as well as between post-confidence and GINI change, these

Table 4.6: Regression Treatment Results Weighted

	No Weight	Naive	Sophisticated	Optimists	Pessimists
INCOME					
#Profitable Noise	**-2.302+**	**-2.512***	**-2.182+**	**-2.897***	-1.976
	(1.210)	(1.251)	(1.227)	(1.262)	(1.219)
#Neutral Noise	**-2.810***	**-3.029***	**-2.683***	**-3.387***	**-2.498+**
	(1.267)	(1.285)	(1.300)	(1.316)	(1.278)
OVERCONFIDENCE					
#Profitable Noise	0.446	0.596	0.333	0.463	0.423
	(0.851)	(0.857)	(0.878)	(0.945)	(0.824)
#Neutral Noise	**1.887***	**1.885***	**1.873***	1.444	**2.131***
	(0.891)	(0.899)	(0.922)	(0.911)	(0.907)
Male	**8.720***	**8.285***	**9.057***	**7.152+**	**9.639***
	(3.951)	(4.038)	(4.007)	(4.125)	(3.948)
IDEOLOGY					
#Modest	**-1.797***	**-1.632+**	**-1.916***	**-1.948***	**-1.711***
	(0.845)	(0.928)	(0.823)	(0.921)	(0.842)
#Overconfident	-0.661	-0.713	-0.630	-0.511	-0.746
	(0.497)	(0.513)	(0.501)	(0.506)	(0.501)
Constant	38.085***	39.371***	37.366***	43.803***	34.943***
	(10.022)	(10.207)	(10.260)	(10.538)	(10.063)
Observations	298	298	298	298	298
Adjusted R^2	0.071	0.080	0.066	0.073	0.072

Standard errors in parentheses. + p<0.1, * p<0.05, ** p<0.01, *** p<0.001.

connections do not fully extend to a robust direct link between redistribution vote and GINI change. Thus, the evidence for a strong association between redistribution votes and perceived inequality is limited.

Table 4.8 presents regressions with post-test confidence as the dependent variable, using the same model specifications as in the fourth column of Table 4.5, but run separately for each treatment due to differences in how post-confidence is estimated between the Profitable Noise and Neutral Noise treatments. In the Profitable Noise treatment, the probability mechanism generally resulted in better participant performance, while in the Neutral Noise treatment, performance closely resembled that of the control group.

Within the Profitable Noise treatment, initial overconfidence has a positive and statistically significant effect on post-test overconfidence, indicating that participants with higher initial overconfidence levels maintained higher levels of post-test overconfidence. This relationship is also present in the Neutral Noise treatment but with greater statistical significance and a larger coefficient, suggesting that overconfidence exerts an even stronger influence on post-test confidence within this group. Apart from overconfidence, gender also shows a statistically significant effect on post-test confidence in the Neutral Noise treatment.

Table 4.7: Regression Results - Relationships between Merit Perception Indicators

	(Post-Confidence)	(GINI Change)	(GINI Change)
TAX VOTE			
#Modest	-0.015		-0.002
	(0.009)		(0.002)
#Overconfident	**0.012****		**-0.003+**
	(0.004)		(0.002)
POST-CONFIDENCE			
#Modest		**-0.059**	
		(0.046)	
#Overconfident		**-0.077***	
		(0.033)	
Male	0.279	0.180	0.176
	(0.272)	(0.119)	(0.121)
Ideology	0.028	-0.004	-0.008
	(0.036)	(0.015)	(0.015)
Constant	0.262	**0.277****	**0.308****
	(0.235)	(0.090)	(0.102)
Observations	298	233	233
Adjusted R-squared	0.051	0.027	0.005

Standard errors in parentheses. + p<0.1, * p<0.05, ** p<0.01, *** p<0.001.

Table 4.8: Regression Results - Post-Confidence

	(Profitable Noise)	(Neutral Noise)
Overconfidence	**0.184***	**0.243*****
	(0.074)	(0.069)
Income	-0.208	-0.102
	(0.128)	(0.096)
Male	0.073	**0.719***
	(0.380)	(0.334)
Ideology	0.014	0.001
	(0.057)	(0.038)
Constant	1.147	0.369
	(1.110)	(0.711)
Observations	124	174
Adjusted R^2	0.149	0.169

Standard errors in parentheses.*p<0.05,**p<0.01,***p<0.001.

Finally, regressions using estimated GINI as the dependent variable did not yield notable results (Appendix Table 4.11).

4.5.3 Hypotheses Evaluation Summary

H1: The higher the initial overconfidence, the higher the redistribution rate votes. The role of expectations will be stronger in the Neutral Noise group. This hypothesis was confirmed, with the Neutral Noise group being the only one exhibiting a statistically significant correlation between redistribution rate voted and overconfidence. The control group exhibited a higher correlation between the two variables than the Profitable Noise group; however, at a statistically insignificant level.

H2: The higher the redistribution rate voted, the higher the post-test confidence (a) and the lower the estimation of true inequality (b). The first part of this hypothesis was confirmed, with a strong correlation between redistribution vote and post-test confidence. The evidence to support that a higher redistribution rate will correlate with a lower estimated true inequality is weak.

H3: The higher the initial overconfidence, the higher the post-test confidence. Initial overconfidence positively correlated with post-test confidence in both treatments. The correlation was higher and more statistically significant in the Neutral Noise treatment.

H4: The higher the initial overconfidence, the lower the participants' estimations of true inequality. No evidence was found that overconfidence directly affected estimations of true inequality.

4.6 Discussion

Income as a Belief Signal The results of this experiment suggest that participants were influenced by the income they achieved, even in the absence of any direct tax concerns for themselves or their group. Across all three treatments, participants who achieved a higher income consistently voted for lower redistribution rates. This finding is intriguing because it indicates that self-belief manipulation may have been at play, as participants' incomes affected their redistribution preferences. Even if we focus solely on this finding and disregard the rest of the results from this research, the result further confirms that the winner's bias must be understood, at least in part, through the lens of self-image maintenance.

Initial Expectations In the Neutral Noise group, initial overconfidence proved to be a significant factor. Participants who exhibited higher overconfidence at the beginning of the experiment were more inclined to vote for increased redistribution rates later on—a pattern uniquely prominent within this treatment. The Neutral Noise treatment was designed to introduce moderate noise into the signal, allowing participants room for belief manipulation while keeping incomes roughly comparable to the control group. Although the control group displayed similar, though less statistically insignificant, patterns, the absence of a probabilistic score mechanism in the control limited participants' capacity to manipulate self-beliefs compared to those in the Neutral Noise group.

In contrast, the Profitable Noise treatment introduced more favorable noise, resulting in higher incomes. Here, the low coefficients for overconfidence and lack of statistical significance suggest that participants, generally satisfied with their results, felt less compelled to support higher redistribution rates. In summary, the control group was given limited scope for expectations-related belief manipulation and, despite the potential for such manipulation being available in both treatment groups, only participants in the Neutral Noise group engaged in it to preserve a positive self-image.

Incorporating expectations into the analysis of winner's bias provides explanatory power for previously puzzling patterns; such as the recent evidence suggesting that individuals who accumulated their wealth tend to oppose redistribution more than those born into wealth (Cohn et al., 2022; Koo et al., 2023). This phenomenon might initially seem counterintuitive, as one might expect those who have experienced lower income levels to empathize with the struggles of the lower classes. However, viewing this behavior through the lens of expectations offers a clearer interpretation. Individuals born into wealth naturally form high expectations for future income, such as an individual in the top 5% aspiring to reach the top 1%. If they remain at the top 5%, they may feel less inclined to attribute this outcome to personal shortcomings. Conversely, someone born in the bottom 5%, with lower expectations, may set an aspirational target of reaching the top 30%; if they instead achieve a position in the top 5%, they are likely to feel a strong sense of personal achievement. Thus, while both individuals may display winner's bias, the self-made top earner may exhibit a stronger bias against redistribution.

Initial expectations also emerged as a robust predictor of post-test overconfidence. Participants with higher pre-test overconfidence tended to retain elevated confidence even after receiving the post-test signal. This finding underscores the persistence of overconfidence and the influence of self-belief management, suggesting that feedback alone may be insufficient to counter inflated self-perceptions among participants.

Finally, ideology only played a role among participants whose expectations were met, a finding consistent with previous experimental findings (Deffains et al., 2016; Fehr and Vollmann, 2022). The finding becomes more compelling when considering the absence of a similar pattern when participants are separated by low and high income. The prominence of expectations in blurring ideological lines implies that previous experimental findings might have mistakenly captured income position as the defining factor. Accounting for the relationship between expectations and ideologically driven behavior could help explain phenomena such as the recent rapid political changes observed in both the US and Europe. In other words, these changes might not reflect an overall ideological shift in the population, but an abandonment of ideology altogether.

Bayesianism: Is All Motivated Thinking Created Equal? The results of the weighted regressions presented in Table 4.6 did not reveal any substantial differences between the behavior of naive and sophisticated participants, either in coefficient size or statistical significance. This finding implies that the capacity for belief manipulation may be more a function of informational access and cognitive search skills rather than advanced statistical reasoning. Although the literature supports that more educated and intelligent individuals are better at managing their beliefs, their ability to follow perfect Bayesian behavior might not be what allows them to succeed in that belief management. Instead, these findings may imply that their advantage lies in their access to a greater amount of information and their ability to selectively sift through this information to identify and emphasize data that supports their existing beliefs.

However, there was a difference in the patterns observed between the optimism-weighted and pessimism-weighted regressions. In the optimism-weighted regression, income played a more significant role in the creation of the winner's bias. In contrast, the pessimism-weighted regression revealed a shift. For the more pessimistic individuals, overconfidence had a more substantial impact on their redistribution voting behavior. This finding suggests that the model and hypotheses of this research resonate more strongly with the behavior of pessimistic motivated thinkers than with their optimistic counterparts.

This shift can potentially be explained by an awareness of self-belief maintenance among pessimistic participants. As described by Bénabou and Tirole (2002), individuals who are conscious of their tendency toward self-belief manipulation may adopt a more cautious stance toward the signals they receive. When these participants receive a higher income, they do not hastily adjust their redistribution rate votes downward, resulting in a reduced impact of income on their voting behavior compared to optimists. While these pessimistic participants manage to moderate their immediate responses, they remain influenced by a deeper, more subconscious layer of self-belief maintenance—expectations. This deeper level of belief management can result in overconfident pessimists ultimately voting for higher redistribution rates as they attempt to navigate and reconcile their self-perceptions. This complex interaction between conscious self-awareness and subconscious overconfidence highlights a nuanced dimension of decision-making that merits further investigation but is in line with our understanding of motivated beliefs. Unlike heuristics and bounded rationality, motivated beliefs are directional (Bénabou and Tirole, 2016) and this result implies that the directionality of motivated beliefs may be deeply embedded into an individual's decision-making process.

These findings highlight the complexity of inequality perceptions. Such differences in reasoning styles have significant societal implications. For instance, optimistic motivated thinkers may be swayed by ”growing pie” arguments, wherein they feel content with nominal income increases, even as real income stagnates. Consequently, they may refrain from supporting redistributive policies that align with their economic interests, inadvertently acting against their own well-being.

Personal and Group Merit Beliefs Higher post-test confidence among participants —suggesting a belief that they deserved better outcomes than they received—was associated with a preference for higher redistribution rates. However, this relationship was less pronounced when extended to group-level beliefs. Although participants who favored higher redistribution rates generally estimated lower levels of true inequality within the group, this correlation did not reach statistical significance.

As outlined in previous sections, the model anticipated a cascading effect between personal meritocratic beliefs and group-level meritocratic perceptions. In terms of redistribution, the model predicts that personal and group merit beliefs may appear congruent. Specifically, if participants perceive high inequality or a lack of meritocracy at the group level, this perception would likely influence their personal beliefs as well—though the reverse may not necessarily hold. This finding aligns with prior research indicating that individuals often assess inequality through a distributive lens by evaluating the personal impact of tax policies (Bartels, 2005). This perspective suggests that individuals may not be solely driven by self-interest in their responses to tax policy; rather, the direct implications of tax policies might be making the distribution of resources more salient. In other words, personal tax gains or losses may act as a lens, enhancing individuals' understanding of fairness within the broader income distribution and shaping their perceptions of meritocracy.

4.7 Conclusion

A belief in the meritocratic nature of income distribution—or the lack thereof—is a strong predictor of support for redistribution. Previous literature has established that individuals positioned at the top of the income distribution often exhibit a winner's bias, leading them to attribute their status to a meritocratic process. This research sought to gauge the extent of motivated thinking in this process by quantifying beliefs through assessing participants' initial expectations.

The findings confirm that the winner's bias endures largely as a function of self-image maintenance. Initial expectations played a significant role, especially within the Neutral Noise treatment group, where conditions were optimized to encourage motivated thinking. Moreover, studying the participants' initial—unmotivated—updating process revealed that its directionality was more strongly associated with redistribution preferences than its accuracy. Finally, personal beliefs about individual merit had a stronger influence on redistribution voting than group merit beliefs; participants who felt they underperformed were more inclined to support higher redistribution rates.

Future research should investigate scenarios where the introduced noise disadvantages participants, providing insight into how outcomes shape perceptions of fairness. Additionally, it would be valuable to test these experimental treatments with redistribution votes that directly impact the participant's group, to better understand how motivations for personal tax profit compare with the influence of self-beliefs and self-image maintenance when implicit measurements of beliefs are involved.

The participants' prediction updating correlated with different patterns of motivated thinking. Subsequent research should investigate the existence of such early indicators of potential motivated thinking and bias patterns. Identifying similar signs can help us understand the problem and tackle it more accurately. For instance, approaches suited to addressing optimistic motivated reasoning may differ from those required to mitigate pessimistic motivated reasoning. Understanding these distinctions will be crucial in developing targeted measures and policies.

Lastly, the observed shift away from ideological commitment among participants who underperformed relative to their expectations, coupled with the absence of a clear trend among low and high-income groups, suggests that expectations may play a crucial role in shaping both economic and political behavior, warranting further investigation. Incorporating expectations into the analysis of contemporary issues in economics and political economy could yield useful insights.

Appendix C

C.1 Tables

Table 4.9: Participants per Treatment

Treatment	Freq.	Percent
Control	163	32.15
Profitable Noise	155	30.57
Neutral Noise	189	37.28
Total	507	100.00

Table 4.10: Summary of Tax Vote by Gender

Gender	Mean	Std. dev.	Freq.
Denied	8.7778	23.4829	9
Female	30.0785	31.5354	242
Male	33.7244	36.9523	254
Other	36.0000	50.9117	2
Total	31.5503	34.3940	507

Table 4.11: Regression Results - Guessed Inequality

	(1)	(2)
Income	0.000	0.054
	(0.103)	(0.036)
Overconfidence	-0.009	0.015
	(0.062)	(0.020)
Male	0.234	0.079
	(0.346)	(0.114)
Ideology	0.028	-0.007
	(0.048)	(0.013)
Constant	0.599	-0.235
	(1.019)	(0.249)
Observations	65	168
Adjusted R^2	-0.041	-0.004

Standard errors in parentheses.

C.2 Figures

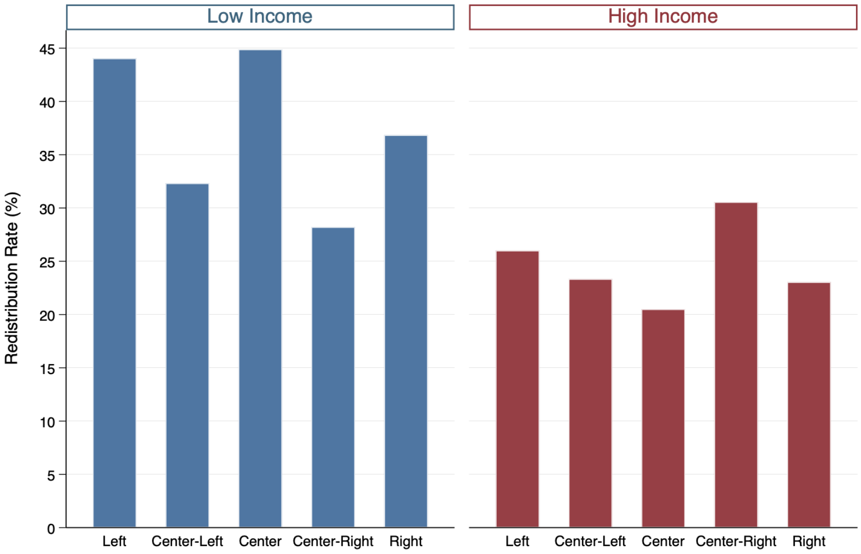


Figure 4.7: Redistribution Rate Votes over Ideology Groups, by Above or Below Session's Median

C.3 Expanded Model Specification

Belief-Dependent Utility ($I(\pi, \chi, \omega)$)

This component represents the utility derived from updating one's belief about one's ability. The belief update is a weighted combination of the prior belief π and the signal S, where χ is the Bayesian weight based on the relative variances of the prior belief and the signal, and ω is the manipulation factor chosen by the individual.

Components of Belief-Dependent Utility

- **Prior Belief** (π): The individual's initial belief about their ability.
- **Signal** (S): New information or evidence about the individual's ability.
- **Bayesian Weight** (χ): The objective weight based on the variances of the prior belief and the signal:

  $$\chi = \frac{\frac{1}{\sigma_0^2}}{\frac{1}{\sigma_0^2} + \frac{1}{\sigma_s^2}}$$

  where σ_0^2 is the variance of the prior belief and σ_s^2 is the variance of the signal.
- **Manipulation Factor** (ω): Chosen by the individual to adjust the influence of the prior belief and the signal on the updated belief.

The belief-dependent utility is given by:

$$I(\pi, \chi, \omega) = \alpha\left(\omega \cdot \chi \cdot \pi + (1 - \omega \cdot \chi) \cdot S\right)$$

where α is the weight of belief-dependent utility.

Cognitive Cost of Manipulating ω

This represents the cognitive cost associated with choosing ω. The cost increases as ω deviates from 1, reflecting the effort required to manipulate the belief weights.

$$\text{Cost}(\omega) = k(1 - \omega)^2$$

where k is a parameter representing the cost sensitivity.

Combining these elements, we have:

$$I(\pi, \chi, \omega) - \text{Cost}(\omega) = \alpha\left(\omega \cdot \chi \cdot \pi + (1 - \omega \cdot \chi) \cdot S\right) - k(1 - \omega)^2$$

Inequality Aversion Utility ($F(r, m)$)

This component represents the utility associated with inequality aversion, where f captures the degree of inequality aversion, m denotes belief in meritocracy, and r is the tax rate.

The function is given by:

$$F(r,m) = -f((1-m-r)^2)$$

Merit and Luck

Merit and luck balance the influence of the prior belief and the signal:

Merit is inversely related to the weighted influence of luck.

$$\mathrm{m} = 1 - \frac{\chi \cdot \omega}{1 + \beta \cdot \sigma_0^2}$$

where β is an adjustment parameter that controls for the influence of the prior belief's variance.

The influence of χ and ω is mediated by the strength of the prior beliefs through β and σ_0^2 because even if part of the update manipulation influences beliefs on merit and luck, another part could always be a result of the conviction someone has regarding their skill. Namely, someone might think a process was relatively fair overall, but simultaneously they might believe their performance does not represent them.

Luck is the complement of merit.

$$\mathrm{l} = 1 - \mathrm{m}$$

Total Utility

The total utility function $U(\omega, r)$ combines task performance utility, belief-dependent utility, inequality aversion. Substituting the expanded functions we have:

$$U(\omega, r) = \alpha\left(\omega \cdot \chi \cdot \pi + (1 - \omega \cdot \chi) \cdot S\right) - k(1-\omega)^2 - f((1-m)-r)^2 \tag{4.4}$$

Equilibrium

The individual aims to maximize their total utility by choosing the optimal values of ω and r:

$$\max_{\omega, r} U(\omega, r) = G(x, e) + \alpha\left(\omega \cdot \chi \cdot \pi + (1 - \omega \cdot \chi) \cdot S\right) - k(1-\omega)^2 - f((1-m)-r)^2$$

Second period: Solving for r

The tax rate r is chosen in the second period, after beliefs about merit have been formed. The belief-based utility does not play a role in choosing the optimal tax rate; therefore, we can focus on the inequality aversion component.

Substituting ω into the function Since m $= 1 - \frac{\chi \cdot \omega}{1+\beta \cdot \sigma_0^2}$, we can rewrite the function $F(r,m)$ as:

$$F(r,m) = -f\left(\left(1 - \left(1 - \frac{\chi \cdot \omega}{1+\beta \cdot \sigma_0^2}\right) - r\right)^2\right)$$

Differentiating $U(\omega, r)$ with respect to r

$$\frac{dU(\omega,r)}{dr} = -2f \cdot \left(\frac{\chi \cdot \omega}{1+\beta \cdot \sigma_0^2} - r\right)$$

Setting the first-order condition to zero

$$-2f \cdot \left(\frac{\chi \cdot \omega}{1+\beta \cdot \sigma_0^2} - r\right) = 0$$

Solving for r

Solving the first-order condition:

$$\frac{\chi \cdot \omega}{1+\beta \cdot \sigma_0^2} - r = 0$$

Thus, the optimal value of r is:

$$r^* = \frac{\chi \cdot \omega}{1+\beta \cdot \sigma_0^2} \tag{4.5}$$

Next, to find the optimal value of ω, we differentiate the utility function with respect to ω and set it to zero. Note, that the optimal value of ω does not depend on the inequality aversion component; therefore, we can focus on the belief-dependent utility.

Differentiating $U(\omega, r)$ with respect to ω

$$\frac{dU(\omega,r)}{d\omega} = \alpha\left(\chi \cdot \pi - \chi \cdot S\right) - 2k(1-\omega)$$

Setting the first-order condition to zero

$$\alpha\chi(\pi - S) + 2k(1-\omega) = 0$$

Solving for ω

$$\alpha\chi(\pi - S) = -2k(1-\omega)$$

The optimal ω is given by:

$$\omega^* == 1 + \frac{\alpha\chi(\pi - S)}{2k} \tag{4.6}$$

Equation 4.5 shows that the tax rate r will be a mediated but increasing function on how much weight one puts on their prior belief. When someone completely discounts the signal they receive, they believe that this signal should also be discounted for others; in other words, that the signal is truly noisy and, therefore, luck has played an important part in its creation. As a result, the individual attempts to fix the resulting inequalities through an increase in the tax rate. In other contexts, the tax rate could be a cost, since being taxed reduces one's income. However, in the present context chosen redistribution rates do not hurt the individual who sets them.

Mediated by k and α, ω increases with the difference $\pi - S$; the lower one's received signal is compared to their prior belief, the more weight they will add to their prior belief when updating (Equation 4.6). When the signal is higher than their prior belief, individuals will attempt to read too much into that signal by decreasing their ω.

Together, equations 4.5 and 4.6 tell us that an individual who under-performed their prior belief will tend to prefer higher tax rates and an individual who over-performed their prior belief will prefer lower tax rates.

C.4 Methodology for Constructing Weights

This section of the appendix outlines the construction of four weights—*Bayesianism*, *Anti-Bayesianism*, *Optimistic Updating*, and *Pessimistic Updating*.

Participants in the two treatments make two predictions. They predict the total number of correct answers they will get in the test and how many points their correct answers will give them. Since they are informed about the structure of the point probability mechanism, there is a perfect Bayesian response to the second prediction in relation to their first prediction. For example, a participant in the Neutral Noise group who has predicted they will answer 10 questions correctly, can calculate the expected received points in the following manner:

$$C_i \cdot P_{Ci} + (15 - C_i) \cdot P_{Mi} = 10 \cdot 0.7 + 5 \cdot 0.15 = 7.75$$

where C_i denotes the total correct answers initially predicted, P_{Ci} represents the probability that the point is awarded by the mechanism when the answer is correct, and P_{Mi} is the probability that the point is awarded by the mechanism when the answer is a mistake.

Since it is not possible to answer only a fraction of a question correctly, I round the result to the closest integer. Therefore, in this case, the *Bayes Perfect Prediction* is equal to 8.

To measure how individual predictions deviate from an ideal Bayesian outcome, I define the following variable, signifying deviation from the *Bayes Perfect Prediction*:

$$\mathrm{D}_i = \frac{\mathrm{Prediction}_i - \mathrm{BayesPerfect}_i}{\mathrm{BayesPerfect}_i},$$

where $\mathrm{Prediction}_i$ represents the individual's point prediction and $\mathrm{BayesPerfect}_i$ denotes the theoretically perfect Bayesian prediction for observation i. Positive values indicate overestimation, while negative values denote underestimation relative to the Bayesian benchmark.

The *Bayesianism* weight is defined as:

$$\text{BayesWeight}_i = \exp(-|\text{D}_i|).$$

I use an exponential decay function to down-weight outliers and ensure that BayesWeight_i attains its maximum at $\text{D}_i = 0$, decreasing symmetrically as the deviation grows, emphasizing observations that align closely with Bayesian predictions.

The *Anti-Bayesianism* weight, is defined as:

$$\text{AntiBayesWeight}_i = \max(\text{BayesWeight}) - \text{BayesWeight}_i + \min(\text{BayesWeight}).$$

This formulation ensures that AntiBayesWeight_i reaches its minimum at $\text{D}_i = 0$ and increases as the deviation from the Bayesian prediction becomes more pronounced, making the weight a mirrored image of its counterpart (Figure 4.8).

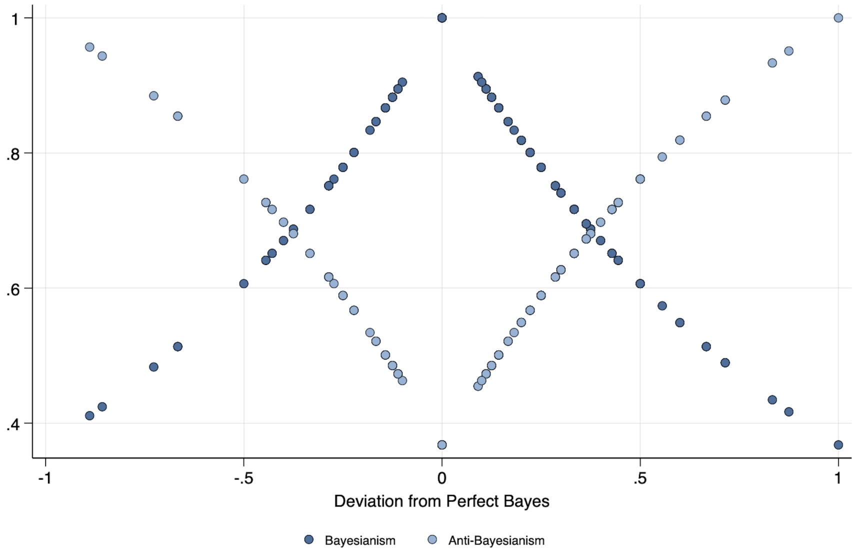


Figure 4.8: Scatter Plot of Bayesianism and Anti-Bayesianism Weights over Deviation from Perfect Bayes

Instead of measuring the absolute distance from the *Bayes Perfect Prediction*, the *Optimistic Updating* weight reflects the numeric distance. The numeric distance incorporates the sign of the deviation, meaning that positive and negative values are treated differently. Therefore, the difference between the two weights is that D_i is now included as is, without transforming it into the negative of its absolute value:

$$\text{OptimistWeight}_i = \exp(\text{D}_i).$$

The *Pessimistic Updating* weight is constructed as a mirrored image of the *Optimistic Updating* weight, utilizing the same process used to build the *Anti-Bayesianism* weight (Figure 4.9).

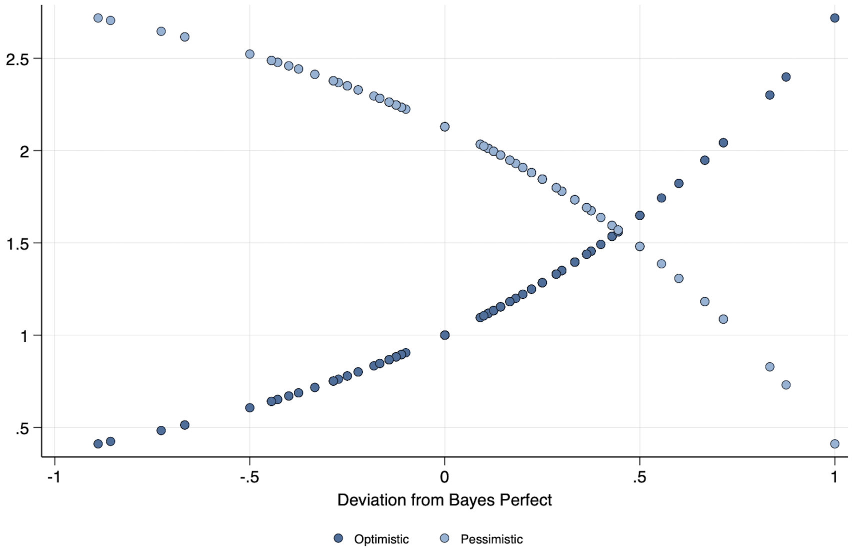


Figure 4.9: Scatter Plot of Optimism and Pessimism Weights over Deviation from Perfect Bayes

Bibliography

Ackerloff, G. (1970). The market for lemons: Quality uncertainty and the market mechanism. *Quarterly Journal of Economics*, 84(3):488–500.

Aghion, P. and Tirole, J. (1997). Formal and real authority in organizations. *Journal of Political Economy*, 105(1):1–29.

Akbaş, M., Ariely, D., and Yuksel, S. (2019). When is inequality fair? an experiment on the effect of procedural justice and agency. *Journal of Economic Behavior Organization*, 161:114–127.

Akerlof, G. A. and Dickens, W. T. (1982). The economic consequences of cognitive dissonance. *American Economic Review*, 72(3):307–319.

Akerlof, G. A. and Kranton, R. E. (2005). Identity and the economics of organizations. *Journal of Economic Perspectives*, 19(1):9–32.

Alesina, A. and Ferrara, E. L. (2005). Preferences for redistribution in the land of opportunities. *Journal of Public Economics*, 89:897–931.

Alesina, A., Glaeser, E. L., and Sacerdote, B. (2001). Why doesn't the united states have a european-style welfare state? *Brookings Papers on Economic Activity*, 2001(2):187–277.

Allcott, H., Boxell, L., Conway, J., Gentzkow, M., Thaler, M., and Yang, D. (2020). Polarization and public health: Partisan differences in social distancing during the coronavirus pandemic. *Journal of Public Economics*, 191:104254.

Allgeier, A., Byrne, D., Brooks, B., and Revnes, D. (1979). The waffle phenomenon: Negative evaluations of those who shift attitudinally. *Journal of Applied Social Psychology*, 9(2):170–182.

Almås, I., Cappelen, A. W., Sørensen, E. , and Tungodden, B. (2022). Global evidence on the selfish rich inequality hypothesis. *Proceedings of the National Academy of Sciences*, 119(3):e2109690119.

Alonso, R. and Matouschek, N. (2008). Optimal delegation. *Review of Economic Studies*, 75(1):259–293.

Amasino, D. R., Pace, D. D., and van der Weele, J. (2023). Self-serving bias in redistribution choices: Accounting for beliefs and norms. *Journal of Economic Psychology*, 98:102654.

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.

Babcock, L., Loewenstein, G., Issacharoff, S., and Camerer, C. (1995). Biased judgments of fairness in bargaining. *American Economic Review*, 85(5):1337–1343.

Backus, P., Cubel, M., Guid, M., Sánchez-Pagés, S., and López Mañas, E. (2023). Gender, competition, and performance: Evidence from chess players. *Quantitative Economics*, 14(1):349–380.

Balcetis, E. and Dunning, D. (2006). See what you want to see: motivational influences on visual perception. *Journal of Personality and Social Psychology*, 91(4):612–625.

Balcetis, E. and Dunning, D. (2010). Wishful seeing: More desired objects are seen as closer. *Psychological Science*, 21(1):147–152.

Bartels, L. M. (2005). Homer gets a tax cut: Inequality and public policy in the american mind. *Perspectives on Politics*, 3(1):15–31.

Bartling, B., Fehr, E., Maréchal, M. A., and Schunk, D. (2009). Egalitarianism and competitiveness. *American Economic Review*, 99(2):93–98.

Bénabou, R. and Tirole, J. (2002). Self-confidence and personal motivation. *Quarterly Journal of Economics*, 117(3):871–915.

Bénabou, R. and Tirole, J. (2004). Willpower and personal rules. *Journal of Political Economy*, 112(4):848–886.

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

Bénabou, R. and Tirole, J. (2011). Identity, morals, and taboos: Beliefs as assets. *Quarterly Journal of Economics*, 126(2):805–855.

Benoit, R. G. and Anderson, M. C. (2012). Opposing mechanisms support the voluntary forgetting of unwanted memories. *Neuron*, 76(2):450–460.

Berns, G. S., Chappelow, J., Cekic, M., Zink, C. F., Pagnoni, G., and Martin-Skurski, M. E. (2006). Neurobiological substrates of dread. *Science*, 312(5774):754–758.

Besley, T. and Ghatak, M. (2005). Competition and incentives with motivated agents. *American Economic Review*, 95(3):616–636.

Bishop, B. and Cushing, R. G. (2009). *The big sort: Why the clustering of like-minded America is tearing us apart.* Houghton Mifflin Harcourt.

Bodner, R. and Prelec, D. (2003). Self-signaling and diagnostic utility in everyday decision making. In Brocas, I. and Carrillo, J. D., editors, *The Psychology of Economic Decisions. Rationality and Well-being*, volume 1, pages 105–123. Oxford University Press.

Brunnermeier, M. K. and Parker, J. A. (2005). Optimal expectations. *American Economic Review*, 95(4):1092–1118.

Burks, S. V., Carpenter, J. P., Goette, L., and Rustichini, A. (2013). Overconfidence and social signalling. *Review of Economic Studies*, 80(3):949–983.

Bénabou, R. (2013). Groupthink: Collective delusions in organizations and markets. *Review of Economic Studies*, 80(2):429–462.

Bénabou, R. and Tirole, J. (2002). Self-confidence and personal motivation. *Quarterly Journal of Economics*, 117(3):871–915.

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

Cain, D. M., Loewenstein, G., and Moore, D. A. (2011). When sunlight fails to disinfect: Understanding the perverse effects of disclosing conflicts of interest. *Journal of Consumer Research*, 37(5):836–857.

Callen, M., Bursztyn, L., Ferman, B., Gulzar, S., Hasanain, A., and Yuchtman, N. (2016). Identifying ideology: Experimental evidence on anti-americanism in pakistan. Technical report, CEPR Discussion Papers.

Caplin, A. and Leahy, J. (2001). Psychological expected utility theory and anticipatory feelings. *Quarterly Journal of Economics*, 116(1):55–79.

Cappelen, A. W., Konow, J., Sørensen, E. , and Tungodden, B. (2013). Just luck: An experimental study of risk-taking and fairness. *American Economic Review*, 103(4):1398–1413.

Carlson, R. W., Maréchal, M. A., Oud, B., Fehr, E., and Crockett, M. J. (2020). Motivated misremembering of selfish decisions. *Nature Communications*, 11(1):2100.

Carrillo, J. D. and Mariotti, T. (2000). Strategic ignorance as a self-disciplining device. *Review of Economic Studies*, 67(3):529–544.

Castagnetti, A. and Schmacker, R. (2022). Protecting the ego: Motivated information selection and updating. *European Economic Review*, 142:104007.

Chance, Z. and Norton, M. I. (2015). The what and why of self-deception. *Current Opinion in Psychology*, 6:104–107.

Charness, G. and Gneezy, U. (2010). Portfolio choice and risk attitudes: An experiment. *Economic Inquiry*, 48(1):133–146.

Cheng, I.-H., Raina, S., and Xiong, W. (2014). Wall street and the housing bubble. *American Economic Review*, 104(9):2797–2829.

Chopra, F., Haaland, I., and Roth, C. (2019). Do people value more informative news? Available at SSRN: https://ssrn.com/abstract=3342595 or http://dx.doi.org/10.2139/ssrn.3342595.

Cohn, A., Jessen, L. J., Klasnja, M., and Smeets, P. (2022). The wealth gap in fairness preferences: Evidence from america's top 5%. Available at SSRN: https://ssrn.com/abstract=3395213.

Cruces, G., Perez-Truglia, R., and Tetaz, M. (2013). Biased perceptions of income distribution and preferences for redistribution: Evidence from a survey experiment. *Journal of Public Economics*, 98:100–112.

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.

Deffains, B., Espinosa, R., and Thöni, C. (2016). Political self-serving bias and redistribution. *Journal of Public Economics*, 134:67–74.

Di Tella, R., Galiant, S., and Schargrodsky, E. (2007). The formation of beliefs: evidence from the allocation of land titles to squatters. *Quarterly Journal of Economics*, 122(1):209–241.

Di Tella, R., Perez-Truglia, R., Babino, A., and Sigman, M. (2015). Conveniently upset: Avoiding altruism by distorting beliefs about others' altruism. *American Economic Review*, 105(11):3416–3442.

Ditto, P. H. and Lopez, D. F. (1992). Motivated skepticism: Use of differential decision criteria for preferred and nonpreferred conclusions. *Journal of Personality and Social Psychology*, 63(4):568–584.

Dorin, C., Hainguerlot, M., Huber-Yahi, H., Vergnaud, J.-C., and de Gardelle, V. (2021). How economic success shapes redistribution: The role of self-serving beliefs, in-group bias and justice principles. *Judgment and Decision Making*, 16(4):932–949.

Echarte, L. E., Bernacer, J., Larrivee, D., Oron, J., and Grijalba-Uche, M. (2016). Self-deception in terminal patients: belief system at stake. *Frontiers in Psychology*, 7:117.

Eil, D. and Rao, J. M. (2011a). The good news-bad news effect: asymmetric processing of objective information about yourself. *American Economic Journal: Microeconomics*, 3(2):114–138.

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

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

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

Engelmann, J. B., Damaraju, E., Padmala, S., and Pessoa, L. (2009). Combined effects of attention and motivation on visual task performance: transient and sustained motivational effects. *Frontiers in Human Neuroscience*, 3:342.

Engelmann, J. B., Lebreton, M., Salem-Garcia, N. A., Schwardmann, P., and van der Weele, J. J. (2024). Anticipatory anxiety and wishful thinking. *American Economic Review*, 114(4):926–960.

Engelmann, J. B. and Pessoa, L. (2014). Motivation sharpens exogenous spatial attention. *Motivation Science*, 1(S):64–72.

Ewers, M. and Zimmermann, F. (2015). Image and misreporting. *Journal of the European Economic Association*, 13(2):363–380.

Eyster, E. (2002). Rationalizing the past: A taste for consistency. *Nuffield College Mimeograph*.

Falk, A. and Zimmermann, F. (2011). Preferences for consistency. (3528). Available at SSRN: https://ssrn.com/abstract=1903622 or http://dx.doi.org/10.2139/ssrn.1903622.

Fehr, D. and Vollmann, M. (2022). Misperceiving economic success: Experimental evidence on meritocratic beliefs and inequality acceptance. (9983). Available at SSRN: https://ssrn.com/abstract=4241623.

Fehr, E. and Charness, G. (2023). Social preferences: fundamental characteristics and economic consequences. Available at SSRN: https://ssrn.com/abstract=4464745.

Fernbach, P. M., Hagmayer, Y., and Sloman, S. A. (2014). Effort denial in self-deception. *Organizational Behavior and Human Decision Processes*, 123(1):1–8.

Finkelstein, E. A., Baid, D., Cheung, Y. B., Schweitzer, M. E., Malhotra, C., Volpp, K., Kanesvaran, R., Lee, L. H., Dent, R. A., Ng Chau Hsien, M., et al. (2021). Hope, bias and survival expectations of advanced cancer patients: a cross-sectional study. *Psycho-Oncology*, 30(5):780–788.

Fischbacher, U. and Föllmi-Heusi, F. (2013). Lies in disguise—an experimental study on cheating. *Journal of the European Economic Association*, 11(3):525–547.

Fong, C. (2001). Social preferences, self-interest, and the demand for redistribution. *Journal of Public Economics*, 82:225–246.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, 19(4):25–42.

Friedrichsen, J. and Engelmann, D. (2013). Who cares for social image? interactions between intrinsic motivation and social image concerns. *CESifo Working Paper Series*.

Friedrichsen, J., König, T., and Schmacker, R. (2018). Social image concerns and welfare take-up. *Journal of Public Economics*, 168:174–192.

Gibbons, R., Roberts, J., et al. (2013). Economic theories of incentives in organizations. *Handbook of organizational economics*, pages 56–99.

Ging-Jehli, N. R., Schneider, F. H., and Weber, R. A. (2020). On self-serving strategic beliefs. *Games and Economic Behavior*, 122:341–353.

Gneezy, U., Niederle, M., and Rustichini, A. (2003). Performance in competitive environments: Gender differences. *Quarterly Journal of Economics*, 118(3):1049–1074.

Gneezy, U. and Rustichini, A. (2004). Gender and competition at a young age. *American Economic Review*, 94(2):377–381.

Gneezy, U., Saccardo, S., Serra-Garcia, M., and van Veldhuizen, R. (2020). Bribing the self. *Games and Economic Behavior*, 120:311–324.

Gödker, K., Jiao, P., and Smeets, P. (2021). Investor memory. *Available at SSRN 3348315*.

Golman, R., Hagmann, D., and Loewenstein, G. (2017). Information avoidance. *Journal of Economic Literature*, 55(1):96–135.

Golman, R., Loewenstein, G., Moene, K. O., and Zarri, L. (2016). The preference for belief consonance. *Journal of Economic Perspectives*, 30(3):165–188.

Golman, R., Loewenstein, G., Molnar, A., and Saccardo, S. (2022). The demand for, and avoidance of, information. *Management Science*, 68(9):6454–6476.

Gottlieb, D. (2014). Imperfect memory and choice under risk. *Games and Economic Behavior*, 85:127–158.

Grossman, Z. and Van der Weele, J. J. (2017). Self-image and willful ignorance in social decisions. *Journal of the European Economic Association*, 15(1):173–217.

Gur, R. C. and Sackeim, H. A. (1979). Self-deception: A concept in search of a phenomenon. *Journal of Personality and Social Psychology*, 37(2):147.

Haeckl, S. (2022). Image concerns in ex-ante self-assessments–gender differences and behavioral consequences. *Labour Economics*, 76:102166.

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

Hansson, K. and Sund, O. (2023). Lucky but confident — how confidence can polarize meritocratic beliefs and preferences for redistribution. Available at SSRN: https://ssrn.com/abstract=4527836.

Hippel, W. v. and Trivers, R. (2011). Reflections on self-deception. *Behavioral and Brain Sciences*, 34(1):41–56.

Huffman, D., Raymond, C., and Shvets, J. (2022). Persistent overconfidence and biased memory: Evidence from managers. *American Economic Review*, 112(10):3141–3175.

Kahan, D. M. (2013). Ideology, motivated reasoning, and cognitive reflection. *Judgment and Decision making*, 8(4):407–424.

Kahan, D. M., Peters, E., Dawson, E. C., and Slovic, P. (2017). Motivated numeracy and enlightened self-government. *Behavioural public policy*, 1(1):54–86.

Kajackaite, A. (2015). If i close my eyes, nobody will get hurt: The effect of ignorance on performance in a real-effort experiment. *Journal of Economic Behavior & Organization*, 116:518–524.

Katok, E. and Siemsen, E. (2011). Why genius leads to adversity: Experimental evidence on the reputational effects of task difficulty choices. *Management Science*, 57(6):1042–1054.

Khalmetski, K. and Sliwka, D. (2019). Disguising lies-image concerns and partial lying in cheating games. *American Economic Journal: Microeconomics*, 11(4):79–110.

Konow, J. (2000). Fair shares: Accountability and cognitive dissonance in allocation decisions. *American Economic Review*, 90(4):1072–1091.

Koo, H. J., Piff, P. K., and Shariff, A. F. (2023). If i could do it, so can they: Among the rich, those with humbler origins are less sensitive to the difficulties of the poor. *Social Psychological and Personality Science*, 14(3):333–341.

Korn, C. W., Sharot, T., Walter, H., Heekeren, H. R., and Dolan, R. J. (2014). Depression is related to an absence of optimistically biased belief updating about future life events. *Psychological Medicine*, 44(3):579–592.

Köszegi, B. (2006). Ego utility, overconfidence, and task choice. *Journal of the European Economic Association*, 4(4):673–707.

Kunda, Z. (1987). Motivated inference: Self-serving generation and evaluation of causal theories. *Journal of Personality and Social Psychology*, 53(4):636–647.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, 108(3):480–498.

Larson, T. and Capra, C. M. (2009). Exploiting moral wiggle room: Illusory preference for fairness? a comment. *Judgment and Decision Making*, 4(6):467–474.

Loewenstein, G. (1987). Anticipation and the valuation of delayed consumption. *The Economic Journal*, 97(387):666–684.

Loughnan, S., Kuppens, P., Allik, J., Balazs, K., De Lemus, S., Dumont, K., Gargurevich, R., Hidegkuti, I., Leidner, B., Matos, L., et al. (2011). Economic inequality is linked to biased self-perception. *Psychological Science*, 22(10):1254–1258.

Lowenstein, G. (1992). The fall and rise of psychological explanations in the economics of intertemporal choice. *Choice over Time. Russell Sage Foundation*, pages 3–34.

Lynch, R. F. and Trivers, R. L. (2012). Self-deception inhibits laughter. *Personality and Individual Differences*, 53(4):491–495.

Mayraz, G. (2011). Wishful thinking. *Available at SSRN 1955644.*

Mei, D., Ke, Z., Li, Z., Zhang, W., Gao, D., and Yin, L. (2023). Self-deception: Distorted metacognitive process in ambiguous contexts. *Human Brain Mapping*, 44(3):948–969.

Mele, A. R. (1997). Real self-deception. *Behavioral and Brain Sciences*, 20(1):91–102.

Mijović-Prelec, D. and Prelec, D. (2010). Self-deception as self-signalling: a model and experimental evidence. *Philosophical Transactions of the Royal Society Biological Sciences*, 365(1538):227–240.

Mollerstrom, J., Reme, B. A., and Sørensen, E. T. (2015). Luck, choice and responsibility — an experimental study of fairness views. *Journal of Public Economics*, 131:33–40.

Molnar, A. and Loewenstein, G. (2020). The false and the furious: People are more disturbed by others' false beliefs than by differences in beliefs. *Available at SSRN 3524651.*

Molnar, A. and Loewenstein, G. (2022). Thoughts and players: an introduction to old and new economic perspectives on beliefs. *The cognitive science of belief: a multidisciplinary approach*, pages 321–350.

Moore, D. A., Tanlu, L., and Bazerman, M. H. (2010). Conflict of interest and the intrusion of bias. *Judgment and Decision Making*, 5(1):37–53.

Müller, M. W. (2022). Selective memory around big life decisions. *Working Paper: https://app.box.com/s/ai7sr1zy648yar6dbq9xpdtvcwhrgvdt.*

Murad, Z., Stavropoulou, C., and Cookson, G. (2019). Incentives and gender in a multi-task setting: An experimental study with real-effort tasks. *Plos One*, 14(3):e0213080.

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

Nyborg, K. (2011). I don't want to hear about it: Rational ignorance among duty-oriented consumers. *Journal of Economic Behavior & Organization*, 79(3):263–274.

Oprea, R. and Yuksel, S. (2022). Social exchange of motivated beliefs. *Journal of the European Economic Association*, 20(2):667–699.

Ortoleva, P. and Snowberg, E. (2015). Overconfidence in political behavior. *American Economic Review*, 105(2):504–535.

Oster, E., Shoulson, I., and Dorsey, E. R. (2013). Optimal expectations and limited medical testing: evidence from huntington disease. *American Economic Review*, 103(2):804–830.

Pittarello, A., Leib, M., Gordon-Hecker, T., and Shalvi, S. (2015). Justifications shape ethical blind spots. *Psychological Science*, 26(6):794–804.

Ploner, M. and Regner, T. (2013). Self-image and moral balancing: An experimental analysis. *Journal of Economic Behavior Organization*, 93:374–383.

Puri, M. and Robinson, D. T. (2007). Optimism and economic choice. *Journal of financial economics*, 86(1):71–99.

Quattrone, G. A. and Tversky, A. (1984). Causal versus diagnostic contingencies: On self-deception and on the voter's illusion. *Journal of Personality and Social Psychology*, 46(2):237–248.

Roy-Chowdhury, V. (2022). Self-confidence and motivated memory loss: Evidence from schools. *Working Paper 2213, Faculty of Economics, University of Cambridge*. Available at: https://ssrn.com/abstract=4527836.

Saucet, C. and Villeval, M. C. (2019). Motivated memory in dictator games. *Games and economic Behavior*, 117:250–275.

Savage, L. J. (1972). The foundations of statistics. Courier Corporation.

Schelling, T. C. (1988). The mind as a consuming organ. *Decision making: Descriptive, normative, and prescriptive interactions*, pages 343–357.

Schrand, C. M. and Zechman, S. L. (2012). Executive overconfidence and the slippery slope to financial misreporting. *Journal of Accounting and Economics*, 53(1-2):311–329.

Schwardmann, P. (2019). Motivated health risk denial and preventative health care investments. *Journal of Health Economics*, 65:78–92.

Schwardmann, P. and Van der Weele, J. (2019). Deception and self-deception. *Nature Human Behaviour*, 3(10):1055–1061.

Sharot, T., Korn, C. W., and Dolan, R. J. (2011). How unrealistic optimism is maintained in the face of reality. *Nature Neuroscience*, 14(11):1475–1479.

Sial, A. Y., Sydnor, J. R., and Taubinsky, D. (2023). Biased memory and perceptions of self-control. Technical report, National Bureau of Economic Research.

Sicherman, N., Law, K., Lipkin, P. H., Loewenstein, G., Marvin, A. R., and Buxbaum, J. D. (2021). Information avoidance and information seeking among parents of children with asd. *American Journal on Intellectual and Developmental Disabilities*, 126(3):249–259.

Sloman, S. A., Fernbach, P. M., and Hagmayer, Y. (2010). Self-deception requires vagueness. *Cognition*, 115(2):268–281.

Smith, A. (1759). *The theory of moral sentiments.* Gutenberg Publishers.

Smith, M. K., Trivers, R., and Von Hippel, W. (2017). Self-deception facilitates interpersonal persuasion. *Journal of Economic Psychology*, 63:93–101.

Spence, M. (1978). Job market signaling, uncertainty in economics. *Quarterly Journal of Economics*, 87(3):355–374.

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

Steele, C. (1988). The psychology of self-affirmation: Sustaining the integrity of the self. *Advances in Experimental Social Psychology*, 21:261–302.

Stiglitz, J. E. and Weiss, A. (1981). Credit rationing in markets with imperfect information. *American Economic Review*, 71(3):393–410.

Thaler, M. (2021). Gender differences in motivated reasoning. *Journal of Economic Behavior & Organization*, 191:501–518.

Thaler, M. (2024). The fake news effect: Experimentally identifying motivated reasoning using trust in news. *American Economic Journal: Microeconomics*, 16(2):1–38.

Thompson, L. and Loewenstein, G. (1992). Egocentric interpretations of fairness and interpersonal conflict. *Organizational Behavior and Human Decision Processes*, 51(2):176–197.

Valdez Gonzalez, N., Brown, A. L., and Palma, M. A. (2023). Social-benefits stigma and subsequent competitiveness. *Available at SSRN 4479804*.

Valero, V. (2022). Redistribution and beliefs about the source of income inequality. *Experimental Economics*, 25(3):876–901.

Verma, R. (2017). Role of self deception, overconfidence and financial aliteracy in household financial decision making. *The Journal of Applied Business and Economics*, 19(8):94–109.

Zimmermann, F. (2020). The dynamics of motivated beliefs. *American Economic Review*, 110(2):337–363.

Títol de la Tesi

Assaigs sobre la Utilitat Basada en Credençes

Resum de la tesi

Aquesta tesi estudia la utilitat basada en creences que estipula que la utilitat de l'individu no només depèn dels resultats materials, sinó també de les seves expectatives i creences.

Capítol 1

El capítol 1 introdueix el concepte de la utilitat basada en creences i resumeix la literatura existent sobre el tema. La utilitat basada en creences planteja que les creences poden proporcionar utilitat directa més enllà del seu valor informatiu o de presa de decisions. Les creences poden ser funcionals—ajudant a assolir objectius—i afectives—proporcionant satisfacció emocional. S'hi analitzen les diferències entre la cognició motivada i els heurístics, destacant biaixos cognitius com la ignorància estratègica, la negació de la realitat i l'autosenyalització. El capítol també examina les implicacions de la utilitat basada en creences a nivell individual i col·lectiu, com ara el seu paper en l'autoengany i la formació de delusions col·lectius. Aquest primer capítol estableix els fonaments conceptuals en els quals es basen els estudis experimentals inclosos a la tesi.

Capítol 2

Els esforços per gestionar la pròpia imatge sovint resulten en autoengany. La literatura existent indica una relació positiva entre aquests dos conceptes. No obstant això, el marc teòric presentat en aquest segon capítol suggereix que la relació monòtona, prèviament observada, no es manté en contextos on les motivacions entren en conflicte. Per construir un context de conflicte, es presenta un disseny experimental innovador que incorpora incentius monetaris, preferències altruistes i preocupacions per la imatge pròpia. Aquest disseny permet modelitzar la imatge pròpia mitjançant una proxy, evitant el biaix associat a les metodologies d'autoavaluació. Els participants rebien un estímul que indicava la seva participació en una de dues tasques d'assignació diferents; una plantejava un dilema moral entre maximitzar les contribucions a organitzacions de caràcter social i el benefici personal, mentre que l'altra no incloïa un dilema moral. Interpretar malament la senyal tenia un cost i no implicava un canvi de la tasca assignada. Els resultats mostren una relació no monòtona entre les preocupacions per la imatge pròpia i l'autoengany, amb la direcció del biaix depenent de l'incentiu dominant: el benefici personal o les preocupacions per la imatge pròpia. Aquesta divergència suggereix dos tipus diferents d'autoengany: motivat pel benefici i motivat per la imatge. Els autoenganyats pel benefici subestimen les oportunitats altruistes i mostren nivells baixos o moderats d'altruisme i preocupacions per la imatge pròpia. Contràriament, els autoenganyats per la imatge sobreestimen els esce-

naris que requereixen un judici ètic i pertanyen a grups amb nivells més elevats d'altruisme i preocupacions per la imatge pròpia. Donades les tendències diferencials dels dos tipus d'autoengany, el biaix mitjà global és nul, cosa que ressalta la importància d'integrar les consideracions per la imatge pròpia en la investigació sobre l'autoengany.

Capítol 3

El tercer capítol de la tesi té com a objectiu determinar si les preocupacions per la imatge social creen conflictes d'interès en una relació principal-agent. Amb aquest objectiu, es dissenya un experiment en què un principal pot delegar l'elecció d'un projecte arriscat a un agent. És desitjable que l'agent s'abstingués d'escollir un projecte quan no disposa d'informació suficient. L'agent només sap quin projecte s'hauria de dur a terme si el seu rendiment relatiu en una prova d'IQ és alt. Tot i així, l'agent pot culpar la mala sort si el projecte no té l'èxit esperat. Les preocupacions per la imatge social sorgeixen quan el principal només pot observar el projecte triat per l'agent, però no el seu rendiment relatiu en la prova. Els resultats mostren que, mentre la interacció sigui anònima, les preocupacions per la imatge social són irrellevants. Per contra, quan la relació no és anònima i el principal observa la foto de l'agent, els agents no informats tenen més probabilitats d'escollir un projecte arriscat. A més, s'observa que aquest efecte està impulsat significativament pels participants masculins.

Capítol 4

Les persones tendeixen a acceptar la desigualtat que perceben com a meritòria però la rebutgen quan l'atribueixen a la sort. En un procés de distribució, els guanyadors tendeixen a sobrevalorar el mèrit, un comportament que sovint atribuït al biaix egocèntric. La investigació experimental que es presenta en l'últim capítol de la tesi, intenta entendre la persistència del biaix del guanyador mitjançant el manteniment de la imatge pròpia. Per quantificar el paper de les creences en aquest procés, s'inclouen expectatives sobre els ingressos futurs dels participants. Les expectatives són un indicador de les creences prèvies. Així, la discrepància entre els ingressos esperats i els obtinguts pot revelar la intensitat dels guanys o pèrdues atribuïbles a la imatge pròpia. Els resultats experimentals indiquen que, quan els ingressos són una senyal sorollosa de l'habilitat d'una persona, els participants que no compleixen les seves expectatives d'ingressos poden protegir la seva imatge pròpia atribuint els resultats a la sort, fet que els porta a advocar per una redistribució més elevada. A més, els participants redistribueixen segons línies ideològiques quan les seves expectatives es compleixen, però aquest efecte desapareix quan no ho fan. Finalment, les creences sobre el mèrit personal tendeixen a eclipsar les creences sobre el mèrit grupal en les preferències per la redistribució.

Paraules clau

Utilitat Basada en Credençes, Autoengany, Altruisme, Imatge Social, Preferències Redistributives, Credençes sobre el Mèrit, Raonament Motivacional, Expectatives, Autoimatge, Preferències Socials, Desigualtat, Sobreconfiança, Comportament Prosocial