UNIVERSITY OF CALIFORNIA SAN DIEGO

Through the looking glass of choice: Conflicting preferences, causal inferences, and social signaling

A Dissertation submitted in partial satisfaction of the requirements
for the degree Doctor of Philosophy

in

Experimental Psychology

by

Xingyu (Shirley) Liu

Committee in charge:

Professor Craig McKenzie, Chair
Professor Jennifer Carr
Professor Michael McCullough
Professor Adena Schachner
Professor Shlomi Sher
Professor Piotr Winkielman

2026

The Dissertation of Xingyu (Shirley) Liu is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2026

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

P1	First-order preferences
P2	Second-order preferences
DM	Decision-maker

ACKNOWLEDGEMENTS

I am grateful to my advisor, Craig McKenzie, for his guidance and mentorship throughout my graduate training. Craig has shaped how I think about rationality and decision-making, and gave me the freedom to explore widely — within decision-making and far beyond it. I deeply appreciate his care, his friendship, and our many engaging conversations about psychology, philosophy, and life, at lunch in San Diego, or conferences around the world.

I am grateful to Michael McCullough, who kindly welcomed me into his lab, and introduced me to the functional perspective of social psychology and to the demands of methodological rigor. I dearly appreciate Mike's generosity, both intellectually – in our conversations ranging from research to literature to music, and socially—in bringing me along to lab retreats and in generously connecting me with researchers in the field.

I am grateful to Adena Schachner for taking me on as one of her own mentees. I am so thankful for our weekly one-on-one meetings and her inexhaustible support. Adena brought the same generosity to our intellectual brainstorming and as with her mentorship and care. She taught me not only about causal inference, but about how to be a better graduate student, mentor, and job candidate. I have learned and grown enormously from our time together.

I am thankful to Shlomi Sher, my professor from Pomona College, whose mentorship has followed me from undergrad into graduate school, and whose intellectual rigor and personal kindness are a constant inspiration for me.

I am thankful to Piotr Winkielman for his often-offbeat but always-illuminating suggestions, his intellectual curiosity, and his whimsy — which remind me why I fell in love with social psychology in the first place.

I am thankful to Jennifer Carr, whose questions and suggestions meaningfully shaped how I framed Chapter 1, and who generously agreed to join my committee without knowing me, and who had more than once Zoomed in from another continent.

I thank my collaborators, labmates and friends, who have made grad school not only intellectually stimulating but also personally meaningful. Thank you for all the brainstorming, draft reading, shenanigans, long chats and hangouts—ultimately, for teaching me how to be a better researcher, colleague, and friend.

Finally, I thank my parents, my friends, and my partner for their love, patience, and support. This dissertation, and who I am today, would not have been possible without you. You are at once the harbor from which I sail and to which I return, and the lighthouse that guides my growth.

Chapter 1, is currently being prepared for submission for publication of the material. Liu, Xingyu S., McKenzie, Craig R.M., Winkielman, Piotr, McCullough, Michael E. The dissertation author was the primary researcher and author of this material.

Chapter 2, is currently being prepared for submission for publication of the material. Liu, Xingyu S.; McKenzie, Craig R.M.; Schachner, Adena. The dissertation author was the primary researcher and author of this material. The data in Study 2a were presented at the 2025 Annual Meeting of the Cognitive Science Society and published in the conference proceedings (see Liu et al., 2025).

Chapter 3, is currently being prepared for submission for publication of the material. Liu, Xingyu S.; McCauley, Thomas G.; McCullough, Michael E. The dissertation author was the primary researcher and author of this material.

VITA

2019 Bachelor of Arts in Psychology & Bachelor of Arts in French, Pomona College

2020 Master of Science by Research in Psychology, University of Edinburgh

2026 Doctor of Philosophy in Experimental Psychology, University of California San Diego

PUBLICATIONS

McKenzie, C. R. M., Sher, S., **Liu, X. S.,** & Kleiman-Lynch, L. J. (2025). When and why framing effects are neither errors nor mistakes. *Mind & Society*. https://doi.org/10.1007/s11299-025-00338-9

Liu, X.S., & McKenzie, C.R., & Schachner, A. (2025). When default options explain away preferences: a causal reasoning account of mental state reasoning from default options. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 47. https://escholarship.org/uc/item/9bq3d5ch

Stewart, R.D., Diaz, A., Hou, X., **Liu, X. S.,** Vainik, U., Johnson, W., & Mõttus, R. (2024). The ways of the world? Cross-sample replicability of personality trait-life outcome associations. *Journal of Research in Personality*, 112, 104515. https://doi.org/10.1016/j.jrp.2024.104515

Liu, X.S. (2021). What motivates us to go on strike? Social and moral norms and their impact on strike participation in individualist and collectivist situations in western cultures. *Journal of European Psychology Students*, 12(1), pp.1–15. http://doi.org/10.5334/jeps.507

ABSTRACT OF THE DISSERTATION

Through the looking glass of choice: Conflicting preferences, causal inferences, and social signaling

by

Xingyu (Shirley) Liu

Doctor of Philosophy in Experimental Psychology

University of California San Diego, 2026

Professor Craig R. M. McKenzie, Chair

In this dissertation, I investigate how people determine the meaning of their own preferences and choices, as well as those of others, especially when those choices are driven by multiple causes such as internal conflicts, social expectations, or concerns about social evaluation. I approach this question through three lines of research: (1) the psychology of second-order preference conflicts, (2) causal inferences about others' preferences, and (3) strategic signaling under social evaluation. Chapter 1 examines how choices can mask second-

order preference conflicts: conflicts between the preferences people have and the preferences they wish to have. I propose a psychological framework for understanding these conflicts and provide initial evidence for an empirical measure of these conflicts. Chapter 2 explores how people infer others' preferences from choices involving default options. I propose a causal reasoning account in which people reason backward from an observed choice to the plausible causes that led to that choice, whether they are preferences or alternative reasons from the choice context. I show that default options can provide such alternative explanations and explain away preferences. Chapter 3 investigates what people think their choices signal to others and how they use this knowledge strategically under social evaluation. I focus on reported self-control as a signal of cooperativeness and find that people inflate their reported self-control when observed by someone who may be a future cooperative partner. Across metacognitive, social, and functional perspectives, my work aims to shed light on how choices and preferences are a window to broader questions about who we are.

INTRODUCTION

When Alice walks through the looking glass, nothing is quite as it seems. What looks like a simple reflection turns out to be a world of its own, with many more layers and more complexity than what meets the eye. Studying the meaning of choices and their relation to preferences has always reminded me of Lewis Carroll's *Alice Through the Looking Glass*. We peer into choices hoping to see preferences clearly, but the relationship between the two is more complex than it appears. In this dissertation, I explore this complexity across three lines of research, each capturing a different way in which people make sense of their own and other people's choices.

Some choices are infinity mirrors that reflect multiple layers of preferences: they reflect not simply preferences, but also the preferences people prefer to have. In Chapter 1 of my dissertation, I explore the psychology of second-order preferences and their behavioral implications. People sometimes wish they had different preferences than they currently do. They may wish they liked healthier food, or they may wish they didn't desire to be unfaithful to their partners. These are second-order preference conflicts: metacognitive conflicts between the preferences people have and the preferences they wish to have. Although widely discussed in philosophy, their psychology is less understood. I bridge this gap by characterizing second-order preferences in terms of their cognitive bases — whether they are reflective or reflexive — and their motivational bases— whether they are driven by concerns about values, self-identity, or instrumental considerations. Understanding the psychology of second-order preference conflicts also involves distinguishing them from other forms of internal conflicts like self-control conflicts. I argue that whereas self-control conflicts are about managing behavior, second-order preference conflicts are about managing preferences themselves. Yet from choice behavior

alone, the two are difficult to tease apart. When a person declines a glass of wine, are they trying to exercise self-control –resisting the urge to drink because they need to drive – or are they trying to live up to a second-order preference –wishing not to like alcohol to begin with because they think it's debauched? While there are well-established measures for self-control conflicts, no empirical measure of second-order preference conflicts currently exists. As a first step toward filling this gap, I use desire for preference change rather than behavior change as a defining marker of second-order preferences, and investigate whether different subtypes of second-order preference conflicts influence this relationship. In a preregistered study ($N = 786$), I found that people with stronger second-order preference conflicts preferred to change their preference rather than their behavior across conflicts in sex- and food-related domains. This work brings psychological depth and empirical substance to the phenomenon of having unwanted preferences, establishing second-order preference conflicts as a distinct psychological construct.

If infinity mirrors describe choices that reflect meta-representations of the self, then one-way mirrors characterize choices that obscure mental states of others, such as their true preferences. Choices are often assumed to reflect preferences, but sometimes features in the choice architecture can provide alternative reasons for choice that explain away preferences. Therefore, observers often need to reason backwards from a choice to infer whether it reflects true preference or other pressures in the choice context. In Chapter 2, I investigate how people incorporate one such feature – default options – to reason about other people's preferences based on their choice. Default options can influence choices for reasons independent of preferences, such as following a default-setter's recommendations or cost avoidance. Thus, people often infer stronger preferences when someone actively switches away from a default than when they passively accept it, even when the choice outcomes are identical. I propose a causal reasoning

account based on Bayesian inference to formalize this process, and make novel predictions about when the asymmetry diminishes: asymmetric preference inferences occur when defaults provide strong plausible alternative explanations for choice that can explain away preference as a reason for choice. Across five preregistered studies ($N = 600$), I find that asymmetric preference inferences only emerged when defaults provided plausible alternative explanations across different types of defaults and default formats. Critically, when no such alternative explanations were available, the asymmetry diminished — a finding that has not been demonstrated in prior work. Together, these findings suggest that asymmetric preference inferences from defaults reflect a form of rational causal inference, and provide a framework for predicting when defaults will, and will not, shape how others' choices are interpreted.

In addition to infinity mirrors that look into the self and one-way mirrors through which we examine others, choices can also be vanity mirrors in front of which we pose for ourselves, leveraging our knowledge about how we appear to others. People are often aware of what their choices and behaviors signal to others, and may leverage this knowledge to appear socially desirable, especially when their reputation is at stake. In Chapter 3, I explore when and how people strategically signal to others, and whether this signal is honest. One such example is signaling cooperativeness with self-control: people who are seen as more self-controlled are also seen as more cooperative, making self-control a signal for cooperativeness. However, it is unclear whether more self-control actually predicts more cooperativeness, since both expressed self-control and cooperation behavior may covary with social evaluation. I experimentally manipulated social evaluation and its cooperative consequences to clarify this relationship, and investigate whether self-control is an honest signal of cooperation or a form of cheap talk in which people appear cooperative without actually behaving cooperatively. Across two

preregistered experiments ($N = 664$) and a Mega-analysis combining data from these experiments, participants reported greater self-control when they thought their self-control scores would be shared with someone who would evaluate their trustworthiness in an economic game, compared to when they thought their scores would be shared with a stranger or no one. However, the participants who inflated their reported self-control did not behave more cooperatively in the ensuing game. This suggests that people are sensitive to the reputational consequences of social evaluation, and inflate their reported self-control under social evaluation when the evaluation is consequential. In turn, this work addresses a methodological gap in prior research where the established association between self-control and cooperation is confounded by social evaluation. More broadly, it suggests that people are not just sensitive to being watched, but to being watched by an audience who matters for their future cooperation.

After all, Alice didn't walk through the looking glass to find the world exactly as it is — she was looking for a more nuanced way to see the world. Like Alice, I walk through the looking glass of choices and preferences to tell a more complex story: the closer we look, the more we find a world that is layered with internal conflicts and intersubjective reasoning, shaped by rich and informative choice architecture, and oftentimes staged for an audience who matters. Understanding this world helps us better understand what it means to want, to choose, and to be seen doing both.

Chapter 1 SECOND-ORDER PREFERENCES: ORIGINS AND IMPLICATIONS

Introduction

People sometimes wish that they had different desires and preferences than they currently do. Anna Karenina wished that she did not prefer passion over marital faith, Frankenstein's monster wished that he did not desire being human. In everyday life, a smoker might wish that they liked cigarettes less, and an aspiring musician might wish that they enjoyed composing more.

A person's existing preferences are their *first-order preferences* (P1s), and their preferences over these P1s are *second-order preferences* (P2s). Put more simply, P1s are preferences in the ordinary sense, and P2s are preferences over preferences. P1s are rankings over objects, people, or actions, and can be denoted as X > Y, whereas second-order preferences (P2s) are preferences over these rankings, and can be denoted as (Y>X) > (X>Y). For instance, an unwilling smoker may have a P1 for smoking over not smoking, Smoking > Not Smoking, but they may wish they didn't have this preference, and have a P2 for preferring not smoking instead, (Not smoking > Smoking) > (Smoking > Not smoking). P2s are conceptually similar to Frankfurt's (1971) second-order desires. Here, we use the term "preferences" rather than "desires," because it more naturally captures a broader range of psychological phenomena that are relevant to our framework, including desires, behavioral patterns, and evaluative judgements.

When a P1 conflicts with a P2, that is, when a person wishes to have a different preference than they currently do, they have a *P2 conflict*. P2 conflicts have mostly been discussed in philosophy, but less is known about their psychology. What are the cognitive bases and psychological motivations that give rise to P2 conflicts? How do P2 conflicts relate to similar internal conflicts, in particular, self-control conflicts? And how do we measure P2

conflicts empirically in a way that does not conflate them with self-control conflicts? The answers to these questions have important theoretical and empirical implications. Theoretically, studying the psychological origins of P2 conflicts can not only provide a more nuanced understanding of these metacognitive conflicts, but also suggest that diverse perspectives in the philosophical literature reflect P2 conflicts with different psychological features. Empirically, examining how people experience and seek to resolve P2 conflicts can clarify their structural distinction from self-control conflicts and their psychological meaning beyond their philosophical significance.

In the first part of this paper, we provide more nuanced, but important, distinctions between different subtypes of P1s and P2s based on their cognitive bases and psychological motivations, and re-examine P2 conflicts in the literature in light of these differences. In the second part of this paper, we investigate the relationship between P2 conflicts and self-control conflicts, and test whether P2 conflicts can be empirically measured as a distinct psychological construct from self-control conflicts.

Part I: Psychological origins of second-order preferences

The word "preference" is multifaceted. It can refer to patterns of choice — "the action of or an act of preferring or being preferred"— or to internal mental states— "a greater liking for one alternative over another or others; predilection" (Oxford English Dictionary, 2025). These different uses of the word reflect two common ways of conceptualizing preferences. On the one hand, preferences can be understood behaviorally as patterns of choice, a view formalized in the framework of revealed preference (Samuelson, 1938). On the other hand, preferences can be interpreted as internal mental states, such as "liking" or "predilection," which are not always directly reflected in behavior. These two ways of conceptualizing preferences apply to P1s and P2s in slightly different ways. P1s are preferences over objects and actions, and can therefore be

understood as either patterns of choice or states of liking and predilection. By contrast, P2s are preferences over preferences. They are therefore less directly tied to observable behavior, and are more naturally conceptualized as evaluative mental states directed at one's own preferences.

In what follows, we provide a psychological account of P2 conflicts by examining P1s and P2s along two dimensions: their cognitive bases and the psychological motivations that give rise to them. We first argue that both P1s and P2s can emerge from distinct cognitive bases — reflexive versus reflective — a distinction that has not been systematically developed in the existing literature. We then distinguish between P2s that are motivated by different reasons: instrumental considerations, values, and self-identity.

We use this framework to clarify how different subtypes of P1s and P2s can give rise to different forms of P2 conflicts, and how people might respond to these conflicts. Finally, we revisit existing perspectives on how P2 conflicts can be resolved, and argue that differences across these perspectives may stem in part from the fact that they focus on different subtypes of P1s and P2s.

Cognitive bases of P1s and P2s

Preferences can emerge from different cognitive processes. Some preferences emerge from *reflexive* processes that operate without conscious deliberation. Other preferences emerge from *reflective* processes that rely on conscious reasoning and deliberation. Most preferences, whether they are P1s or P2s, likely involve both reflective and reflexive processes. However, these processes take slightly different forms across P1s and P2s depending on whether they are directed at choice options themselves or at one's own preferences.

When reflexive processes are applied to P1s, they are often expressed as spontaneous inclinations toward choice options – whether they are objects, actions, or people. We will refer to P1s that mostly rely on reflexive processes as *reflexive P1s*. Typical reflexive P1s include

appetites for food or substances, automatic attraction or aversion toward people, concern about others' social evaluation, and moral intuitions about behaviors (Haidt, 2001). By contrast, when reflexive processes are applied to P2s, they often take the form of automatic judgements of one's P1s or internalized norms about what P1s are desirable. We will refer to P2s that mostly rely on reflexive processes as *reflexive P2s*. Typical reflexive P2s include internalized rejection of one's sexuality, or spontaneous moral discomfort about one's unwanted desires.

When reflective processes are applied to P1s, they tend to take the form of explicit evaluations of choice options. We will refer to P1s that mostly rely on reflective processes as *reflective P1s*. Typical reflective P1s involve deliberate comparisons across multiple attributes of choice options – such as the quality, the price, and the artistic statement of a fashion item. In comparison, when applied to P2s, reflective processes are mostly expressed as deliberate reasoning about one's P1s based on beliefs, values, and self-identity. We will refer to P2s that mostly rely on reflective processes as *reflective P2s.* Typical reflective P2s involve introspective reasoning about why one wishes to desire, or not desire, an object, a behavior, or a person.

Many paradigmatic cases of P2 conflict involve reflexive P1s and reflective P2s. Typically, the P1s that people want to change are spontaneously generated, such as unhealthy cravings for food, inappropriate sexual desires, or concerns about others' opinions. The corresponding P2s are often based on reflected evaluations, such as a dieter's wish to have healthier tastes in consideration of their health goals, a person's disgust at their sexual desire given their moral values, or someone's wish not to care so much about others' opinions because it feels inconsistent with their sense of self. Similarly, previous literature tends to characterize P1s as more reflexive and P2s as more reflective: P2s have been defined as endorsement of P1s

(Bruckner, 2011; Carballo, 2018), critical assessments of P1s (Stanovich, 2008), value-based rankings of P1s (Sen, 1977), or utility-based judgements about P1s (Bethel, 1980; George, 1998).

These common associations can create the misleading impression that P2s are more normatively authoritative than P1s. This impression resonates with Frankfurt's (1971) account in which second-order desires, rather than first-order desires, play the central role in agency and personhood. It can also make P2 conflicts appear superficially similar to self-control conflicts, which are typically described as conflicts between higher-order goals and lower-order impulses. We explore the relationship between these two types of conflicts in the second part of the paper.

However, the reflexive–reflective distinction is orthogonal to the distinction between P1s and P2s. The reflexive-reflective distinction compares two cognitive bases that give rise to a preference, whereas the P1-P2 distinction highlights two different objects of preference: P1s are directed at objects, actions, or people, whereas P2s are directed at other preferences. For instance, one can form a reflective P1 after deliberate consideration, such as preferring induction stoves to gas stoves for environmental reasons. Conversely, a person may have a reflexive P2 when they immediately feel shame, disgust, or alienation toward a desire before engaging in explicit reasoning about it. For example, a homosexual person may reflexively reject a same-sex attraction due to internalized homophobia. Thus, P2s are not reflective by definition, and P1s are not reflexive by definition. Rather, both P1s and P2s can be more reflexive or more reflective, although some combinations may be more common than others.

Reasons that motivate P2s

In addition to understanding how P1s and P2s arise cognitively, we are also interested in the reasons that motivate a person to have a P2. Here, we focus on the reasons that underlie P2s rather than P1s, because reasons play different roles in P1s and P2s: reasons are central to P2s but incidental to P1s.

P2s are evaluative in nature: they are metacognitive assessments of one's P1s against some standard, and therefore are grounded in reasons for endorsing or rejecting a P1. One cannot have a P2, whether reflective or reflexive, without some reason for deeming their P1 as desirable or undesirable. By contrast, although P1s can involve reasoning, as in the case of reflective P1s, reasons are not a requirement or core feature of P1s. One can simply prefer an object, an action, or a person without having a reason why.

Here, we distinguish between three broad types of reasons that can motivate P2s: instrumental reasons, which concern whether one's P1s are beneficial; value-based reasons, which concern whether one's P1s reflect one's values; and identity-based reasons, which concern whether one's P1s are consistent with one's sense of self.

Instrumental P2s arise from practical considerations about whether having different P1s than one's current P1s would make one's life better or worse overall. These P2s are concerned with the consequences of having certain preferences, rather than the intrinsic qualities of these preferences. For example, consider Charlie, a chocolate-lover on a calorie-restricted diet. Charlie may reflect on whether liking chocolate less would make their life easier and may conclude that they wish they liked chocolate less—forming an instrumental P2. In previous literature, P2s described as judgments about what makes one better off (e.g., Bethel, 1980; George, 1998) can be seen as instrumental P2s.

Instrumental P2s are typically reflective, as they often involve weighing the costs and benefits of having certain preferences, but they can be more reflexive when one's assessment of their existing preferences is more instinctive. Consider Will, a perfectionist writer facing a tight deadline. As the deadline approaches, Will may feel frustration with their desire to get

everything perfect even if they have not reflectively weighed the consequences of their perfectionism.

Value P2s arise from beliefs about whether one's existing P1s align with their values. They are driven by whether one's preferences reflect or embody what one finds important and worthwhile. In previous literature, P2s defined in terms of value-based evaluations can be considered as value P2s. For instance, Carballo's (2018) account of P2s as endorsement of or identification with P1s can be interpreted in this way. Similarly, Smith et al. (1989) claimed that "valuing is just desiring to desire" (p.116), linking P2s with values. Sen's (1977) notion of P2s as moral rankings of P1s also corresponds to value P2s, with an explicit emphasis on moral values. Value P2s can be both reflective or reflexive. For example, consider Lola, a pop-music fan, who, based on reflected artistic values, may wish to prefer classical music instead—forming a reflective value P2 based on deliberate reasoning. By contrast, consider Harry, who is homosexual but has been raised with homophobic values. Harry may have a spontaneous distaste of their same-sex attraction without having reflected on those values—forming a reflexive value P2.

Identity P2s arise from beliefs about whether certain preferences fit with one's self-identity. These preferences are driven by whether one's current preferences reflect who they are or who they want to be. In Korsgaard's (1996) words, we "endorse or reject our impulses by determining whether they are consistent with the ways in which we identify ourselves" (p.120). People may hold different beliefs about who they are. Some of these beliefs concern one's current self, which encompasses one's existing traits, values, roles, and commitments. Others concern a deeper sense of self — who one takes oneself to be, deep down, or who one aspires to become. The most psychologically interesting version of this deeper self is the true self, which

reflects who people think they really are, even when that self is not fully expressed in their current traits or behavior (Newman et al., 2014; Strohminger et al., 2017).

Identity P2s can be grounded in one's current self. Consider Max, a musician who likes to write crowd-pleasing songs that are commercially successful, even if these songs are less personally meaningful to them. Max may wish that they preferred writing music that is more personally meaningful so they can be truer to their musical identity, thus developing an identity P2 grounded in his current sense of self.

Identity P2s can be grounded in one's true self, which may differ from the current self: it can reflect traits and values that one does not yet possess (Strohminger et al., 2017), taking the form of an aspiration that may not yet be fully realized. In Callard's (2018) words, "aspiring is the method by which she works to become her true self" (p.144). The true self is often idealized — people tend to perceive it as good and morally virtuous, a tendency that is cross-culturally stable and holds regardless of whether one is evaluating themselves or others (De Freitas et al., 2018; Newman et al., 2014; Strohminger et al., 2017). This idealized quality connects the true self to Higgins' (1989) *ideal self* — a representation of "one's hopes, aspirations, or wishes" (p.321), and gives identity P2s their aspirational characteristic. Consider the Greek warrior, Alcibiades, who valued gallantry but, inspired by Socrates, wished to be someone who valued knowledge instead. Alcibiades can be seen as having an identity P2 that reflects the aspirations of his true self rather than his current sense of self (Callard, 2018).

These moral and idealized associations of the true self create an overlap between identity P2s and value P2s: when one's true self is defined in moral terms, wanting to be truer to oneself may seem equivalent to wanting to live up to one's moral values. Value P2s ask whether a preference is valuable; identity P2s ask whether a preference is "me." A person can wish they

preferred honesty because honesty is good, which would reflect a more of a value P2, or because being honest is who they really are, which would reflect more of an identity P2. These standards can coincide, but they need not. Crucially, identity P2s do not need to be morally grounded. Someone might wish they preferred solitude over socializing — not because solitude is morally superior, but because their sociable habits feel at odds with their introverted true self.

Finally, identity P2s can also be either reflective or reflexive. Alcibiades' identity P2 is reflective, grounded in explicit reflections about the type of person he aspires to be (Callard, 2018). By contrast, identity P2s are more reflexive when one's sense of self arises from non-deliberate intuition. Think of moments when a protagonist in a movie suddenly rejects their own selfish desires—not through explicit reasoning, but with an inarticulable sense of "this simply isn't me". These moments illustrate reflexive identity P2s.

Table 1.1 summarizes the subtypes of P1s and P2s we discussed based on their cognitive bases and the reasons that motivate P2s. These distinctions capture how people come to form preferences for their own preferences, while also shaping the types of conflicts that can arise between P1s and P2s.

Table 1.1: Subtypes of P1s and P2s.

Preference	Cognitive Basis	Reasons	Driven by
P1	Reflexive	—	Inherent taste, liking (e.g., craving chocolate)
P1	Reflective	—	Explicit considerations between options (e.g., choosing a car based on its features)
P2	Reflexive	Instrumental	Intuitive consequence-based judgements (e.g., Will the writer who wishes not to be a perfectionist when facing deadlines)
P2	Reflective	Instrumental	Deliberate consequence-based evaluations (e.g., Charlie the chocolate-lover who wishes to like chocolate less to make their diet easier)
P2	Reflexive	Value	Internalized values (e.g., Harry who wishes not to have same-sex attractions due to internalized homophobia)
P2	Reflective	Value	Endorsed values (e.g., Lola who wishes to prefer classical music to pop due to artistic values)
P2	Reflexive	Identity	Intuition that "this isn't me" (e.g., a movie protagonist's moment of self-realization)
P2	Reflective	Identity	Reflected sense of one's current self (e.g., Max the musician who wishes to write music that is true to themselves), or true self (e.g., Alcibiades the warrior who wishes to prefer knowledge to gallantry so they can be their aspirational self)

P2 conflicts

When a person wishes to have a different P1 than they currently do, they experience a P2 conflict. Different types of P2s give rise to different types of P2 conflicts. Some P2 conflicts are *intrinsic*: they concern the properties of P1s themselves, such as whether these preferences reflect one's values or self-identity, as in the case of value P2s and identity P2s. Because the source of these intrinsic conflicts lies in the P1s themselves, resolving them typically requires changes in the P1s themselves, or the evaluative standards that give rise to them. In comparison,

other P2 conflicts are *extrinsic*: they concern the constraints that P1s impose on one's goals, as in the case of instrumental P2s. In these conflicts, the P1s themselves are not necessarily problematic, but become problematic in light of their consequences or the contexts that give rise to them. Because the source of these extrinsic P2 conflicts lies in the consequences and contexts rather than the P1s themselves, they can be resolved without changing one's P1s, but by altering the external conditions that give rise to them.

Intrinsic conflicts

Value P2 conflicts arise when value P2s conflict with either reflective or reflexive P1s; *identity P2 conflicts* arise when identity P2s conflict with either reflective or reflexive P1s. Both types of P2 conflicts are typically intrinsic. Because value and identity P2s concern one's values and sense of self, these conflicts cannot be resolved solely through external changes, but instead require internal changes in one's preferences, values, or self-conception.

For example, recall Lola, who is conflicted between their obsession with pop music (reflexive P1) and their wish to enjoy classical music instead given their artistic values (value P2). This conflict can only be resolved if Lola alters their music taste or their artistic values. This is the case whether Lola's value P2 is based on deliberate endorsement (reflective) or internalized beliefs (reflexive).

A similar argument applies to identity P2 conflicts. These P2 conflicts highlight a mismatch between one's existing preferences and one's sense of self, or a discrepancy between one's current self and their ideal self (Higgins, 1989). Therefore, they can only be resolved if one changes their preferences or their sense of self to fit one another. Alcibiades the warrior is conflicted between their valuation of gallantry (reflexive P1) and their wish to cherish

knowledge instead because that is who they aspire to be (identity P2). Alcibiades can only resolve their conflict through changing their existing values or their sense of true self.

The same pattern holds when the P1 is reflective rather than reflexive. Consider Finn, who, after objective assessment of different car options, prefers to purchase a Lexus (reflective P1). However, Finn sees themselves as a "Ferrari person," which is part of their value system (conflicting value P2) and their self-conception (conflicting identity P2). Finn's conflict can only be resolved if they revise their car purchase assessment, or no longer sees themselves as a "Ferrari person".

Extrinsic conflicts

Instrumental P2 conflicts arise when instrumental P2s conflict with either reflective or reflexive P1s, and these conflicts are often extrinsic. Because instrumental P2 conflicts concern the consequences of having certain preferences, they can often be resolved through altering those consequences or the circumstances surrounding them, without changing the P1s themselves.

For example, recall Charlie, who has a conflict between their love for chocolate (reflexive P1) and their wish to like chocolate less (instrumental P2). This conflict can be managed through changing the circumstances or contexts that give rise to these conflicts. For instance, if Charlie had no access to chocolate or abandoned their diet, they would no longer need to refrain from chocolate, and would thus no longer wish to not like chocolate – their conflict would be resolved without Charlie needing to change their food preferences. This is the case regardless of whether Charlie's instrumental P2 is based on articulated reasons (reflective) or gut feelings (reflexive).

The same logic applies when the P1 is reflective rather than reflexive. Consider Ray, who prefers watching movies with higher overall ratings because they see ratings as reasonable

indicators of quality (reflective P1). However, Ray wishes that they didn't rely on ratings so much because then it would make their social life easier – Ray has some artsy friends who would mock them for their by-the-book movie preferences (conflicting instrumental P2). This conflict can also be managed through changes in circumstances (e.g., if Ray spent more time with less judgy friends) without Ray having to change their movie preferences. Table 1.2 summarizes the different subtypes of P2 conflicts we discussed. Note that whether P2 conflicts are intrinsic or extrinsic depends primarily on the reasons underlying the P2s, rather than on the cognitive bases of P1s and P2s. The cognitive bases of preferences instead shape how these conflicts are experienced—for example, whether they are immediate and intuitive or more deliberative and articulated. Extrinsic conflicts can be resolved through changes in behavior, or changes in preference, whichever is more easily achievable. Intrinsic conflicts are only resolved through changes in preference.

Table 1.2: Subtypes of P2 conflicts.

	Conflict Type	P2 Reasons	P1 Cognitive basis	Example
Instrumental P2 conflicts	Extrinsic	Instrumental	Reflexive	Charlie loves chocolate but wants to like it less because it would make their diet easier
			Reflective	Ray prefers highly rated movies but wishes that they didn't because it would make social life easier with their artsy judgy friends.
Value P2 conflicts	Intrinsic	Value	Reflexive	Lola likes pop music but wants to like classical music instead given their artistic values
			Reflective	Finn wants a Lexus after multi-faceted considerations, but wishes that they wanted a Ferrari instead because they value Ferraris
Identity P2 conflicts	Intrinsic	Identity	Reflexive	A movie protagonist instinctively rejects a desire as 'not me.'
			Reflective	Consider Finn again, except they don't just value Ferraris, they are a Ferrari *person*

Conflicts with reflexive preferences and higher-order preferences

So far, we have treated preferences as relatively stable. However, some preferences, especially those that people have not reflected on, may be subject to change upon reflection. In particular, reflexive preferences can sometimes be seen as "cocoon" preferences that can either develop into reflective preferences upon endorsement, or undergo change upon dis-endorsement.

This is especially likely when the reflexive preferences emerge from internalized values or beliefs that one has not critically examined, whether they are P1s or P2s.

For instance, consider Watson, a workaholic who has a reflexive P1 to prioritize work over leisure. Upon reflection, Watson may come to endorse this preference. In doing so, they not only form a P2 that endorses their P1, but their reflexive P1 may also become more reflective in nature—grounded in explicit reasoning and beliefs. Alternatively, Watson may not endorse their P1 upon reflection. In this case, they not only form a conflicting P2 but, over time, their reflexive P1 may shift towards a greater emphasis on work–life balance.

A similar process can apply to reflexive P2s that are driven by internalized values. In other words, these P2s arise automatically from values or beliefs that have not yet been critically examined. Upon further reflection, people may form preferences over these P2s—which we will refer to as *third-order preferences* (*P3s*). For example, imagine Paul, a puritan who is tempted by earthly pleasures, but automatically feels that he ought to prefer asceticism over pleasure. Paul's P2 is reflexive insofar as it arises from internalized religious commitments before he has explicitly reflected on whether he endorses them. Upon reflection, Paul may either endorse this reflexive P2 — transforming it into a reflective P2 — or reject it, forming a P3 that rejects his earlier P2 to prefer asceticism. Over time, this P3 may reshape his P2 into a new one that is more accepting of his pleasure-seeking disposition.

The formation of P3s, or of even *higher-order preferences*, does not introduce fundamentally new types of preferences. Rather, the formation of higher-order preferences reflects an iterative application of the same evaluative process with which people form P2s. Therefore, the same distinctions between instrumental, value-based, and identity-based motivations that drive P2s can also apply to P3s and higher-order preferences as people

iteratively reflect on what they want, what they want to want, what they wish they wanted to want, etc.

This opens the door to the possibility of an infinite regress of preferences. If people can form preferences over preferences at successive levels, when does this process stop? Frankfurt (1971) suggests that the regress can come to a halt when a person "identifies himself decisively with one of his first-order desires" (p. 322). That is, when a person fully commits to acting on a particular desire, whether it is the desire they currently have or the desire that they want to have, they no longer need higher-order evaluations of these desires. Cummings & Roskies (2020) describe P3s as judgments about whether to align one's actions with their P1s or P2s, and argue that P3s reflect a longing to make free agentic decisions. They further suggest that higher-order preferences beyond P3s, while possible, are not very psychologically realistic because these preferences would reflect even deeper existential longings.

Consistent with these views, we suggest that although higher-order preferences may arise when people reflect on their own preferences over preferences, they are unlikely to extend this process indefinitely. Each level of higher-order preference involves an additional layer of meta-representation of one's own preferences, which can quickly become cognitively demanding and less psychologically meaningful.

Deciding on a course of action

It is not always possible to resolve P2 conflicts, whether they are intrinsic or extrinsic. Sometimes, people may not be able to alter external circumstances or generate the internal changes needed to resolve their P2 conflicts. Nevertheless, they need to decide on what to do.

In a given P2 conflict, a person can typically act in one of two ways: they can either act on their P1, or they can *act on their P2*, that is, act as if they have their preferred preference,

even if their actual P1 has not changed. For instance, in a P2 conflict where the P1 is X > Y and the P2 is (Y>X) > (X>Y), acting on P1 means choosing X, and acting on P2 means acting as if one's P1 is Y > X, and hence choosing Y. The appropriate decision about what to do may depend on the subtype of P2 conflicts. Below, we outline possible reasonable approaches for different subtypes of P2 conflicts, and compare them with those proposed in the existing literature. Our goal is not to make normative claims about what one should do *per se*, but to explore whether different proposals in the literature may reflect different underlying types of P2 conflicts.

Instrumental P2 conflicts: Utility maximization

In instrumental P2 conflicts, a natural approach is to decide what makes one better off overall, given their existing P1s. Because these conflicts are about the overall consequences of having certain P1s, rather than intrinsic desirableness of the P1s themselves, the decision about what to do can be seen as a question about overall well-being. In this respect, this approach fits into a broader form of utility maximization, where utility is understood as overall well-being or satisfaction.

For instance, recall Charlie the chocolate-lover on a diet, who has an instrumental P2 for liking chocolate less. When deciding on whether to eat chocolate, Charlie can compare the value of dieting and the enjoyment they derive from chocolate, and decide whether acting on their P1 (eating chocolate) or acting on their P2 (abstaining) makes them better off. Note that what makes one better off may depend on the time frame considered. For example, when P1s for short-term benefits conflict with P2s for long-term benefits, maximizing utility at each moment may consistently favor acting on P1s, whereas taking into account longer-term goals may favor acting on P2s.

George's (1998) approach to P2 conflicts fits naturally within this utility-based framework. More specifically, he argues that when one cannot change their existing P1s, they are better off acting on it than not acting on it simply because of having a conflicting P2. As he says, "the agent having a preference to smoke is, *by definition,* better off smoking than not, regardless of whether or not his second-order preference is to have a different preference" (George, 1998, p.189). At the same time, George (1998) suggests that people can pre-emptively restrict their future choice sets so that they will not act on their unpreferred P1, but that this only improves overall well-being when doing so changes one's P1 itself at the time of choice. He illustrates this with the example of a smoker who has a P2 for not wanting to smoke, and who is deciding whether to throw away all their cigarettes before a party. If the smoker would still want to smoke even after throwing away the cigarettes, George (1998) argues, then preventing themselves from smoking at the party leaves them worse off, and they should not throw away their cigarettes. But if removing the cigarettes also changes the smoker's P1 so that they would no longer want to smoke at the party, then this pre-emptive restriction leaves them better off and they should throw away the cigarettes. In line with our account, George's (1998) discussion mainly concerns instrumental P2s, since he treats P2s as implying that one is "better off" having a different preference (p.183).

This utility-based approach may suggest an assumption that both the emergence of instrumental P2 conflicts, and the decision about what to do the face of these conflicts, are based on the same evaluative framework – where one is better off acting on their P1 at the time of choice, even if having a different P1 and acting on it may further improve their utility.

However, this assumption does not necessarily generalize to other types of P2 conflicts. Carballo (2018), whose account of P2s is more closely aligned with value P2s, argues that

expected utility does not allow room for a rational DM to have P2 conflicts to begin with. He argues that the formal structure of expected utility theory requires one's P1s and P2s to align: if a DM knows that they are a utility maximizer, they also know that they will do what they most prefer, which means that they will assign the same utility to choosing X and most preferring X. This means that the DM's P1 (X>Y) should always align with their P2 ((X > Y) > (Y>X)). This leaves no coherent space for P2 conflicts. To allow for P1s and P2s to conflict, he suggests, one may need to posit different "systems of preferences" (p.209) that do not assess one's P1s and preferred preferences on the basis of the same evaluative framework. As we show below, in value P2 conflicts and identity P2 conflicts, one's P1s and P2s may be based on different systems of values, or different ways of identifying with oneself.

Value P2 conflicts: Commit to values, act on P2s

Indeed, for value P2 conflicts, utility maximization may not be the best approach. Unlike instrumental P2 conflicts, value P2 conflicts are concerned with value-alignment rather than practical outcomes. In this case, the decision is not necessarily about choosing which action yields better outcomes, but about choosing which preference to commit to. A potential approach is to commit to one's values, and act on one's P2s.

For instance, recall Lola, who has a reflexive P1 for pop music, but a value P2 for preferring classical music instead given their artistic values. When deciding whether to listen to pop or classical music, Lola is not simply weighing the pros and cons of listening to one type of music over another. Rather, they are deciding on whether to act in accordance with a preference that they see as valuable. If Lola acts on their P2 and listens primarily to classical music, they are committing to their artistic values by cultivating their preferred musical taste, even if their P1 has not changed.

This perspective resonates with Sen (1977), who focused on conflicts between self-interested preferences and rankings of possible preferences based on one's moral values. Here, selfish preferences resemble reflexive P1s, and rankings of preferences based on one's moral values resemble value P2s with an emphasis on moral values. Sen (1977) argues that people should act in accordance with their highest-ranked preference. For instance, he suggests that people should act as if they prefer group welfare over selfish gain in a Prisoner's Dilemma, and always cooperate. Similarly, Callard (2018) argues that "the realization that it would be rational to gain or lose a value may not itself amount to value-change, but it could constitute an impetus to value-change" (p.187). In other words, the motivation to change one's preferences to align with certain values can provide a reason to act in ways that bring about that change.

However, acting on value P2s is not always justified. Sometimes, one's values may be shaped by problematic social views, or may have not been reflectively endorsed. Recall Harry, who is homosexual but has a value P2 for not having same-sex attraction because of internalized homophobia. In this case, it is not obvious that Harry should act on their value P2s before reflectively examining their homophobic values.

Identity P2 conflicts: Align with oneself, act on P2s

In identity P2 conflicts, one is concerned about whether their preferences reflect their sense of self. Therefore, the decision of what to do is neither about practical consequences, nor about cultivating values per se, but about acting in ways that most reflect who one is and wants to be. Here, like in value P2 conflicts, it is reasonable to align one's actions with one's sense of self, and act on one's P2s.

For instance, recall Max, the musician who has a P1 for writing crowd-pleasing songs, but an identity P2 for preferring to make music that is true to themselves, even if it means less

commercial success. When deciding on what type of music to write, Max is not simply thinking about their career outcomes, nor are they focused on what they find more valuable. Rather, they are deciding whether to act like the type of musician they see themselves as. If Max acts on their P2 and creates less popular but more personally meaningful work, they are aligning with their identity as a musician.

This perspective resonates with Frankfurt's (1971) account of second-order desires. For instance, he states that the unwilling addict "identifies himself through the formation of a second-order volition" (p.13), that is, through committing to act on the desire they want to have (i.e., not to take drugs). In other words, Frankfurt (1971) associates acting on P2s with identification. This perspective also resonates with Callard's (2018) account of aspiration, in which she argues that the aspiration to become one's true self provides sufficient justification to act in that way.

As with value P2s, it is not always obvious that one should act on their identity P2s, especially when one's sense of self or aspirations are distorted, or less reflected on. For instance, imagine Bonnie, who is peaceful at heart, but is raised in a cult that romanticizes violence. Bonnie has an identity P2 for preferring violence because that's the person they aspire to be, but Bonnie is not justified to act on their identity P2 without further deliberation.

Conflicts involving reflexive P2s: Reflect!

In these situations where one may not be justified to act on their value or identity P2s, it could be that they come from a value system that is typically regarded as unjustified, or it could be that these P2s are reflexive "cocoon" preferences that require further deliberation. When people have reflexive P2s—especially those that they have not introspectively examined or that conflict with broader social norms—one course of action may be to reflect on them first.

Through such reflection, these “cocoon” preferences can transform into reflective preferences that one can more fully endorse.

It is not obvious how one should know that further deliberation is needed. Because reflexive preferences often feel self-evident, and because people often have limited access to the causes of their own preferences (Wilson & Dunn, 2004), they may not spontaneously see reflexive preferences as objects of reflection. Instead, the need for reflection may arise from other forms of value or social conflicts such as moral dumbfounding (Haidt, 2001) – where one has trouble explaining their moral intuitions – or negative social feedback. When one cannot articulate the reasons for a reflexive moral P2s, similar to Haidt's (2001) participants who could not rationally explain their disgust for consensual incest, this may be a cue that the P2s themselves warrant further examination. Likewise, consistent negative social feedback about one’s P2s may reveal a gap between one’s self-perception and others’ perception of them, prompting reflection on whether one truly endorses their beliefs. This is consistent with Wilson and Dunn’s (2004) suggestion that people can gain self-knowledge by considering how others view them. Therefore, P2 conflicts involving reflexive preferences may not always elicit reflection on their own, but other forms internal or external friction may nudge a person to reflect on their P2s.

Reflexive P2s are rarely discussed in the existing literature. Bruckner (2011) can be interpreted as allowing for reflexive P2s, describing P2s as “not necessarily endorsement[s] that [are] adequately reflective” (p. 377). However, rather than suggesting that these preferences require further reflection, Bruckner (2011) proposes an alternative approach: to treat P1s and P2s as competing preferences, rank them based on their strength, and act on the highest-ranked option.

This rank-order approach is inspired by Jeffrey (1974), although it is unclear what subtype of P2 he had in mind. In this rank-order approach, both Jeffrey (1974) and Bruckner (2011) suggest that if one's P1 is ranked higher than their P2 (Figure 1.1a), one should act on their P1; if one's P2 is ranked higher than their P1 (Figure 1.1b), one should act on their P2.

a.
S
$\sim(S \text{ pref } \sim S)$
$S \text{ pref } \sim S$
$\sim S$

b.
$\sim(S \text{ pref } \sim S)$
S
$\sim S$
$S \text{ pref } \sim S$

Figure 1.1: Ranking of P1s and P2s from (Jeffrey, 1974). *S* = smoking; ~*S* = Not smoking; *S pref* ~ *S* = Smoking > Not smoking; *~(S pref ~S)* = (Not smoking > smoking) > (Smoking > Not smoking). *S* ranked above ~*S* = a P1 for Smoking > Not smoking; *~(S pref ~S)* ranked above *S pref ~S* = P2 for (Not smoking > smoking) > (Smoking > Not smoking). Distance between the ranked objects = strength of preference (e.g., *~(S pref ~S)* is ranked further above *S pref ~S* than *S* is above *S pref ~S*, indicating a stronger preference for not preferring to smoke than for smoking)

Although the rank-ordering of P1s and P2s may seem similar to forming higher-order preference like P3s, the two approaches differ in an important way. P3s are preferences over P2s, and therefore compare preferences only at the second order—one P2 against another P2. By contrast, Jeffrey (1974) and Bruckner (2011) "level" P1s and P2s on a single scale, comparing different orders of preferences directly. We argue that there is little justification for leveling P1s and P2s in this way because they have qualitatively different objects: P1s concern *objects, actions, people, or outcomes in the world,* whereas P2s concern *one's own preferences and motivational states.* Therefore, these preferences elicit structurally different relationships to the self and the world. P1s represent what one wants in relation to the world, and P2s represent one's relation to their own desires, and oftentimes, themselves. It follows that P1s and P2s serve different psychological functions, and directly comparing them on the same scale risks treating these structurally different psychological states as interchangeable.

In sum, the diverse perspectives in the literature on what to do in the face of P2 conflicts—acting on P1s, acting on P2s, reflecting on one's preferences, or rank-ordering them—may not reflect competing perspectives to the same types of P2 conflicts, but rather responses to different types of P2 conflicts.

Part II: P2 conflicts and self-control conflicts

The cognitive bases and psychological motivations that characterize P2 conflicts bring to mind another type of internal conflict: self-control conflicts, which occur when "personally valued longer-term goals" clash with "conflicting urges that are more potent in the moment" (Duckworth & Steinberg, 2015). P2 conflicts can resemble self-control conflicts in psychological characteristics and behavioral outcomes, but the relationship between the two is rarely explored. Are they the same type of conflict dressed in different words, subsets of each other, or distinct types of conflicts? We argue that they are distinct types of conflicts that can often occur at the same time — and that to establish this distinction, we need an empirical measure of P2 conflicts that can distinguish from self-control conflicts as its own psychological construct.

Consider the following example. In an interview, Jimmy Carter disclosed that he had "looked on a lot of women with lust" and "committed adultery in my heart many times," but believed that "God forgives" him because he did not act on his fantasies (Carter, 1976). Carter's conflict is a self-control conflict: he feels the temptations of lust and resists acting on it, but he does not seem to reject the lusting itself, even if he considers it a sin—he disclosed: "I try not to commit a deliberate sin. I recognize that I'm going to do it anyhow, because I'm human and I'm tempted." Now consider a faithful husband who feels distressed about desiring another woman, even if he does not act on his desire. Both men face temptation and refrain from committing

adultery — but their conflicts feel different in an important way. Carter is mainly concerned about his actions; but the faithful husband is also concerned about his desire.

Structural difference

This distinction is precisely what separates self-control conflicts from P2 conflicts. In a self-control conflict, the objects of conflict are two behavioral motivations — to X versus to Y — both are about what to do. P2 conflicts have a different structure. The objects of conflict are two preferences, where one takes the other as its object: a first-order preference (X > Y) conflicts with a meta-preference (Y > X) > (X > Y), which rejects it. This recursive structure is what makes P2 conflicts about preferences rather than behaviors.

This structural distinction resonates with Frankfurt's (1971) concept of second-order desires. Frankfurt's unwilling addict has a first-order desire to take the drug, but a second-order desire not to have that desire — "it is the latter desire, and not the former, that he wants to constitute his will" (p.12). The addict's conflict is not simply about what to do; it is about what to want. We argue that P2 conflicts and self-control conflicts are distinct types of conflicts, but they are empirically difficult to tease apart.

Psychological similarities

Despite their structural difference, the two conflicts can evoke similar phenomenological experiences — not because they are the same, but because they share overlapping psychological features. Understanding these similarities can help explain why these conflicts are difficult to distinguish empirically.

One main account of self-control conflicts describes a push-and-pull between two opposing forces: one pulls toward immediate desire, while another seeks to regulate it. Successful self-control has canonically been characterized as the triumph of the regulatory force — historically known as "willpower" or "ego" — over immediate desires (Hoch & Loewenstein,

1991; Holton, 2009; Rachlin, 2000; Schechtman, 2004). P2 conflicts engage a similar phenomenology: an existing desire pulling in one direction, a meta-preference pushing in another. A second account explains self-control through temporal discounting — the phenomenon in which "valuation of a future reward decreases as the delay to that reward increases" (Doyle, 2012; Green & Myerson, 2004, 2004; Rachlin, 2000). This utility-based framework can resemble some instrumental P2 conflicts, especially when people wish they had desires that aligned with their long-term interests (e.g., enjoying healthy food) rather than desires that indulge short-term benefits (e.g., enjoying sweets). Furthermore, both accounts share a hierarchical structure in which the conflict occurs between one force associated with more authority and identification with the self, regulating or overriding the other force.

Thus, it might be tempting to conclude that the two conflicts are simply variations on the same theme. But the key distinction is not whether one motivation is more authoritative, more long-term-oriented, or more self-associated than the other — both conflicts can involve all of these features. What matters is whether the conflict includes a preference about the desire itself. Recall that Carter's conflict, even read charitably, does not seem to include a rejection of his desire, but this rejection is the center of the faithful husband's conflict.

Co-occurrence

In addition to their psychological similarities, P2 conflicts and self-control conflicts tend to co-occur because preferences and choices are inherently linked. Having a P2 conflict about one's desires almost inevitably places a person, at some point, in situations where they must act on or against that desire — situations that in turn give rise to self-control conflicts. Frankfurt's (1971) unwilling addict is an illustrative example. The addict faces a self-control conflict: "he wants to take the drug, and he also wants to refrain from taking it" (p.12). At the same time, he

has a P2 conflict: "it is the latter desire, and not the former, that he wants to constitute his will" (p.12).

However, co-occurrence does not mean that the two types of conflicts merge into one. After all, their structural difference remains: one is about what to do, the other about what to want. This distinction is evidenced by situations in which P2 conflicts and self-control conflicts empirically come apart. In Table 1.3, we illustrate the different combinations of P2 conflicts and self-control conflicts – when they co-occur and come apart—with a toy example of a chocolate lover. The bottom left cell illustrates a person who wants to eat less chocolate but has no wish to like chocolate less — they have a self-control conflict without a P2 conflict. For a more mundane example, consider a student who needs to resist going to parties before exams (self-control conflicts) but does not want to like partying less (no P2 conflict). The top right cell illustrates a person who wants to like chocolate less but does not wish to eat it less – they have a P2 conflict without a self-control conflict. For another example, consider a person who is in love with a fictional character and wishes that they were not (P2 conflict), but does not have a problem with acting on literary fantasies (no self-control conflict). The top left cell illustrates a person who likes chocolate and eats chocolate with no conflict, and the bottom right cell illustrates a person who wishes that they ate less chocolate and liked it less –a typical example in which P2 conflicts and self-control conflicts co-occur. Together, these examples highlight that despite their co-occurrence, P2 conflicts and self-control conflicts are distinct conflicts that can occur independently of each other.

Table 1.3: Combinations of self-control conflicts and P2 conflicts

	No P2 conflict	**P2 conflict**
No Self-control conflict	I like chocolate, eat it when I want, and have no problem with wanting it.	I like chocolate and wish I didn't enjoy it so much, but I do not have a problem with how much chocolate I actually eat.
Self-control conflict	I like chocolate and often need to resist eating it, but I do not want to enjoy chocolate less.	I like chocolate, often need to resist eating it, and also wish I did not enjoy chocolate in the first place.

Empirical challenge

The aforementioned co-occurrence creates an empirical challenge for measuring P2 conflicts as a distinct construct from self-control conflicts. There is an abundance of empirical measures for self-control conflicts – such as temporal discounting tasks –and they are mostly designed to capture behavioral patterns. In contrast, there is little empirical work on the psychology of P2 conflicts, which have mostly been discussed at a conceptual level, let alone a measure of P2 conflicts that focuses on preferences. Related work on ambivalent attitudes has examined the psychological experience and behavioral correlates of internal conflicts in general (Priester & Petty, 2001), but it does not distinguish between structurally different sources of conflict. Both self-control conflicts and P2 conflicts can produce internal conflict and ambivalence — but these accounts do not capture the fundamental difference in their structure and object: one is about what one wants to do, the other about what one wants to want.

One way to get an empirical handle on P2 conflicts is to examine how people prefer to resolve them — through changes in desires or in behaviors. This question is related to, but different from, the distinction we made in Part I between intrinsic and extrinsic P2 conflicts. There, we used the question of conflict resolution to illustrate a theoretical distinction: whether a

P2 conflict arises from internal sources such as preferences and values, or from external circumstances that make the preference problematic. Here, the focus is descriptive: do people prefer to resolve their P2 conflict by changing what they do or what they want. If P2 conflicts are genuinely about managing preferences rather than behaviors, they should be empirically associated with a preference for desire change over behavior change. However, simply measuring the association between a reported P2 conflict and a preference for desire change is not very informative on its own, since some amount of preference for desire change is almost by definition entailed by identifying P2 conflict. A more diagnostic test of this is to examine whether stronger P2 conflicts predict a greater preference for desire change over behavior change. This is a first step to substantiating P2 conflicts as a psychologically real and measurable construct of its own right.

This relationship may not be the same across all subtypes of P2 conflicts, though. People with instrumental P2 conflicts, which often arise from a need to achieve a behavioral goal, may prefer behavior change to desire change. If changing one's desires is a means to changing one's behavior, then behavior change would get rid of the need for desire change to begin with – if Charlie the chocolate lover can easily resist chocolate, they can enjoy it without worrying about their diet. By contrast, value and identity P2 conflicts are more closely tied to desire change by definition: mere behavior change cannot resolve a conflict about what one wants to want. At the same time, these reasons that drive P2 conflicts are not mutually exclusive. A pedophile may reject their sexual desire because it is practically harmful, morally disturbing, and inconsistent with who they are all at once, and behavior change alone leaves many aspects of the conflict unresolved. Thus, while P2 conflicts are about preferences, the relationship between conflict strength and preference for desire change may vary depending on the subtype of P2 conflict.

Different subtypes of P2 conflicts may be more or less likely to emerge in different domains. For instance, food-related P2 conflicts are probably more instrumental in nature, since unwanted food desires are often acted on and the conflict typically centers on behavioral habits and dieting goals. Sex-related P2 conflicts, by contrast, probably carry more moral and identity weight, since unwanted sexual desires tend to have deeper implications for who we are and how we want to live. To explore domain differences in subtypes of P2 conflicts, we first needed to identify domains in which people naturally experience P2 conflicts. Importantly, if people spontaneously think of conflicts in different domains when asked to describe different subtypes of P2 conflicts, that itself serves as evidence that the subtypes map onto people's intuitions rather than being purely theoretical distinctions. We therefore conducted a pilot survey:

We asked 60 Prolific participants to provide at least one example of things they *do*, *desire*, or *care about* that they wish they didn't — or at least not as much. These prompts were intended to approximate different ways of describing P2 conflicts in everyday language: things people "do" may be more closely tied to behavior and therefore instrumental P2s; things people "care about" — Frankfurt's own term for desires people identify with (Frankfurt, 1992) — may be more closely tied to self-identity; and things people "desire" may apply more broadly to P2s in general. While value P2s may not be specifically tied to any one description, they can occur across all three because people can have value-based concerns about what they do, what they desire, and what they care about.

We manually coded the responses into different categories. The prompts elicited partially distinct response patterns, suggesting that participants can intuitively distinguish between the subtypes of P2 conflicts we have identified. "Do" responses most often involved behavioral or mental habits such as negative emotional states, procrastination, unhealthy diet, and substance

use. "Desire" responses more often included acquisitions or possessions, unhealthy diet, sex, and sports. "Care about" responses were especially likely to involve others' opinions, concern for others, and appearance. Figure 1.2 summarizes the main categories that are unique to each prompt. The full prompt wording, coding scheme, and response-category breakdown are reported in the Supplemental Materials.

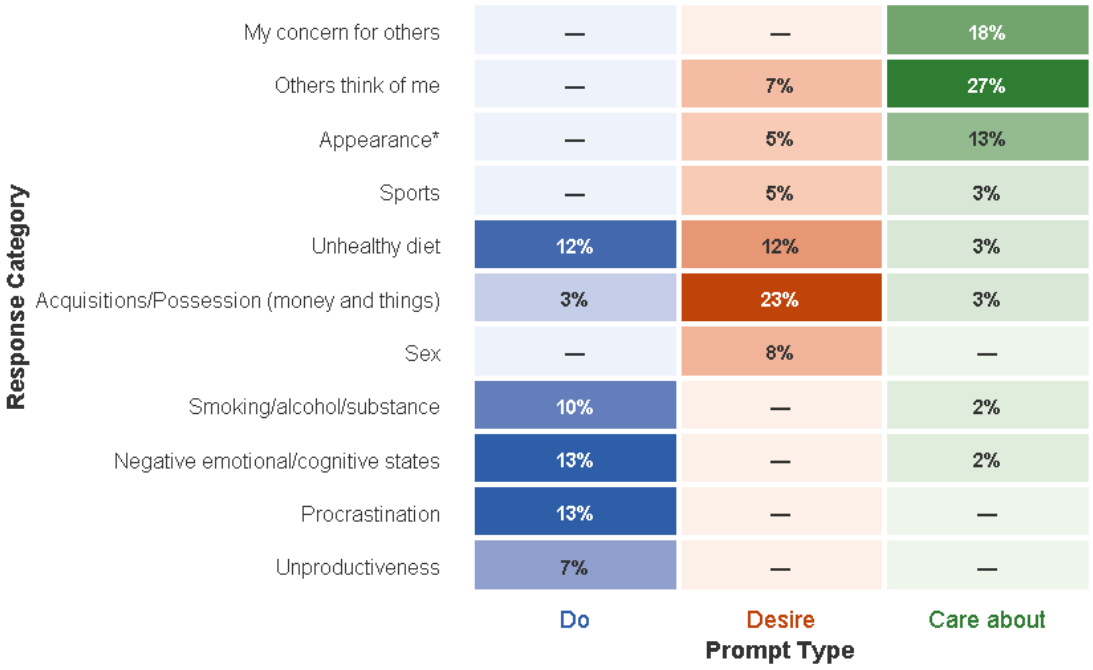


Figure 1.2: Common response categories by prompt type in the pilot survey. Percentages are out of N = 60. Categories appearing fewer than three times across all prompts and idiosyncratic responses are excluded; see Supplemental Materials for the full breakdown.

Notably, as expected, food-related P2 conflicts occurred more in conflicts of behavioral patterns, suggesting their instrumental nature. By contrast, sex-related conflicts occurred more in conflicts of desires per se. Therefore, in the current study, we focused on food- and sex-related domains as a starting point, because they appeared frequently in participants' responses and because they may differ in the extent to which they relate to different subtypes of P2 conflicts.

These two domains are not meant to be representative or comprehensive, but they offer a starting point for exploring how the psychology of P2 conflicts varies across different areas of desires.

Current Study

The current study examines the link between P2 conflicts and preference for desire change over behavior change, and the factors that shape this relationship. Do stronger P2 conflicts predict a greater preference for desire change over behavior change? Does this association vary based on the extent to which the conflict is instrumental, value-driven, or identity-driven? Here, because value P2 conflicts can take many forms, we focused specifically on moral values as one important and tractable instance, consistent with prior work on the role of morality in shaping preferences and desires (e.g., Sen, 1977). Finally, do the subtypes of P2 conflicts differ across food- and sex-related domains — and if so, how?

Methods

Participants

Participants were recruited from www.prolific.com. A total of 786 participants were included in the analyses (349 men, 410 women, 22 non-binary gender, and 5 preferred not to say; M_{Age} = 38.36, SD_{Age}= 11.88). Of these participants, 73.7% (N =579) completed the study in the Food domain and 26.3% (N = 207) in the Sex domain.

We aimed for N = 713 to detect a small incremental effect of the three P2 subtype predictors above P2 Strength, Frequency, and Condition, f^2 = .02, with α = .05 and 90% power. We oversampled to ensure a sufficient final sample size, anticipating that some participants would not meet the preregistered inclusion criteria described below.

As preregistered, participants were included if they reported having a behavior that they engaged in but wished they engaged in less, and also wished that they had less of the desire to engage in that behavior. A total of 181 participants were screened out for not meeting these

criteria. An additional 14 participants were excluded based on the preregistered exclusion criterion of failing Prolific's bot authenticity checks.

Screening and condition assignment

In the Prolific study description, we advertised for participants "with unwanted desires, related to food or sex, that they have acted on at least once but wish they did not act on." We also emphasized that participants would not be asked to disclose what their desire was about.

Participants were initially randomly assigned to one of two between-subjects conditions (domain: Food vs Sex). They were first asked whether they could think of a behavior in their assigned domain that they engaged in but wished they did not. If so, they indicated whether they also wished they engaged in this behavior less and wished they desired it less. Participants proceeded with the study in their assigned condition if they had such a behavior in mind and met both criteria. If they did not have such a behavior in mind, or if they met only one of the two criteria, they were reassigned to the other condition and answered the same screening questions. Participants who failed to meet both criteria in either condition were excluded from the study and received minimum compensation for their time. Figure 1.3 illustrates the screening and exclusion process.

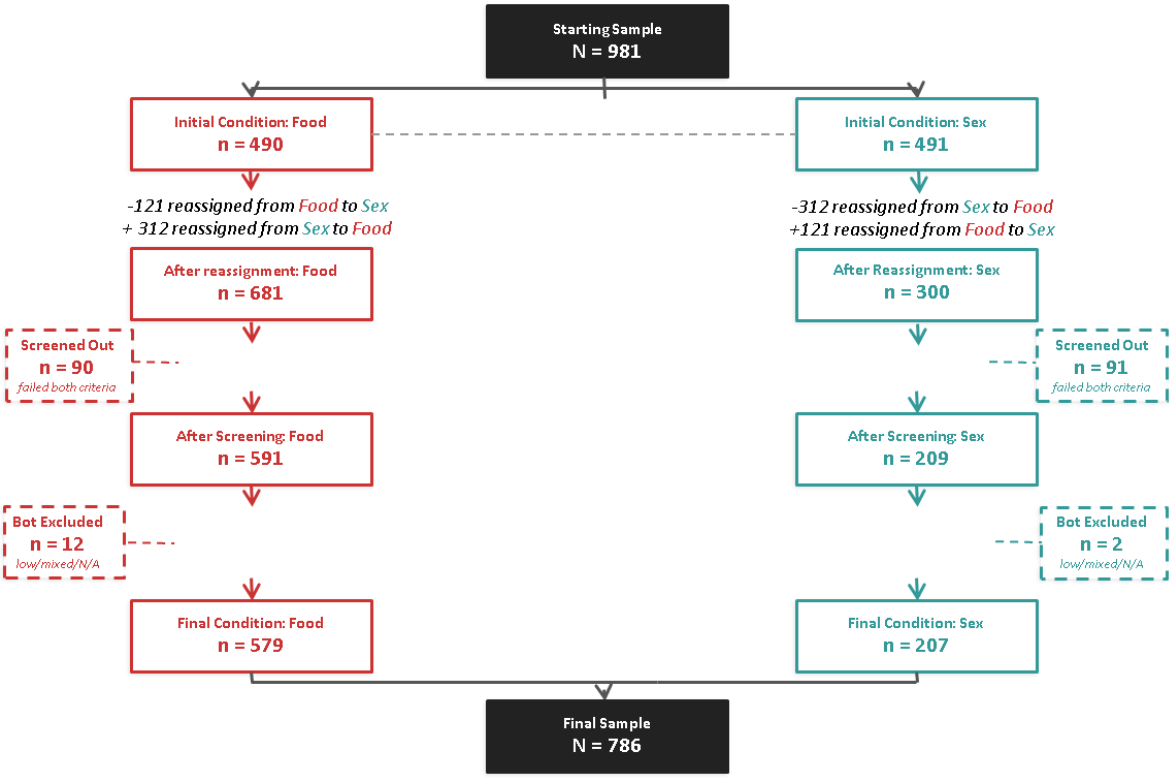


Figure 1.3: Screening and exclusion process based on preregistered criteria.

Procedure and measures

Participants were asked to keep in mind the specific desire/behavior they had identified while answering the following questions.

P2 Strength. Participants indicated the strength of their P2 conflict by rating the extent to which they wished not to have the relevant desire ("To what extent do you wish to not have this desire, or at least not as much?") Ratings were made on a scale from 1 (*slightly*) to 7 (*strongly*).

Behavioral Frequency. Participants reported how frequently they acted on the relevant desire ("When you have the opportunity to act on this desire, how often do you do so?"). Ratings were made on a scale from 1 (*almost never*) to 7 (*almost always*).

Subtypes of P2s. Participants indicated the extent to which their P2 conflict was driven by each subtype of P2: Identity P2, Moral P2, and Instrumental P2. Participants rated their agreement with each statement on a scale from 1 (*completely disagree*) to 7 (*completely agree*).

***Identity P2*:** "I wish I didn't have this desire, or at least not as much, because having this desire is not who I *really* am."

***Moral P2*:** "I wish I didn't have this desire, or at least not as much, because having this desire is morally wrong."

***Instrumental P2*:** "I wish I didn't have this desire, or at least not as much, because having this desire has negative practical consequences for me."

Treatment Preference. Participants then imagined two hypothetical treatments targeting their food-related (Food condition) or sex-related (Sex condition) desires and behaviors. The Behavior Modification Treatment targeted what participants did without influencing what they wanted. The Desire Modification Treatment directly targeted what participants wanted. The order in which the two treatments were presented was counterbalanced across participants. Participants were also told that both treatments were effective long-term, reversible, and had no negative side effects. The descriptions with the Behavior Modification Treatment presented first is shown below:

*The **Behavior Modification Treatment** targets what you do, without influencing what you want. With this treatment, you will be much less likely to act on the food- (sex-) related desire that you've been thinking about here (but the desire itself will not change).*

*The **Desire Modification Treatment** directly targets what you want. With this treatment, the food- (sex-) related desire you've been thinking about here will be greatly reduced (and as a result, you will also be much less likely to act on it).*

After reading both descriptions, participants indicated their relative preference for the two treatments on a 7-point scale with counterbalanced endpoints: For participants who read the Behavior Modification Treatment first, the scale ranged from "definitely prefer the Desire Modification Treatment" to definitely prefer the Behavior Modification Treatment, with 0 indicating "no preference". The endpoints were reversed for the participants who read the Desire Modification Treatment first.

Because P2 conflicts are defined as conflicts about one's preferences, relative preference for Desire Modification over Behavior Modification served as our key measure of whether participants preferred to resolve the conflict by changing the desire itself rather than only changing their behavior.

Results

P2 Strength

P2 Strength was generally high (M = 5.55, SD = 1.27) but differed across conditions. P2 strength was significantly higher in the Food condition (M = 5.63, SD = 1.22) than in the Sex condition (M = 5.33, SD = 1.37), t(330.84) = 2.83, *95% CI* [0.09, 0.52], p = .005.

Behavioral Frequency

Behavioral Frequency was also high in general (M = 4.79, SD = 1.34). Preregistered comparison showed that participants acted on their relevant desire significantly more often in the Food condition (M = 4.87, SD = 1.28) than in the Sex condition (M = 4.56, SD = 1.48), t(322) = 2.65, 95% CI [0.08, 0.53], p = .008.

Subtypes of P2s

As preregistered, we examined whether participants' agreement ratings for each subtype of P2 conflicts differed by Condition. Participants were more likely to agree with having Identity P2s in the Sex condition (M = 5.01, SD = 1.53) than in the Food condition (M = 4.03, SD = 1.65),

t (387.82) = −7.75, *95% CI* [−1.23, −0.73], *p* < .001. They were also significantly more likely to agree with having Moral P2s in the Sex domain (*M* = 4.11, *SD* = 2.15) than in the Food domain (*M* = 2.28, *SD* = 1.59), *t* (290.40) = −11.22, *95% CI* [−2.15, −1.51], *p* < .001. By contrast, they were significantly more likely to agree with having Instrumental P2s in the Food condition (*M* = 5.68, *SD* = 1.35) than in the Sex condition (*M* = 4.61, *SD* = 1.92), *t*(282.64) = 7.42, 95% CI [0.79, 1.36], *p* < .001. Figure 1.4 shows mean ratings for each P2 subtype across Conditions.

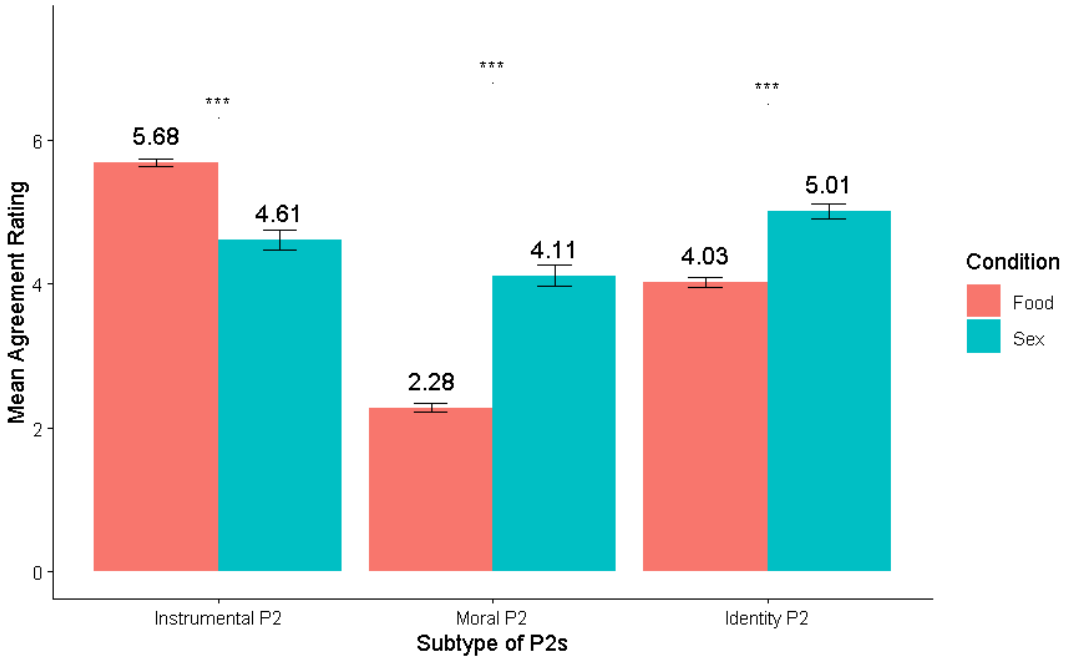


Figure 1.4: Subtypes of P2s across the Food and Sex conditions. Higher ratings indicate stronger agreement with statements describing each subtype of P2 conflict. Error bars represent ±1 SE.

In an exploratory analysis, we further examined correlations among the three P2 subtypes. Across the two conditions, Identity P2 and Moral P2 were moderately positively correlated (*r* = .42, *p* < .001). This pattern was robust in both conditions. By contrast, Instrumental P2 was only weakly and non-significantly correlated with Identity P2 and Moral P2 (*ps* > .208) in general, but it was weakly and significantly correlated with Identity P2 in the Food condition, and moderately and significantly correlated with both Identity P2 and Moral P2in the

Sex condition. In particular, Instrumental P2 was significantly correlated with both Identity P2 (r = .19, p < .001) and Moral P2 (r = .28, p < .001) in the Sex condition, suggesting that practical concerns about sex-related desires may be more closely tied to identity and moral concerns. Full correlations are reported in Tables 1.4 and 1.5.

Overall, these findings suggest that the P2 subtypes are related but distinguishable constructs, and that their associations may differ across domains. As predicted, P2 conflicts are more likely to be driven by instrumental concerns in the food domain, and by moral or self-identity concerns in the sex domain. Furthermore, across both domains, P2 conflicts that are driven by moral values also tend to be driven by self-identity; in the sex domain in particular, P2 conflicts that are driven by instrumental concerns also tend to be driven by moral values.

Table 1.4: Intercorrelations among P2 subtypes in the overall sample. Values are Pearson's r. Only the lower triangle is reported because the matrix is symmetric. N = 786. *** p < .001.

	Identity P2	**Moral P2**	**Instrumental P2**
Identity P2	—		
Moral P2	.42***	—	
Instrumental P2	.04	.01	—

Table 1.5: Intercorrelations among P2 subtypes by condition. Values are Pearson's r. Only the lower triangle is reported because the matrix is symmetric. Food condition n = 579; Sex condition n = 207. † p = .055. ** p < .01. *** p < .001.

	Food			**Sex**		
	Identity P2	**Moral P2**	**Instrumental P2**	**Identity P2**	**Moral P2**	**Instrumental P2**
Identity P2	—			—		
Moral P2	.35***	—		.38***	—	
Instrumental P2	.11**	.08†	—	.19**	.28***	—

Treatment Preference

Treatment Preference was rated on a scale from - 3 to 3, where lower scores indicated stronger preference for Desire Modification and higher scores indicated stronger preference for

Behavior Modification. Overall, participants preferred Desire Modification over Behavior Modification ($M = -0.97$, $SD = 2.06$). This relative preference differed significantly from the midpoint of no preference, $t(785) = -13.25$, 95% CI [−1.12, −0.83], $p < .001$.

This relative preference emerged in both Conditions. In the Food condition, Treatment Preference scores were significantly below the no-preference midpoint (0), indicating preference for Desire Modification over Behavior Modification ($M = -0.94$, $SD = 2.04$), $t(578) = -11.09$, 95% CI [−1.11, −0.77], $p < .001$. The same pattern emerged in the Sex condition ($M = -1.07$, $SD = 2.12$), $t(206) = -7.25$, 95% CI [−1.36, −0.78], $p < .001$. Treatment Preference did not significantly differ by Condition, $t(351.78) = 0.74$, 95% CI [−0.21, 0.46], $p = .458$, or by the order in which the two treatments were presented, $t(762.09) = -1.81$, 95% CI [−0.56, 0.02], $p = .071$.

Does P2 strength predict relative preference for desire change?

We first explored the raw correlation between P2 Strength and Treatment Preference. P2 Strength was negatively and significantly correlated with Treatment Preference overall, $r(784) = -.11$, 95% CI [−.18, −.04], $p = .002$, indicating that stronger P2 conflict was associated with greater relative preference for Desire Modification over Behavior Modification. This association was significant in the Food condition, $r(577) = -.16$, 95% CI [−.24, −.08], $p < .001$, but not in the Sex condition, $r(205) = -.01$, 95% CI [−.14, .13], $p = .935$. Figure 1.5 illustrates these associations by Condition.

These findings provide evidence that stronger P2 conflict predicted greater preference for Desire Modification over Behavior Modification, supporting the theorized association between P2 conflicts and preference for desire change over behavior change.

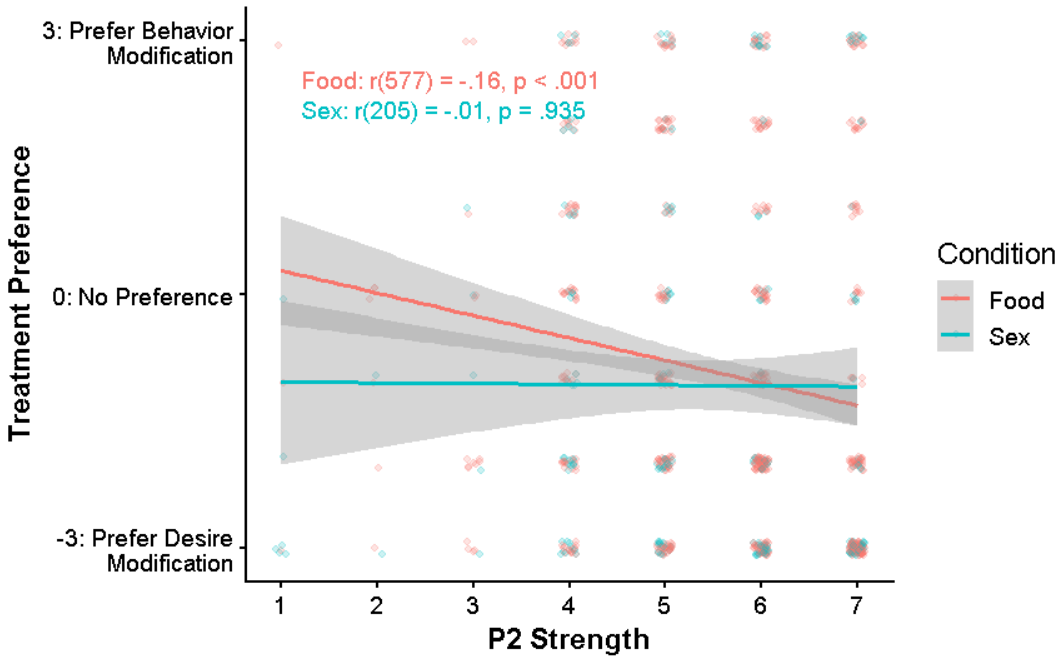


Figure 1.5: Association between P2 Strength and Treatment Preference by Condition. Lower Treatment Preference scores indicate stronger preference for Desire Modification; higher scores indicate stronger preference for Behavior Modification. Shaded bands represent 95% CI.

Do P2 subtypes influence preference for desire change vs behavior change?

To explore whether the relationship between P2 conflicts and relative preference for desire change varied depending on the subtypes of P2 conflicts, we first examined whether different subtypes of P2s predicted Treatment Preference within each Condition, controlling for P2 Strength and Behavioral Frequency. We fit a linear regression predicting Treatment Preference from Identity P2, Moral P2, Instrumental P2, P2 Strength, and Frequency. All continuous predictors were mean-centered.

In the Food condition, stronger P2 Strength predicted lower Treatment Preference scores, indicating greater relative preference for Desire Modification over Behavior Modification (b = −0.22, SE = 0.08, t = −2.84, p = .005). None of the P2 subtypes significantly predicted Treatment Preference (Identity P2: p = .307; Moral P2: p = .552; Instrumental P2: p = .077), neither did Frequency (p = .164).

In the Sex condition, none of the predictors significantly predicted Treatment Preference (Identity P2: $p = .861$; Moral P2: $p = .944$; Instrumental P2: $p = .103$; P2 Strength: $p = .639$; Frequency: $p = .689$).

These results suggest P2 Strength rather than specific subtypes of P2s seem to be the clearest predictor of Treatment Preference. To more directly test the role of P2 subtypes in predicting Treatment Preference, we conducted another exploratory analysis on whether the P2 subtypes explained additional variance in Treatment Preference beyond overall P2 Strength and Behavioral Frequency across both conditions. We compared a baseline model containing only P2 Strength, Frequency, and Condition to a model that additionally included Identity P2, Moral P2, and Instrumental P2. Adding the three subtypes of P2s to a model containing P2 Strength, Frequency, and Condition did not significantly improve model fit, $F(3, 779) = 0.76$, $p = .518$. Thus, although the P2 subtypes differed descriptively by Condition, they did not explain additional variance in Treatment Preference beyond overall P2 Strength and Behavioral Frequency. Therefore, we found little evidence that any specific subtypes of P2s predicted Treatment Preference above and beyond P2 Strength.

Together, these findings suggest that the overall strength of P2 conflict was a stronger predictor of Treatment Preference than specific subtypes of P2 conflicts. Greater strength of P2 conflict predicted greater preference for desire change over behavior change, but this association did not seem to depend on different subtypes of P2 conflicts.

Do these associations vary by Food vs Sex domain?

As preregistered, to test whether the associations between different subtypes of P2 conflicts and Treatment Preference differed by Condition, we fit a regression model predicting Treatment Preference from Identity P2, Moral P2, Instrumental P2, P2 Strength, Frequency, Condition, and the interaction between each predictor and Condition. Food was the reference

category for Condition. All continuous predictors were mean-centered before interaction terms were computed. Collinearity diagnostics did not indicate problematic multicollinearity (all VIFs < 3). A nested model comparison indicated that the interaction model fit significantly better than the main-effects model, $F(5, 774) = 2.26$, $p = .047$.

As shown in Table 1.6, P2 Strength significantly and negatively predicted Treatment Preference, indicating that stronger P2 conflict was associated with greater relative preference for Desire Modification over Behavior Modification.

Table 1.6: Regression of Treatment Preference on Desire Characteristics and Condition. Reference category for Condition = Food. *b* = unstandardized regression coefficient. * $p < .05$. ** $p < .01$. *** $p < .001$.

Predictor	***b***	***SE***	***t***	***p***	
Intercept	-0.85	0.09	-9.33	< .001	***
Identity P2	0.06	0.06	1.01	.315	
Condition (Sex)	-0.09	0.22	-0.41	.685	
Moral P2	0.03	0.06	0.59	.558	
Instrumental P2	-0.12	0.07	-1.74	.082	
P2 Strength	**-0.22**	**0.08**	**-2.80**	**.005**	**
Behavioral Frequency	-0.09	0.07	-1.37	.171	
Identity P2 × Condition	-0.08	0.12	-0.63	.527	
Moral P2 × Condition	-0.04	0.10	-0.41	.680	
Instrumental P2 × Condition	**0.26**	**0.11**	**2.43**	**.015**	*
P2 Strength × Condition	0.16	0.14	1.09	.278	
Behavioral Frequency × Condition	0.13	0.12	1.13	.258	
R^2	.032				
Adjusted R^2	.019				
$F(11, 774)$	2.35			.007	

The only significant interaction was between Instrumental P2 and Condition ($b = 0.26$, $SE = 0.11$, $t(774) = 2.43$, $p = .015$), such that Instrumental P2 was more positively associated with preference for Behavior Modification in the Sex condition than in the Food condition (See Figure 1.6). However, as shown in the within-Condition models reported above, Instrumental P2 did not significantly predict Treatment Preference in either Condition. Thus, this interaction is best interpreted as a relative difference in the association between Instrumental P2 and Treatment Preference across Conditions.

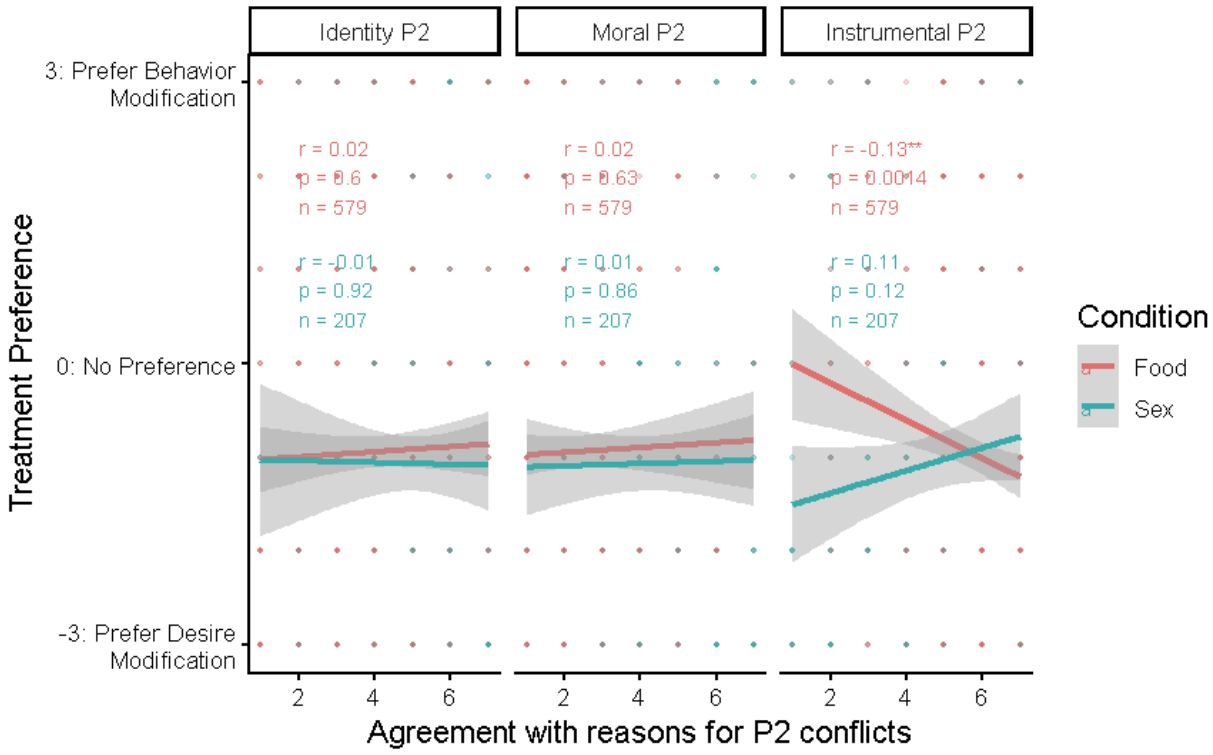


Figure 1.6: Association between P2 subtype endorsement and Treatment Preference by Condition. Lower Treatment Preference scores indicate stronger preference for Desire Modification; higher scores indicate stronger preference for Behavior Modification. Shaded bands represent 95% confidence intervals.

Together, these results suggest that the relationship between different subtypes of P2 conflicts and Treatment Preference varied by domain, but only for Instrumental P2: the more a P2 conflict is instrumental, the greater the preference for behavior change in the Sex domain compared to the Food domain.

Discussion

The current study brings P2 conflicts from the conceptual realm to the psychological realm by offering an initial empirical operationalization of P2 conflict as preference-directed rather than behavior-directed. Specifically, we examined whether stronger P2 conflict predicted a greater preference for desire change over behavior change, and explored whether different subtypes of P2 conflicts, and their associations with that preference, varied across food- and sex-related domains.

Consistent with our key distinction that P2 conflicts are about preferences rather than behaviors, participants who reported stronger P2 conflicts showed a greater preference for Desire Modification over Behavior Modification. This provides preliminary evidence that defining P2 conflicts as preference-directed is psychologically meaningful: the stronger participants' conflict about their desire, the more they preferred changing the desire itself rather than only changing the behavior.

However, the association between P2 conflict strength and preference for desire change over behavior change did not vary based on whether the P2 conflict was driven by instrumental, moral, or identity-based reasons. When predicting treatment preference, the overall strength of P2 conflict appeared to matter more than the specific reason underlying it: participants with stronger P2 conflicts showed a greater relative preference for Desire Modification, whereas instrumental, moral, and identity concerns did not explain additional variance in this preference beyond P2 conflict strength.

The subtypes of P2 conflicts did vary meaningfully across domains. Food-related P2 conflicts were more strongly associated with instrumental concerns, whereas sex-related P2 conflicts were more strongly associated with moral and identity concerns. This pattern aligns with the idea that unwanted food-related desires are often framed as matters of habit, health, and self-regulation, whereas unwanted sexual desires more often implicate moral values and self-concept. Notably, in the sex domain, instrumental concerns were also significantly correlated with moral and identity concerns, suggesting that for sex-related desires, the practical and the personal may be more tightly intertwined. Consistent with this pattern of associations, we found a significant interaction in which instrumental P2 conflicts were more strongly associated with a preference for behavior change over desire change in the sex domain than in the food domain. One possibility is that people are more willing to change food-related desires than sex-related desires because sexual desires are more central to one's self-identity. Thus, when P2 conflicts are driven by instrumental concerns, people may be more willing to change food-related desires to resolve the practical problem "once and for all," but less willing to change sex-related desires even when those desires create practical difficulties. Future work should investigate these possibilities.

This study has several limitations. First, because we selected for participants who had both behavioral conflicts and P2 conflicts, the study cannot by itself fully establish that P2 conflicts and self-control conflicts come apart empirically; but it does provide a starting point for doing so. The next step is to compare participants who have both types of conflicts with those who have self-control conflicts without P2 conflicts, or P2 conflicts without self-control conflicts (i.e., the bottom-left and top-right cells in Table 1.3). Future research that experimentally distinguishes these two types of conflict would provide cleaner and more robust evidence that the

two types of conflicts are structurally and empirically distinct. Second, our measure of treatment preference is hypothetical: whether people would actually seek out Desire Modification, and whether it is even possible, remains an empirical question. Third, the current study focused on food- and sex-related domains as a starting point; future work should examine P2 conflicts across the broader range of domains identified in the pilot survey, including some that were more commonly reported, such as social image, negative emotional states, and acquisitions. Finally, we operationalized each subtype of P2 as agreement with one statement describing that subtype, which may not capture the full richness or variability of these reasons that motivate P2s. Developing validated multi-item scales for these subtypes would be a productive next step.

Concluding thoughts

In this paper, we proposed a psychological framework for second-order preferences (P2s) — preferences about one's existing first-order preferences (P1s) — and the conflicts that arise between them. P2 conflicts have been richly theorized in philosophy, but rarely studied empirically in psychology. This paper takes a first step toward bridging that gap by clarifying the psychological structure of P2 conflicts and offering an initial way to measure them.

Part I developed the conceptual landscape of P2 conflicts by examining both how preferences emerge cognitively and what reasons motivate them. We argued that both P1s and P2s can arise through reflexive or reflective processes, and that P2s in particular can be driven by instrumental, value-based, and identity-based reasons. These distinctions shape the kind of P2 conflicts that can arise, and how these conflicts can be resolved. Instrumental P2 conflicts are typically extrinsic—they arise in service of other instrumental goals – and can in principle be resolved through changes in external circumstances. Value and identity P2 conflicts, by contrast, are intrinsic – they concern the preferences themselves, and can only be resolved through internal change. We then revisited prior work in philosophy in light of this framework, arguing

that the different positions in the field on how to act in the face of P2 conflicts may reflect different responses to distinct subtypes of P2 conflicts.

Part II turned to the empirical question of how P2 conflicts can be distinguished from self-control conflicts — another type of internal conflict they often co-occur with. We argued that the two are structurally distinct: P2 conflicts take preferences as their object, whereas self-control conflicts take behavioral motivations as their object. In a preregistered study, we found that stronger P2 conflict predicted a greater preference for changing desires rather than behavior, providing initial evidence that P2 conflicts can be meaningfully operationalized as preference-directed rather than merely behavior-directed. We further examined whether different subtypes of P2 conflicts were differentially associated with preference for desire versus behavior change, and whether these subtypes varied across domains. We found that food-related P2 conflicts were more strongly tied to instrumental reasons, whereas sex-related P2 conflicts were more strongly tied to moral and identity-based reasons. Moreover, instrumental P2 conflicts were more strongly associated with a preference for behavior change over desire change in the sex domain than in the food domain.

Theoretical implications

To our knowledge, this paper is among the first to study P2 conflicts as a psychological phenomenon in their own right, providing a broader and more nuanced perspective on internal conflicts. This framework suggests that the canonical narrative of internal conflict in psychology — a regulatory force battling an immediate desire — does not tell the whole story. In many cases, the conflict runs deeper: people may not only have a self-control conflict about what to do, but also an additional conflict about what to desire.

Studying the psychology of P2s opens the door to many interesting theoretical connections, one of which we briefly touched on in Part I: the true self. Identity P2 conflicts, as

we discussed, often involve a perceived gap between one's current preferences and a deeper sense of self. Research on the true self has found that people tend to perceive it as morally good, that this perception is cross-culturally stable, and that people are motivated to align their actual selves with it (Newman et al., 2014; Strohminger et al., 2017). This raises an interesting complication for the empirical picture: if the true self is idealized, people may be reluctant to change preferences they experience as fundamental to it — even when those preferences are unwanted in other respects. In other words, perhaps not every unwanted desire is one that a person wants to change, and the relationship between P2 conflict strength and preference for desire change may be moderated by how central the desire feels to one's sense of self.

A related topic is transformative experience – experiences that may change a person in a fundamental way that cannot be foreseen (Paul, 2014). Some P2 conflicts may be especially consequential because acting on them would not merely change a person's preference, but transform fundamentally who they are. Consider a P2 for desiring parenthood. In this case, a person's decision to act on their P2 – for instance, having a child in the hopes of developing a genuine desire for parenthood — constitutes a decision to embark on a transformative experience in which one cannot know in advance how they will be transformed. Future research can explore how people make decisions about whether to act on these "transformative P2s".

Clinical and applied implications

The distinction between P2 and self-control conflicts has important practical and clinical implications. Self-control issues are canonically addressed through interventions that help people resist or regulate behavior, such as precommitment devices or behavioral reinforcement. These approaches focus more on behavioral patterns rather than the preferences that lead to them per se. But if people with P2 conflicts are ultimately troubled by their desires, even in the absence of behavioral issues — then purely behavioral interventions may miss the point for them: The

faithful husband wouldn't be satisfied with precommitment devices that keeps him from acting unfaithfully because he ultimately wants to tame his unwanted desires.

It follows that interventions targeting people's relationship to their desires, such as some forms of acceptance and commitment therapy, may be especially relevant for people who have P2 conflicts. Therapy, coaching, and behavioral interventions could benefit from clarifying whether a client's goal is to change what they do or to change what they want — a question the client may not have considered explicitly.

Future directions

This paper provides a first step toward a broader program of research on P2 conflicts and related metacognitive phenomena. Many questions remain open. One important question is how to experimentally distinguish cases in which people have P2 conflicts without self-control conflicts, self-control conflicts without P2 conflicts, and both. This would provide a cleaner test of the structural distinction that we proposed here. Another intriguing question that emerged from our findings is why instrumental P2 conflicts appeared more behavior-oriented in the sex domain than in the food domain, and how these patterns observed in food- and sex-related desires generalize to other domains. The pilot survey reported here found P2 conflicts across a wide range of domains that would be interesting to explore.

Beyond this study, many open questions arise about the stability and development of P2s over time. Some P2 conflicts may be responses to context, whereas others may be more stable, like dispositional traits. Understanding these differences would paint a broader and more nuanced picture of the psychology of P2 conflicts.

More broadly, P2 conflicts are one instance of a wider class of metacognitive phenomena that are richly theorized in philosophy but remain empirically underexplored in psychology and cognition. Other examples include akrasia, epistemic emotions, and higher-order beliefs. These

phenomena all involve representations of, and stances toward, one's own mental states. A research program that systematically studies such representations can shed light on fundamental questions about agency and self. These questions are not confined to human cognition: as artificial intelligence systems become increasingly capable of modeling their own learning processes and uncertainty, questions about whether they can form representations of themselves, or higher-order preferences about their own learning processes, become increasingly meaningful. In this light, P2 conflicts are not just an isolated phenomenon that happens to be interesting. They are a window into the deeper question of who we are — and what it means to want to be otherwise.

Acknowledgements

Chapter 1, is currently being prepared for submission for publication of the material. Liu, Xingyu S.; McKenzie, Craig R.M.; Sher, Shlomi; Winkielman, Piotr; and McCullough, Michael E. The dissertation author was the primary researcher and author of this material.

REFERENCES

Bethel, A. C. W. (1980). Wanting to want. *Philosophy Research Archives*, *6*, 118–125. https://doi.org/10.5840/pra198066

Bruckner, D. W. (2011). Second-order preferences and instrumental rationality. *Acta Analytica*, *26*(4), 367–385. https://doi.org/10.1007/s12136-010-0113-x

Callard, A. (2018). *Aspiration: The Agency of Becoming*. Oxford University Press.

Carballo, A. P. (2018). Rationality & Second-order preferences. *Noûs*, *52*(1), 196–215. https://doi.org/https://doi.org/10.1111/nous.12155

Carter, J. (1976). *Interview with "Playboy" Magazine | The American Presidency Project* [Interview]. https://www.presidency.ucsb.edu/documents/interview-with-playboy-magazine

Cummings, R., & Roskies, A. L. (2020). Frankfurt and the problem of self-control. In A. R. Mele (Ed.), *Surrounding Self-Control* (p. 0). Oxford University Press. https://doi.org/10.1093/oso/9780197500941.003.0022

De Freitas, J., Sarkissian, H., Newman, G. E., Grossmann, I., De Brigard, F., Luco, A., & Knobe, J. (2018). Consistent belief in a good true self in misanthropes and three interdependent cultures. *Cognitive Science*, *42*(S1), 134–160. https://doi.org/10.1111/cogs.12505

Doyle, J. (2012). Survey of time preference, delay discounting models. *Judgment and Decision Making*, *8*. https://doi.org/10.2139/ssrn.1685861

Duckworth, A. L., & Steinberg, L. (2015). Unpacking self-control. *Child Development Perspectives*, *9*(1), 32–37. https://doi.org/10.1111/cdep.12107

Frankfurt, H. G. (1971). Freedom of the will and the concept of a person. *The Journal of Philosophy*, *68*(1), 5–20. https://doi.org/10.2307/2024717

George, D. (1998). Coping rationally with unpreferred preferences. *Eastern Economic Journal*, *24*(2), 181–194.

Green, L., & Myerson, J. (2004). A discounting framework for choice with delayed and probabilistic rewards. *Psychological Bulletin*, *130*(5), 769–792. https://doi.org/10.1037/0033-2909.130.5.769

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, *108*(4), 814–834. https://doi.org/10.1037/0033-295X.108.4.814

Higgins, E. T. (1989). Self-Discrepancy theory: What patterns of self-beliefs cause people to suffer? In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology* (Vol. 22, pp. 93–136). Academic Press. https://doi.org/10.1016/S0065-2601(08)60306-8

Hoch, S. J., & Loewenstein, G. F. (1991). Time-inconsistent preferences and consumer self-control. *Journal of Consumer Research*, *17*(4), 492–507. https://doi.org/10.1086/208573

Holton, R. (2009). *Willing, Wanting, Waiting*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199214570.001.0001

Jeffrey, R. C. (1974). Preference among preferences. *The Journal of Philosophy*, *71*(13), 377–391. https://doi.org/10.2307/2025160

Korsgaard, C. (1996). Reflective endorsement. In C. M. Korsgaard (Ed.), *The Sources of Normativity* (pp. 49–89). Cambridge University Press. https://doi.org/10.1017/CBO9780511554476.004

Newman, G. E., Bloom, P., & Knobe, J. (2014). Value judgments and the true self. *Personality and Social Psychology Bulletin*, *40*(2), 203–216. https://doi.org/10.1177/0146167213508791

Oxford English Dictionary. (2025). *Preference, n.* Oxford University Press. Oxford English Dictionary. https://doi.org/10.1093/OED/7827183124

Paul, L. A. (2014). *Transformative Experience*. Oxford University Press.

Priester, J. R., & Petty, R. E. (2001). Extending the bases of subjective attitudinal ambivalence: Interpersonal and intrapersonal antecedents of evaluative tension. *Journal of Personality and Social Psychology*, *80*(1), 19–34.

Rachlin, H. (2000). *The science of self-control* (p. 220). Harvard University Press.

Samuelson, P. A. (1938). A note on the pure theory of consumer's behaviour. *Economica*, *5*(17), 61–71. https://doi.org/10.2307/2548836

Schechtman, M. (2004). Self-expression and self-control. *Ratio*, *17*(4), 409–427. https://doi.org/10.1111/j.1467-9329.2004.00263.x

Sen, A. K. (1977). Rational Fools: A Critique of the behavioral foundations of economic theory. *Philosophy & Public Affairs*, *6*(4), 317–344.

Smith, M., Lewis, D., & Johnston, M. (1989). Dispositional theories of value. *Proceedings of the Aristotelian Society, Supplementary Volumes*, *63*, 89–174.

Stanovich, K. E. (2008). Higher-order preferences and the master rationality motive. *Thinking & Reasoning*, *14*(1), 111–127. https://doi.org/10.1080/13546780701384621

Strohminger, N., Knobe, J., & Newman, G. (2017). The true self: a psychological concept distinct from the self. *Perspectives on Psychological Science*, *12*(4), 551–560. https://doi.org/10.1177/1745691616689495

Wilson, T. D., & Dunn, E. W. (2004). Self-knowledge: Its limits, value, and potential for improvement. *Annual Review of Psychology*, *55*, 493–518. https://doi.org/10.1146/annurev.psych.55.090902.141954

Chapter 1 Supplemental Materials

Correlations between Behavioral Frequency and Subtypes of P2s

To examine whether Behavioral Frequency is related to Treatment preference and to the different subtypes of P2 conflicts, we conducted a preregistered secondary analysis testing the correlations between Frequency and Treatment Preference and Frequency and agreement with each subtype of P2s within each Condition. In the Food condition, Behavioral Frequency was negatively and significantly correlated with Treatment Preference, $r = -.09$, $p = .023$, indicating that participants who acted on food-related desire more showed stronger preference for Desire Modification over Behavior Modification. Behavioral Frequency was also positively and significantly correlated with Instrumental P2 ($r = .18$, $p < .001$), suggesting that more frequent food-related behaviors were associated with stronger practical concerns. Behavioral Frequency was not significantly correlated with Moral P2 ($p = .056$) or Identity P2 ($p = .709$).

In the Sex condition, Behavioral Frequency was not significantly correlated with Treatment Preference, Moral P2, Identity P2, or Instrumental P2 ($ps > .065$).

Overall, these correlations suggest that Behavioral Frequency was more closely tied to Instrumental P2 and Treatment Preference in the Food condition than in the Sex condition.

Participants with Frequency ratings of 4

Because Behavioral Frequency and its correlation with P2 subtypes differed by Condition, we conducted a preregistered secondary analysis restricting the sample to participants who gave the Behavioral Frequency rating of 4 (31.3%, $N = 246$). This analysis tested whether the main pattern of results replicate among participants who reported the same moderate level of Behavioral Frequency.

In this restricted sample, we fit a regression model predicting Treatment Preference from Identity P2, Moral P2, Instrumental P2, P2 Strength, Condition, and the interaction between each predictor and Condition. All continuous predictors were mean-centered before interaction terms were computed. A nested model comparison indicated that including the interaction terms did not significantly improve model fit, F (4, 236) = 0.73, p = .570.

P2 Strength negatively and significantly predicted Treatment Preference, b = −0.30, SE = 0.13, $t(236)$ = -2.26 p = .025, indicating that stronger P2 conflict was associated with greater relative preference for Desire Modification over Behavior Modification. None of the P2 subtypes significantly predicted Treatment Preference (Identity P2: p = .283; Moral P2: p = .420; Instrumental P2: p = .982).

These results suggest that the association between P2 Strength and Treatment Preference was robust when comparing participants with the full sample and when restricting the sample for moderate behavioral frequency. However, the Instrumental P2 × Condition interaction observed in the full sample did not emerge in this restricted analysis.

Model including all two-way interactions

As a preregistered secondary analysis, we fit a model including all two-way interactions among Identity P2, Moral P2, Instrumental P2, P2 Strength, Frequency, and Condition. All continuous predictors were mean-centered before interaction terms were computed. The overall model was significant, F(21, 764) = 1.58, p = .048, R^2 = .042, adjusted R^2 = .015. However, P2 Strength was the only significant predictor of Treatment Preference, b = −0.26, SE = 0.08, t (764) = −3.13, p = .002. No two-way interactions were significant (ps > .180).

These results were consistent with the primary analyses: stronger P2 conflict was associated with greater relative preference for Desire Modification over Behavior Modification,

whereas the specific P2 subtypes and their interactions did not consistently predict Treatment Preference.

Pilot Survey Details

Participants and procedure

Participants were recruited from www.prolific.com (N = 60, 30 male, 27 female, 3 non-binary, M_{age} = 37.48, SD_{age}= 13.95).

Participants were asked to provide at least one example of something they do, desire, or care about that they wished they did not, or at least not as much. The three prompts were presented on the same page, and participants were instructed to read all three prompts before responding.

The exact wording of the prompts was as follows:

"Are there things you **do** that you wish you didn't – or at least not so much? If so, please provide at least one example below:"

"Are there things you **desire** that you wish you didn't – or at least not so much? If so, please provide at least one example below:"

"Are there things you **care about** that you wish you didn't – or at least not so much? If so, please provide at least one example below:"

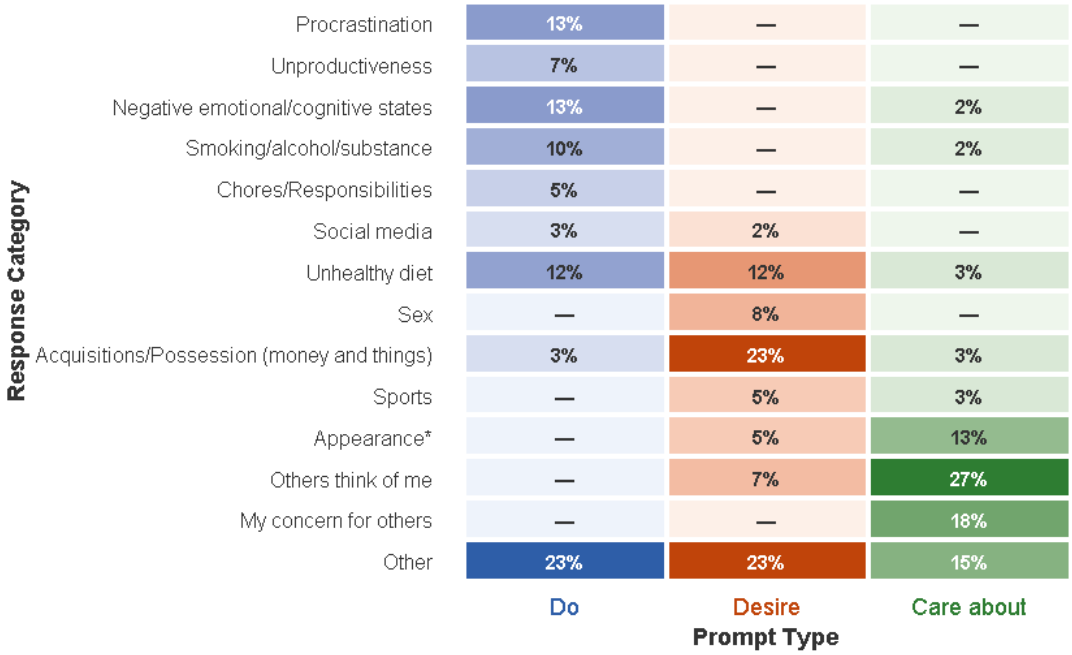
Coding Scheme

Responses were manually coded into categories by the first author. Categories were defined as themes that appeared in the responses from at least three participants; responses that did not appear from more than two participants were grouped into an "Other" category. 20 responses indicated no relevant conflict, and 3 responses seemed to have misinterpret the question or were otherwise unclear; together these responses were grouped into a "None/Unclear" category. Supplemental Table 1.1 below describes each category and any relevant coding notes.

Supplemental Table 1.1: *Survey response categories and coding notes.*

Category	Notes
Procrastination	
Unproductiveness	Includes: idling, not doing work on time, getting distracted
Social media	
Acquisitions/Possession (money and things)	
Chores/Responsibilities	
Appearance*	Includes looks and weight; *concerns about weight may or may not be appearance-related but were categorized here for simplicity
Sex	
Others think of me	Includes: social approval, being liked, social media likes, fame
My concern for others	
Unhealthy diet	
Smoking/alcohol/substance	
Negative emotional/cognitive states	Includes: anxiety, anger, over-worrying, rumination, negativity, defensiveness
Sports	
Other	Includes any response that did not recur for more than two participants
None/Unclear	Includes responses that appear to misinterpret the question or are unclear, or indicated no relevant conflicts

Supplemental Figure 1.1 shows the most common response categories in each prompt. The prompts elicited partially distinct response patterns. “Do” responses more often involved behavioral habits or self-regulation concerns, whereas “desire” and “care about” responses more often involved categories tied to attraction, appearance, others’ opinions, or self- and identity-relevant concerns. Although “Other” was the most common category across prompts, this category reflected the diversity of idiosyncratic responses that did not occur more than twice.



Supplemental Figure 1.1: Common response categories elicited by the "do," "desire," and "care about" prompts. Each cell shows the percentage of responses of each category within each prompt. "Other" includes idiosyncratic responses that did not occur more than twice. *Appearance includes concerns about looks and weight. "None" and unclear responses not presented in the table. Percentages are out of total N = 60.

Chapter 2 WHEN DEFAULTS EXPLAIN AWAY PREFERENCES: A CAUSAL REASONING ACCOUNT OF MENTAL STATE REASONING FROM DEFAULT OPTIONS

Introduction

Imagine that you and a friend are at a restaurant choosing between fries and salad as a side dish, and by default, all orders come with salad. Your friend orders the salad. How much does your friend like salad compared to fries? What if by default, all orders come with fries instead? How much does your friend like salad compared to fries now? In this scenario, while your friend's choice is always salad, your inference about their food preferences may vary based on the default option.

In this way, people often infer that decision-makers (DMs) who actively switch from a default option have a stronger preference for their choice than those who passively accept the default (Davidai et al., 2012; Leong et al., 2020; Lin et al., 2018). For example, people rated customers who chose salad over fries as caring more about healthy eating when the default dish was fries rather than salad (Leong et al., 2020). Similarly, people perceived organ donors as preferring organ donation more – or finding it more meaningful – in countries where the default is to not be an organ donor than in countries where the default is to be an organ donor (Davidai et al., 2012; Lin et al., 2018). In other words, people make *asymmetric inferences* about a DM's preference given the same choice outcome when the DM switches from versus accepts a default (Leong et al., 2020).

Inferring other people's preferences is an essential part of social life: Successful social relationships often require people to take into account others' preferences, which often are not directly observable but must be inferred from behavior, such as others' choices. Typically, one infers that a DM's choice reflects their preference. However, in the presence of defaults this link is more complex, leading to asymmetric inferences. Here, we propose and test a new account of

why asymmetric inferences from defaults occur, suggesting that asymmetric inferences reflect structured and rational processes of causal inference (Tenenbaum et al., 2006), and in particular, inverse decision-making (Jern et al., 2017).

Causal reasoning about preferences from default options

In our proposed causal reasoning account, the asymmetric preference inferences can be characterized as a form of causal inference and inverse decision-making (Jern et al., 2017; Tenenbaum et al., 2006). By this account, people infer a DM's preference by reasoning backward from an observed choice to the possible causes that could have produced it (Jern et al., 2017). In many cases, preference is a highly plausible cause of choice. However, in the presence of a default, the default itself may provide reasons for choice (other than preference), which have been used to explain why people tend to choose an option more often when it is the default than when it is not (Jachimowicz et al., 2019). Three main reasons for choosing the default have been suggested (Dinner et al., 2011). One explanation is *ease* – accepting the default typically requires less physical and mental effort than deciding on a different option (Sunstein & Thaler, 2003). Another explanation is *endorsement* – defaults can reflect the default-setter's implicit recommendation, providing relevant information that may push the DM to accept the default (McKenzie et al., 2006). The final explanation is *endowment* – DMs may see the default as their "endowment", or what they already own, which makes switching from the default seem like a loss to be avoided (Dinner et al., 2011).

In line with a causal reasoning account, we propose that the extent to which people make asymmetric preference inferences depends on the extent to which the default provides a plausible alternative reason for accepting the default – one that can "explain away" any link to a DM's preference. This pattern of reasoning, in which one explanation weakens the evidence of others,

is a signature of structured, rational causal inference (Gopnik & Sobel, 2000; Pearl, 2000; Tenenbaum et al., 2006).

To formalize this reasoning process, imagine a task in which an observer watches a DM choose between two options, one of which is set as a default option. Let L (for "liking") be a DM's relative preference between the two options, where zero is equal preference. An observer does not know the true value of L, but instead begins with a prior distribution of possible values of L, $p(L)$, which captures their general belief about the DM's probable preferences before observing their choice. This prior belief may be based on the observer's general knowledge or stereotypes of people similar to the DM. For instance, consider a child choosing between chocolate and broccoli. Here, L will be the child's relative preference between chocolate and broccoli, such that larger L stands for stronger preference for chocolate over broccoli. In this situation, an observer may begin with a prior, $p(L)$, that is centered above zero, reflecting their belief that children in general prefer chocolate over broccoli.

In addition, before observing the DM's choice, the observer may evaluate the specific choice architecture with regard to the default option. In particular, the observer may estimate the amount of pressure the default option is putting on the DM's choice behavior, or its strength (e.g. large vs. small switching cost; strong vs. weak implicit recommendation). This contextual information allows observers to estimate the extent to which the default provides alternative reasons for choice, which crucially matters for their preference inferences. This factor is thus included in our model as parameter δ, expected strength of the default's influence on choices.

After observing the DM's choice C (for "choice", one of the available options), given a default option D (for "default"), the observer updates their belief about the DM's preference,

taking into account the expected default strength, δ. That is, they update their prior to a posterior distribution, $p(L|C, D, \delta)$, based on Bayes' rule:

$$p(L|C, D, \delta) = \frac{P(C|L,D,\delta)\, p(L)}{P(C|D,\delta)}$$

In this equation, the likelihood term $P(C|L, D, \delta)$ captures how likely the observed choice would be given a specific preference and a particular default, and a particular expected default strength, δ. This likelihood trades off against the marginal likelihood, $P(C|D, \delta)$, which captures how likely the observed choice would be overall, averaging over all possible preference values, and all possible default strengths. Here, choosing the default is highly diagnostic if the default strength δ is low, because then the choice would be likely only if the DM truly prefers it, $P(C|L, D, \delta) \gg P(C|D, \delta)$. By contrast, choosing the default is not as diagnostic when the default strength δ is high, because then, the choice would be likely regardless of the DM's preference, $P(C|L, D, \delta) \approx P(C|D, \delta)$. This is because the choice would be driven by the strong default even if they do not prefer it.

For instance, imagine a scenario in which a child has chosen broccoli, either by accepting or switching from a default snack set by their parents. When the default is broccoli, this default provides a strong plausible reason for accepting the default – following the parents' recommendation to eat healthy, a form of implicit recommendation (McKenzie et al., 2006). The pressure to follow a parents' recommendation to eat healthy may be strong, and impact the child's choice regardless of their true preference. In this context, the expected default strength $\delta_{broccoli}$ is high, and thus the choice of broccoli is likely across a wide range of possible preference values, $P(C = broccoli\ |L, D = broccoli, \delta_{broccoli}) \approx P(C = broccoli\ |D = broccoli, \delta_{broccoli})$. As a result, choosing broccoli when it is the default option provides weak evidence that the child actually prefers broccoli, and the posterior distribution should shift only

slightly, if at all, thus retaining a belief that the child prefers chocolate even though broccoli was chosen. In contrast, when the child switches away from a default of chocolate to choose broccoli, this choice is highly diagnostic of preference, as the choice of broccoli when chocolate is the default is much more likely when the child actually prefers broccoli than it is overall, $P(C = broccoli\,|L, D = chocolate, \delta_{chocolate}) \gg P(C = broccoli\,|D = chocolate, \delta_{chocolate})$. Therefore, the posterior distribution should shift towards broccoli preference dramatically more when the choice has been made by switching, rather than accepting a default - in line with classic asymmetric preference effects.

However, the situation differs when the chosen item is chocolate. Given that people typically expect children to prefer chocolate over broccoli, setting chocolate as the default option likely signals a form of licensed indulgence – licensing behavior in line with a preference that the observer assumes the child has *a priori*. As a result, the default does not provide an alternative reason for accepting the default, and does not explain away preference for the chosen item. In other words, when the prior $p(L)$ of chocolate is high, and especially when the expected default strength $\delta_{chocolate}$ is also low, then the likelihood of choosing chocolate when the child prefers it is always higher than the marginal likelihood, whether the default was chocolate or broccoli. Regardless of whether the choice of chocolate was arrived at by accepting or switching away from the default, the choice is consistent with prior beliefs about the child's preference for chocolate, and the posterior should retain a strong chocolate preference in both cases.

Predictions of the causal reasoning account

The causal reasoning account and inverse decision-making model above makes predictions about when and why asymmetric preference inferences occur and diminish in the presence of defaults. Specifically, the causal reasoning account predicts that asymmetric preference inferences will occur when there are strong and plausible alternative reasons for

accepting the default, in line with existing literature. However, the causal reasoning account makes the distinct and novel prediction that asymmetric inferences should diminish, or even disappear, when accepting the default remains better explained by preferences than by any alternative reasons.

In terms of our example above, because $\delta_{broccoli} > \delta_{chocolate}$, the choice of broccoli is more diagnostic of preference when chocolate is the default than when broccoli is the default. Therefore, the expected value of the posterior should be larger (weaker inferred preference for broccoli) when child accepts the broccoli default than when they switch from the chocolate default, leading to asymmetric inference based on the same chosen food. By contrast, because the prior $p(L)$ of liking chocolate is high, when a child chooses chocolate, the choice is similarly diagnostic of preference regardless of the default. Therefore, the expected value of the posterior should be similar regardless of the default.

This account predicts that this type of diminished asymmetry should occur in multiple contexts, which have not been examined in previous studies. One such context is exemplified in the chocolate default scenario described above, where the default aligns with a DM's likely existing preferences. In this case, although the default can still signal endorsement, it is endorsing the DM's existing preference rather than providing an alternative reason for choice other than preference. Similarly, when a default is seen as reflecting a status quo, such as continuing a prescription (Jachimowicz et al., 2019); or when the default-setter's recommendations is irrelevant to the DM, such as accepting the default seat selection on a flight –while the default can signal endorsement (McKenzie et al., 2006), this endorsement would not exert much pressure on choice (low δ), and would therefore again not be sufficient to explain away preference.

In addition, some defaults exert pressure on choice due to ease, or cost of switching (Sunstein & Thaler, 2003). The causal reasoning model predicts that in this case, inferred preferences should be more or less asymmetric depending on the cost of switching from the default, which would impact the value of parameter δ. When the cost of switching from the default is low (e.g., one can switch from the default by clicking a button when online shopping), δ is low, and accepting the default should remain better explained by preference.

In both the ease and endorsement examples, δ is low and there are no strong alternative reasons for accepting the default. Therefore, the DM's choice is much more likely when they actually prefer the chosen item than it is overall, leading to similar posterior distributions regardless of the default, and reducing the asymmetry of preferences – the choice is better explained by preference either way.

While asymmetric preference inferences are well-explored in the literature, existing studies mostly focused on situations in which defaults provided strong alternative explanations for accepting it, including both ease (high cost) and strong endorsement. For example, in Leong et al. (2020), participants read vignettes in which the DMs chose salad by either accepting or switching from a default; the DM who switched from the default was on average rated as caring about healthy eating more. Here, a salad default seems to imply an endorsement of healthy eating norms, providing a plausible alternative reason for choice (δ_{salad} is high). Crucially, this work did not ask participants to consider the other possible situation, when fries were chosen. Like chocolate in the example above, a default of fries may signal licensed indulgence, allowing people to choose in line with their own prior preferences (δ_{fries} is low, assuming most people think that restaurant-goers tend to prefer fries to salad). It follows that, if the DM chooses fries, the causal reasoning account predicts that the choice is better explained by preference regardless

of the default, leading to reduced asymmetry – but we cannot test this prediction without situations in which fries are chosen.

While prior work on default organ donation policies examined both situations in which "being a donor" was chosen and "not being a donor" was chosen, unlike in the salad-fries example above, both default policies can provide alternative explanations for choice (Davidai et al., 2012; Lin et al., 2018). People may accept either default policy because of ease: switching requires the effortful navigation of paperwork, phone calls, or trips to the relevant offices. They may also accept either default policy in order to follow social norms: The default policy of donation may signal an implied norm of altruism (see Davidai et al., 2012, Exp. 3), while the default policy of not donating organs may signal norms of conservatism or religious practice. Therefore, the organ donation context does not allow for situations in which the default policy does *not* provide strong alternative reasons for choice, and therefore cannot assess when asymmetric inferences diminish.

In sum, the existing literature only focused on situations in which accepting the default can be explained by a plausible alternative reason other than preference, supporting the causal reasoning account's predictions about when asymmetric inferences arise – but not when they diminish. We address this gap by testing the causal reasoning account's key predictions on when asymmetric inferences occur, but more importantly, when they diminish.

The Current Studies

In five preregistered studies, we asked participants to make inferences about DMs' preferences from their choices in the presence of default options, where the defaults either provided, or did not provide, strong alternative reasons for choice other than preference. We focused on scenarios in which it was possible to manipulate the extent to which default options provided plausible alternative explanations for accepting the default, based on ease and

endorsement. To ensure that these alternative explanations can indeed explain away the DM's preferences as reason for choice, default options were first explicitly explained, operationalized as reflecting ease (the most physically accessible option, Study 1), and as reflecting endorsement (a relevant default-setter's recommendation, Studies 2a and 2b). Then, we tested whether the same causal reasoning process generalized to a traditional default format, in which the meaning of the default option was not explicitly stated, but simply presented as a pre-selected option (Studies 3 and 4).

If the causal reasoning account is correct, then asymmetric preference inferences should occur when accepting the default can be better explained by ease or endorsement than by the DM's preferences. By contrast, the asymmetry should diminish in strength or even disappear when the default does not provide a plausible alternative explanation for choice, such as when switching from the default is easy, or when the default conveys licensed indulgence.

Transparency and Openness

Reproducible codes and data for all studies can be found on OSF (https://osf.io/b2ysh/overview?view_only=03691719f693423fb8f9cd28159c5926).

The pre-registration for Study 1 can be viewed online at https://osf.io/6ersc/overview?view_only=4214ad19f330486a97d1d23b7bdba238

The pre-registration for Study 2a can be viewed online at https://osf.io/prb84/overview?view_only=2964153d39e041d99cc68d439ad75dbd

The pre-registration for Study 2b can be viewed online at https://osf.io/nku6r/overview?view_only=b44ba808eee44496995860ec9c5c050a

The pre-registration for Study 3 can be viewed online at https://osf.io/fr2ge/overview?view_only=b8e56a9b6d1748449e064121fe20325b

The pre-registration for Study 4 can be viewed online at

https://osf.io/84gbq/overview?view_only=4940545144e04791922115e3c890dd55

Data were analyzed using R version 4.3.0 (R Core Team, 2023), and plots were made using the package ggplot2, version 3.5.1. All materials and procedures were approved by the institutional review board of the University of California, San Diego. We report all measures, manipulations, and exclusions in these experiments (additional measures not included in analyses are described in the Supplemental Materials).

Study 1

In Study 1, we test scenarios in which *ease* serves as a plausible alternative explanation for choice other than preference (Sunstein & Thaler, 2003). In these scenarios, the default option is more easily accessible than the alternative, which provides an alternative explanation for accepting the default–avoiding the cost of switching. Crucially, this allows us to manipulate the cost of switching. As the cost increases, the expected strength of the default δ increases, and this alternative explanation becomes more plausible. The causal reasoning account therefore predicts that participants will make asymmetric preference inferences only when it is costly to switch from the default, and that this asymmetry will parametrically vary with the switching cost. When the cost of switching is negligible, the asymmetry in preference inference should diminish.

Methods

Participants

As preregistered, $N = 120$ adults from the U.S. participated, recruited from www.prolific.com ($M_{Age} = 41.29$, $SD_{Age} = 12.35$, Range = 18-75; 62 men, 56 women, 1 non-binary gender, 1 prefer not to say).

Design

The experiment used a 4 (Cost: No, Low, Medium, High) by 2 (Choice: Accept vs. Switch) within-subject design, where each participant participated in four trials, and each trial

was a unique condition. The trials were grouped into four pairs by Cost condition, with each pair containing one of each Choice trial (an Accept trial and a Switch trial). The order of the Cost conditions was counterbalanced across participants, and the order of Choice conditions was counterbalanced within each Cost condition.

Procedure and Stimuli

In each trial, participants read about a character choosing between two flavors of a food item at the grocery store. Specifically, the character was at the check-out line and was last in line, and remembered that they needed a food item (e.g. "potato chips"). One flavor was more easily accessible than the other (located "on a rack by the cashier that the character can grab without getting out of the line"), and can therefore be considered the default option. The other flavor required more physical effort to obtain; the additional level of effort varied by Cost condition.

In the Low-Cost condition, the alternative flavor was located at the back of the same rack, and the character could "*reach to the back of the rack*" to get it. In the Medium Cost condition, it was located "in the middle of the store" and the character needed to leave the checkout line to get it. In the High-Cost condition, it was located at "*another store that is a few minutes-drive away*" and the character needed to leave the store and drive to get it. We also included a No-Cost condition in which the two flavors were equally accessible (side by side on the same rack). This condition served as a baseline. Across all trials, participants saw four types of food items: potato chips (sour cream vs. BBQ), chewing gum (mint vs. bubblegum), chocolate bars (dark vs. milk), and juice (orange vs. apple). We used a Latin square design to counterbalance the correspondence between food items and Cost conditions, as well as the order in which Cost conditions were presented. This resulted in four between-subject presentation orders which were randomly assigned to participants. Each participant saw the food items in the

same order (potato chips, chewing gum, chocolate bars, juice), and the corresponding Cost condition varied according to the Latin square (for details, see Supplemental Materials).

Within each Cost condition, participants read about two characters who were in very similar situations: one who chose the default option (Accept trial), and one who chose the alternative option (Switch trial). The order of these two trials was counterbalanced across participants. After each trial, participants inferred the extent to which the character preferred one flavored food item over the other on a scale of -100 to 100, where -100 was the character "likes [the default flavor] a lot more", 0 was the character "likes both flavors equally", and 100 was the character "likes [the alternative flavor] a lot more". The slider handle was initially positioned at 0, and participants moved the slider to indicate their inferred preference.

After completing both trials in one Cost condition, participants proceeded to the remaining Cost conditions, which were identical in structure, but described new characters, food items, and switching costs.

Results

To test whether the asymmetry in preference inferences increased parametrically with switching cost, we fit a linear mixed-effects regression predicting the absolute value of the Inferred Preference from Cost, Choice, and their interaction, with random intercepts for participants. Cost was coded numerically (No-Cost = 0, Low-Cost = 1, Medium-Cost = 2, High-Cost = 3). As predicted, there was a significant interaction between Cost and Choice (b = 19.94, SE = 1.36, t (837) = 14.71, $p < .001$, *95% CI* = [17.29, 22.60]), indicating that the asymmetry in the strength of inferred preference between Accept and Switch trials increased parametrically with switching cost. Nested model comparison confirmed that including the interaction significantly improved mode fit, χ^2 (1) = 193.08, $p < .001$.

To examine the asymmetry within each cost condition without assuming linearity, we conducted follow-up analyses treating Cost as a categorical variable. Comparisons based on estimated marginal means revealed that the difference between Accept and Switch trials increased across Cost conditions (see Figure 2.1). As expected, there was no significant difference in the No Cost condition (b = -2.57, p = .39). In contrast, participants inferred significantly stronger preferences in the Switch trials than in the Accept trials in all other Cost conditions (Low: b = -26.56; Medium: b = -51.35; High: b = -60.77; all ps< .001; see Table 2.1), indicating a parametric increase in asymmetry as switching cost increased.

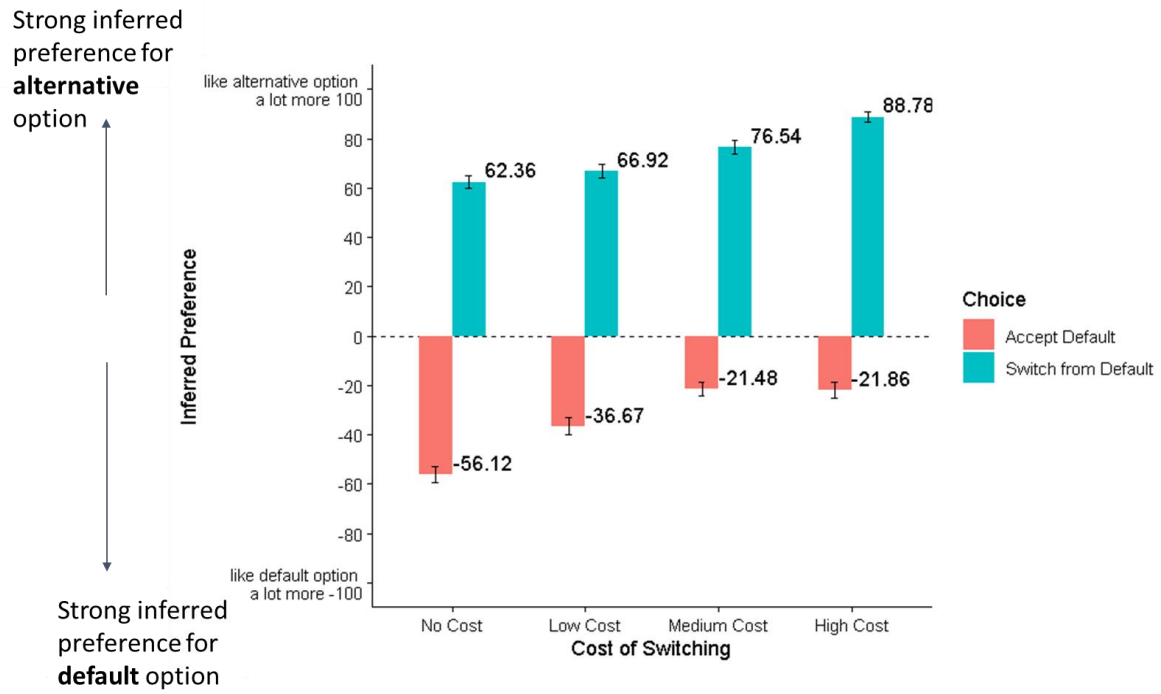


Figure 2.1: Average inferred preference across Cost conditions in Study 1. The asymmetry in inferred preferences increased parametrically with switching cost. Error bars represent ±1 SE.

Table 2.1: Difference in inferred preference strength (Stay - Switch) across Cost conditions in Study 1. Negative values indicate stronger inferred preference in the Switch trials than in the Accept trials.

Cost Condition	**Difference (b)**	***SE***	***95% CI***	***t***	***p***
No Cost	−2.57	3.00	[−8.45, 3.32]	−0.86	.39
Low Cost	−26.56	3.00	[−32.45, −20.67]	−8.86	< .001
Medium Cost	−51.35	3.00	[−57.24, −45.46]	−17.12	< .001
High Cost	−60.77	3.00	[−66.66, −54.89]	−20.26	< .001

Discussion

In Study 1, participants made increasingly asymmetric preference inferences between accepting versus switching away from the default as the cost of switching increased. As predicted, participants made weaker preference inferences when the characters accepted the default than when they switched away from it, and this difference increased parametrically with switching cost. These findings support our causal reasoning account, suggesting that ease — in particular, avoidance of switching cost — served as a plausible alternative explanation for accepting the default, therefore leading to increasingly asymmetric preference inferences as this explanation became more plausible. This shows that participants' preference inferences reflect a causal reasoning process that takes into account physical constraints in the choice environment. We next investigated whether participants can account for *social* constraints as alternative explanations within this same reasoning process.

Study 2a

Study 2a focused on scenarios in which *endorsement* can serve as a plausible alternative explanation for choice, capturing social features of the choice environment rather than the physical ones examined in Study 1. In these scenarios, the default option can be interpreted as

the implicit recommendation from the default-setter, which provides an alternative explanation for accepting the default independent of preferences (McKenzie et al., 2006). This alternative explanation is particularly plausible when the default-setter's opinion is important to the DM; we therefore operationalized this situation as a parent setting a default snack for their child. In addition, choices should be less diagnostic when the default is consistent with the participants' strong prior expectations of the DM's preferences. We therefore tested a situation in which we expected adults would hold strong a priori expectations about preferences (a child's preference for chocolate over broccoli). The causal reasoning account predicts that when the default is broccoli (the *a priori* dispreferred item), participants will infer stronger preferences when the DM chooses it by switching from the default than when they accept it. Crucially, it further predicts that this asymmetry will decrease when the chosen option is consistent with the participants' prior expectations of the DM's preference (e.g., a child choosing chocolate over broccoli, in the context of either default).

Methods

Participants

As preregistered, N=120 adults from the U.S. participated, recruited from www.prolific.com (M_{Age} = 37.29, SD_{Age} = 13.23, Range 18-84; 44 men, 74 women, 2 non-binary gender). This sample size was chosen based on bootstrapping analyses from a pilot sample, which suggested that 120 participants would provide sufficient power to detect the predicted differences in effect size when the characters chose broccoli versus chocolate. 16 additional participants were also tested, but excluded and replaced due to the preregistered exclusion criterion of not passing one or both attention checks.

Design

The experiment used a 2 (Default: Broccoli vs. Chocolate) by 2 (Choice: Accept vs. Switch) within-subject design, where each participant participated in four trials, and each trial was a unique condition. The four trials were grouped into two blocks by Default condition, with each block containing an Accept trial and a Switch trial. One block of trials involved girl characters, and the other one boy characters. The order of the Default conditions was counterbalanced across participants, as was the gender of the characters within each Default condition.

Procedure and Stimuli

In each trial, participants read and viewed illustrated vignettes depicting child characters choosing between two snacks at school. To allow for future comparison of adults' reasoning to children's, all stimuli and measures were designed to be comprehensible to children, as well as adults. Illustrations were created using Vyond (www.vyond.com) and Microsoft PowerPoint. In these vignettes, two child characters stood near a table with two bowls, one of which was marked with a star. In the Broccoli Default condition, the star was on the broccoli bowl (see Figure 2); in the Chocolate Default condition, the star was on the chocolate bowl. Participants read that the star indicated the default option had been selected and endorsed by the children's parents ("Before snack time, the parents picked out which snack they think is a good choice for the kids. All the parents of the kids picked [broccoli or chocolate] for them – that's why there's a star on that bowl"). Participants also learned that the child characters "get to choose which snack they want", but could see which one had a star on it.

Within each Default condition, participants saw two trials: in one trial, a child chose the default option (Accept trial), and in the other, a different child chose the alternative option (Switch trial). The Accept trial was always presented first.

After each trial, participants indicated which snack the child liked more (broccoli or chocolate, two-alternative choice), followed by whether the child liked that snack "a lot more" or "a little more". These two binary responses were combined into a 4-point scale indicating participants' inferred preference strength, ranging from 1 ("likes broccoli a lot more") to 4 ("likes chocolate a lot more"). This simplified measure was used to allow future comparison of data to child participants.

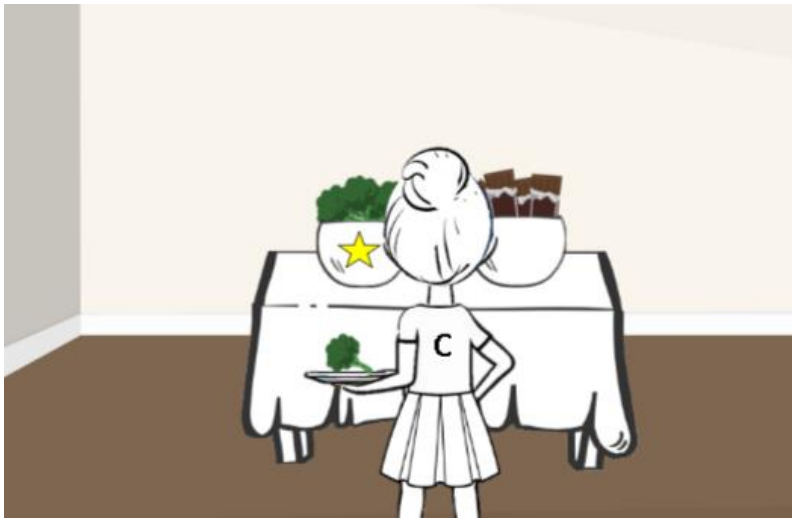


Figure 2.2: The snack with a star was described as the default option in Study 2a. On each trial, one character chose the default option (pictured); the other chose the non-default option. Images are copyrighted by and used by permission of VYOND, trademark of GoAnimate, Inc.

To ensure that participants paid attention to the vignettes, they completed one memory check within each Default condition: Before the second trial (i.e., the Switch trial), participants were asked to explain why there was a star on the bowl in free-response format. Two researchers independently coded these responses for accuracy (e.g. "the parents picked it"; "it was what the parents suggested"). As preregistered, participants who failed one or both memory checks were excluded and replaced (see *Participants*).

After completing both trials in one Default condition, participants proceeded to the other Default condition, which was identical except for depicting new child characters, and the opposite snack marked as the default. After the final trial, participants explained why they gave their response on the final trial in free-response format. As preregistered, two researchers independently screened these responses to identify and exclude bots and non-fluent English speakers; no participants were excluded on this basis.

Results

To test the prediction that the asymmetry in preference inferences would be stronger when child characters chose the broccoli (vs. chocolate), we fit an ordinal logistic regression predicting the absolute value of Inferred Preference from Default, Choice, and their interaction, with random intercepts for participants. As predicted, there was a significant interaction between Default and Choice (b = -10.98, SE = 0.69, z = -15.84, $p < .001$, *95% CI* = [-12.34, -9.62]), indicating that the extent of asymmetry in inferred preference between default trials (Accept vs. Switch) varied based on which kind of snack was chosen. A nested model comparison confirmed that the interaction term significantly improved model fit, χ^2 (1) = 490.74, $p < .001$.

To examine the magnitude of asymmetry for each chosen snack, we conducted Wilcoxon's signed rank tests on inferred preferences across the two trials when each snack was chosen. As predicted, participants inferred weaker preferences in Accept trials than in the Switch trials when broccoli was chosen, and this asymmetry disappeared when chocolate was chosen (see Figure 2.3). When broccoli was chosen, participants inferred significantly weaker preferences in the Accept trials (M = 2.05, SD = 0.89) than in the Switch trials (M = 1.43, SD = 0.64), V = 2375.5, $p < .001$, r = .52, *95% CI* = [0.39, 0.64], indicating strong asymmetry.

As predicted, this asymmetry disappeared when chocolate was chosen. Participants inferred similar levels of preference in the Accept trials (M = 3.87, SD = 0.34) and the Switch

trials ($M = 3.84$, $SD = 0.41$), $V = 72$, $p = .528$, $r = .04$, *95% CI* = [0.00, 0.23], indicating little to no asymmetry.

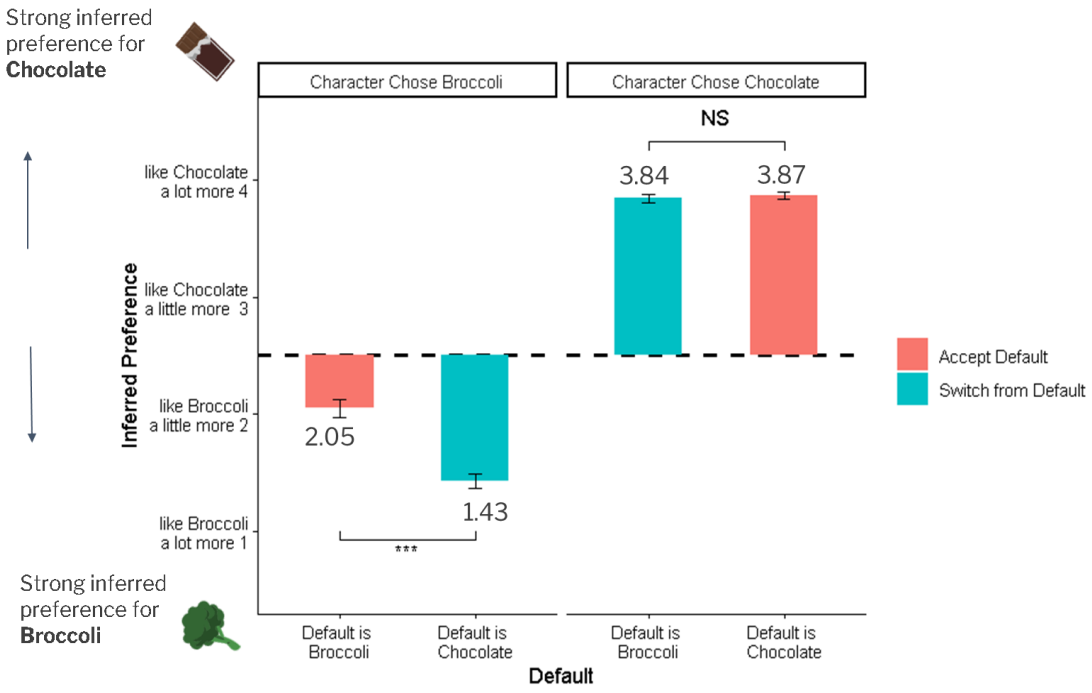


Figure 2.3: Participants' average inferred preference across Default conditions in Study 2a. Participants made asymmetric inferences when broccoli was chosen, and this asymmetry disappeared when chocolate was chosen. Error bars represent ±1 *SE*.

An exploratory analysis confirmed that the asymmetry diminished when chocolate was chosen. We fit an ordinal logistic regression predicting the absolute value of Inferred Preference from Choice, Chosen Snack (broccoli vs. chocolate), and their interaction, with random intercepts for participants. We found a significant interaction between Choice and Chosen Snack, $b = 1.50$, $SE = 0.46$, $z = 3.24$, $p = .001$, *95% CI* [0.59, 2.40], indicating that the magnitude of asymmetry differed depending on which snack was chosen.

Discussion

In Study 2a, participants made more asymmetric preference inferences for accepting versus switching from the default when the chosen snack was an *a priori* dispreferred snack (broccoli) than when it was an *a priori* preferred snack (chocolate). As predicted, when the

characters chose broccoli, participants inferred weaker preferences when the characters accepted the default than when they switched away from it, but this asymmetry disappeared when chocolate was chosen. These findings support our causal reasoning account, suggesting that endorsement — the default-setter's recommendations — served as a plausible alternative explanation for accepting the default, therefore leading to asymmetric inferences when this explanation was most plausible. When the endorsement aligned with the character's expected preferences, it did not present an alternative explanation, and this asymmetry did not occur. This provides initial evidence that participants' preference inferences in the context of defaults reflect a causal reasoning process that takes into account not only physical but also social constraints in the choice environment, and these predict and explain when asymmetric preference inferences will versus will not occur.

One limitation of Study 2a is a potential ceiling effect: participants inferred extremely strong preferences for chocolate when the characters chose chocolate, regardless of whether they accepted (M = 3.87) or switched from the default (M = 3.84). On the one hand, these similarly high inferred preferences are consistent with our prediction that the asymmetry should be reduced in the absence of a plausible alternative explanation for choice. On the other hand, these ceiling effects may have masked any residual asymmetry, if there were any. We reasoned that using a more continuous dependent measure with greater range would give us more power to detect this asymmetry, should it exist. We implemented this dependent measure in Study 2b.

Study 2b

To address the potential ceiling effects in Study 2a, we replicated Study 2a with a continuous dependent measure in Study 2b.

Methods

Participants

As preregistered, $N = 120$ adults from the U.S. participated, recruited from www.prolific.com ($M_{Age} = 37.21$, $SD_{Age} = 11.68$, Range = 19-79; 49 men, 69 women, 2 non-binary gender). Thirty-six additional participants were also tested, but excluded and replaced due to the preregistered exclusion criterion of not passing one or both attention checks.

Design, Procedure and Stimuli

The design and the procedure in Study 2b were identical to Study 2a, except that the binary forced-choice measures of inferred preference were replaced by a continuous slider scale. The scale ranged from -100 to 100, where -100 was the character "likes broccoli a lot more than chocolate", 0 was the character "likes both snacks equally", and 100 was the character "likes chocolate a lot more than broccoli". The slider handle was initially positioned at 0, and participants moved the slider to indicate their inferred preference.

Results

As in Study 2a, we fit a linear mixed-effects regression predicting the absolute value of Inferred Preference from Default, Choice, and their interaction, with random intercepts for participants. In line with our predictions and the results in Study 2a, there was a significant interaction between Default and Choice ($b = -33.88$, $SE = 3.69$, $t\,(357) = -9.18$, $p < .001$, *95% CI* = [-41.09, -26.66]), indicating that the extent of asymmetry in inferred preference between default trials (Accept vs. Switch) differed based on which snack was chosen. A nested model comparison confirmed that the interaction term significantly improved model fit, $\chi^2\,(1) = 76.33$, $p < .001$.

To examine the magnitude of asymmetry for each chosen snack, we conducted paired-samples *t*-tests on inferred preferences across the two trials when each snack was chosen.

Replicating Study 2a, participants inferred weaker preferences in Accept trials than in the Switch trials when broccoli was chosen, and this asymmetry diminished when chocolate was chosen (see Figure 2.4). When broccoli was chosen, participants inferred significantly weaker preferences in the Accept trials ($M = -25.83$, $SD = 59.78$) than in the Switch trials ($M = -70.23$, $SD = 40.77$), $t(119) = 7.40$, $p < .001$, $d = 0.68$, *95% CI* = [0.48, 0.88], indicating strong asymmetry. As predicted, this asymmetry was diminished when chocolate was chosen. Although participants again inferred significantly weaker preferences in the Accept trials ($M = 81.40$, $SD = 25.53$) than in the Switch trials ($M = 85.54$, $SD = 20.19$), this difference was smaller in magnitude, $t(119) = 2.32$, $p = .02$, $d = 0.21$, *95% CI* = [0.05, 0.37], indicating weaker asymmetry.

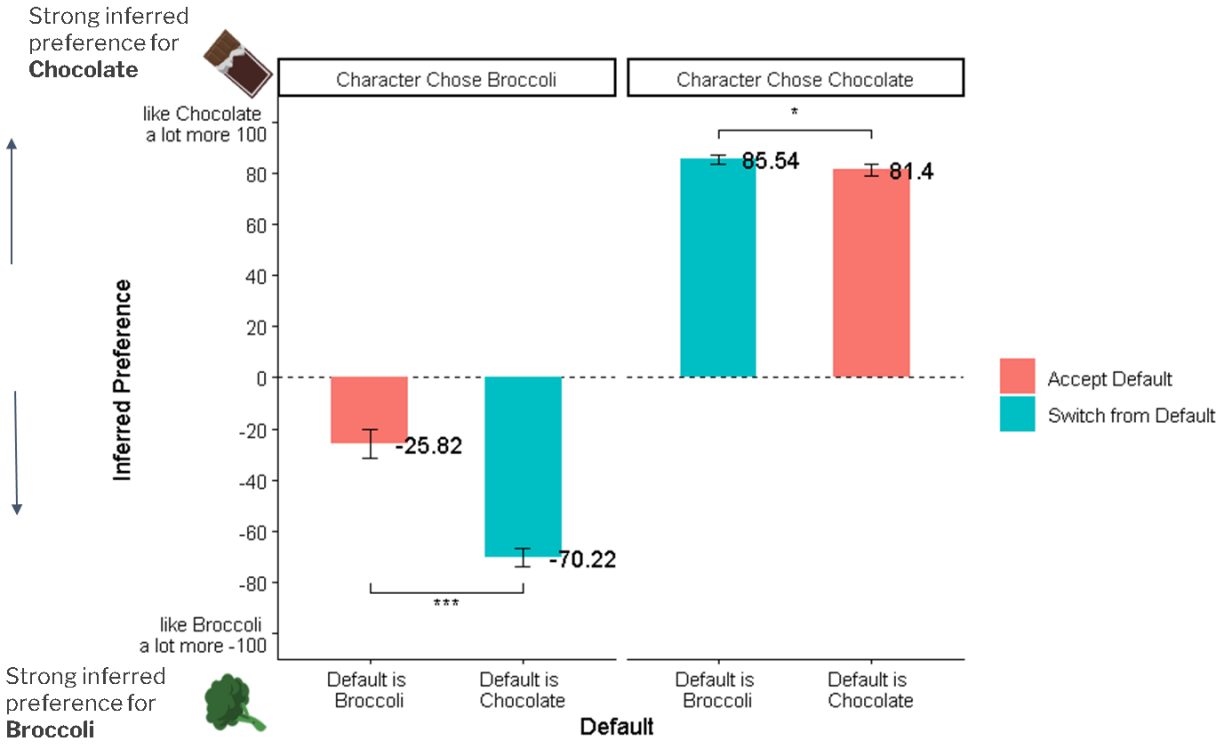


Figure 2.4: Average inferred preference across Default conditions in Study 2b. Replicating Study 2a, participants made asymmetric inferences when broccoli was chosen, and this asymmetry diminished when chocolate was chosen. Error bars represent ±1 *SE*.

An exploratory analysis confirmed that the asymmetry diminished when chocolate was chosen. We fit a linear mixed effects model predicting the absolute value of Inferred Preference from Choice, *Chosen Snack* (broccoli vs. chocolate), and their interaction with random intercepts for participants. We found a significant interaction between Choice and Chosen Snack, $b = -15.44$, $SE = 3.69$, $t(357) = -4.19$, $p < .001$, *95% CI* = [-22.66, -8.22], indicating that the magnitude of asymmetry differed depending on which snack was chosen.

Discussion

Study 2b replicated the findings in Study 2a using a more continuous measure of inferred preference. When the characters chose the food that was expected to be dispreferred *a priori* (broccoli), participants inferred weaker preferences when the characters accepted the default than when they switched away from it, but this asymmetry diminished substantially when the characters chose a food that was expected to be preferred *a priori* (chocolate).

In sum, together with Studies 1 and 2a, Study 2b provided converging evidence in favor of the causal reasoning account. That is, when inferring preferences from default options, people integrate both physical and social features of the choice environment to infer the best explanation for others' choices, and this can predict and explain when people make asymmetric preference inferences based on observed choices.

In all three studies, we made explicit what the alternative explanations a default option provides for accepting the default. For instance, in Study 1, the default option is explicitly operationalized as the grocery item closest to the customer. In Studies 2a and 2b, the default option is explicitly described as the snack that the parents chose for the children at school. However, in many real-world situations, default options are simply presented as pre-selected options without explicit explanations for why they are selected. To test whether the predictions

of our causal reasoning account generalize to these more typical formats of default options, we conducted further experiments (Studies 3 and 4).

Study 3

Study 3 tests whether the findings from Study 1 generalize to more typical default options, in which the default is presented simply as a pre-selected option, without further explanation. As in Study 1, the default option required less effort to obtain than the alternative, and across conditions, we manipulated the additional effort required to switch. To test if these effects generalized to more traditional defaults, the situation presented was choosing an appointment time from one of two options, one of which was pre-selected. Switching to the other required taking actions that required differential amounts of effort (sending a one-word text vs. making a phone call). We also included a no-cost condition in which no default was set, and options could be equally-easily chosen; this served as a baseline for comparison.

We also included manipulation checks to assess whether participants shared our prior assumptions about the relative amount of effort required to switch from the default in each Cost condition. We expected that participants would believe *a priori* that for most other people, sending a one-word text is easier than making a phone call; having this expectation is crucial for the effectiveness of our Cost manipulation. We analyzed our data both with the full sample, and also with a restricted sample of only participants who passed this manipulation check.

If the findings from Study 1 generalize to this more traditional default situation, participants should make more asymmetric preference inferences when switching away from the default is costly, and this asymmetry should vary parametrically with switching cost. When the cost of switching is negligible, the asymmetry in preference inference should be reduced.

The particular stimuli used in Studies 3 and 4 were informed by an additional preregistered study and pilot studies, the findings of which highlighted methodological

limitations. These studies led us to make the following design choices: (1) including illustrations of children making the choice in Study 4, and (2) counterbalancing the endpoints of the dependent measure in both Studies 3 and 4. We describe these additional findings in detail in the Supplemental Materials (Study 2c), and on OSF (https://osf.io/97zb3/files/rpncx?view_only=2d808a5930fc4ab5b722a23279e3b2ce).

Methods

Participants

As preregistered, $N = 120$ participants from the U.S. participated, recruited from www.prolific.com ($M_{Age} = 39.32$, $SD_{Age} = 14.15$, Range = 18-74; 66 men, 52 women, 2 non-binary gender). Six additional participants were also tested, but excluded and replaced due to the preregistered exclusion criterion of not passing the LLM or bot authenticity checks on Prolific. One participant did not receive a result for the LLM authenticity check due to a Prolific technical issue, but they passed the bot authenticity check. We did not exclude this participant.

Design

The experiment used a 3 (Cost: No, Low, High) by 2 (Choice: Accept vs. Switch) within-subject design, where each participant participated in six trials, and each trial was a unique condition. The trials were grouped into three pairs by Cost condition, with each pair containing both Choice trials (an Accept trial and a Switch trial). The order of the Cost conditions was counterbalanced across participants, and the order of Choice conditions was counterbalanced within each Cost condition.

Procedure and Stimuli

In each trial, participants read about a character choosing between two time slots for an event. Each character was automatically assigned a time slot – which presented a default option in its typical format. The character could switch to the other time slot with some effort, which

varied by Cost condition: by sending a one-word text (Low-Cost condition), or by calling (High-Cost condition). We also included a No-Cost condition in which there was no default option – the character simply chose a time slot. This condition served as a baseline.

Across all trials, participants saw three types of events: a dentist appointment, a doctor's appointment, and a class. The character chose between a morning and an afternoon time slot for each event. We used a Latin square design to counterbalance the correspondence between event type and Cost conditions, as well as the order in which Cost conditions were presented. This resulted in three between-subject presentation orders, which were randomly assigned to participants. Each participant saw the event type in the same order (dentist's appointment, doctor's appointment, class), but the corresponding Cost condition varied according to the Latin square.

Within each Cost condition, participants were asked to imagine the character in two situations: one in which they chose the default option (Accept trial), and one in which they chose the alternative option (Switch trial). The order of these two trials was counterbalanced across participants.

After each trial, participants inferred the extent to which the character preferred one time slot over the other on a scale of -100 to 100. Based on pilot findings that suggested some participants had a rightward response bias (see Supplemental Materials & OSF), the endpoints of the scale were also counterbalanced across participants. For half of the participants, -100 indicated that the character 'likes [the default time slot] a lot more' and 100 indicated that the character "likes [the alternative time slot] a lot more" (these numeric values were displayed on the left and right, respectively); for the other half, this mapping was reversed. The slider handle

was initially positioned at 0, indicating that the character 'likes both equally', and participants moved the slider to indicate their inferred preference.

After completing both trials in one Cost condition, participants proceeded to the remaining Cost conditions, which were identical in structure, but described new characters and different event types, and different switching costs.

Finally, after completing all trials, participants completed a manipulation check that assessed their beliefs about others' priors about the relative effort required to call versus text as a means of rescheduling appointments, on a scale of 1 to 5. A score of 1 indicated that "almost all people find making a phone call more effortful", 2 that "most people find making a phone call more effortful", 3 that "the same number of people find either one more effortful", 4 that "most people find sending a text message more effortful", and 5 that "almost all people find sending a text message more effortful".

Results

We first examined participants' responses to the manipulation check to assess whether they perceived the intended differences in switching costs. The majority of participants (87.5%, $N = 105$) met the preregistered criteria, defined as selecting either 1 or 2, indicating that they expect most people to find calling more effortful than texting. Here, we report the results with the restricted sample of participants. Results with the full sample qualitatively replicated the results with the restricted sample, and are reported in Supplemental Materials.

We fit a linear mixed-effects regression predicting the absolute value of the Inferred Preference from Cost, Choice, and their interaction, with random intercepts for participants. Cost was coded as numerically (No Cost = 0, Low Cost = 1, High Cost = 2). As predicted, there was a significant interaction between Cost and Choice ($b = 16.01$, $SE = 2.08$, $t\,(522) = 7.72$, $p < .001$,

95% CI = [11.95, 20.08]), indicating that the asymmetry in the strength of inferred preference between Accept and Switch trials increased parametrically with switching cost.

To examine the asymmetry within each cost condition without assuming linearity, we conducted follow-up analyses treating Cost as a categorical variable. Comparisons based on estimated marginal means revealed that the difference in inferred preference between Accept and Switch trials increased across cost conditions (see Figure 2.5). As expected, there was no significant difference in the No Cost condition ($b = 1.11$, $p = .70$). In contrast, participants inferred significantly stronger preferences in the Switch trials than in the Accept trials in the other two Cost conditions (Low: $b = -15.57$; High: $b = -30.91$; both ps $< .001$; see Table 2.2) indicating a parametric increase in asymmetry as switching cost increased.

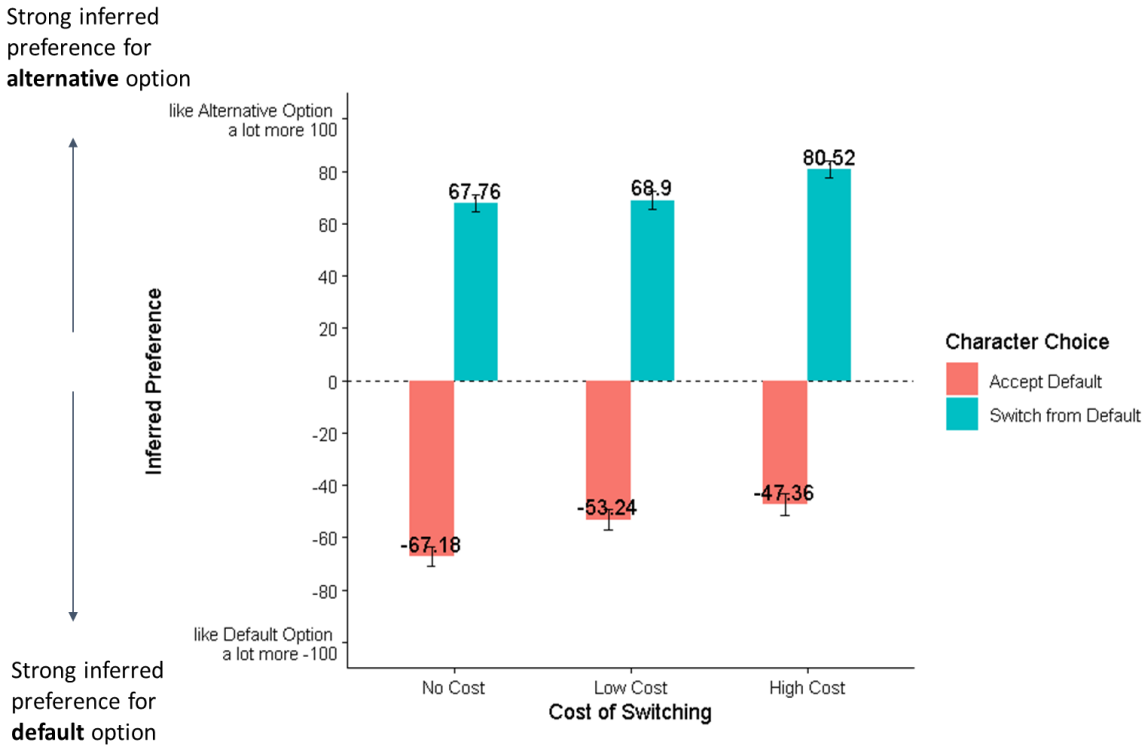


Figure 2.5: Replicating Study 1, the asymmetry in inferred preferences increased parametrically with switching cost. Error bars represent ±1 *SE*. The sample includes participants who passed the manipulation check (N = 105).

Table 2.2: Difference in inferred preference strength (Stay − Switch) across Cost conditions in Study 3. Negative values indicate stronger inferred preference in the Switch trials than in the Accept trials.

Cost Condition	**Difference (b)**	**SE**	**95% CI**	**t**	**p**
No-Cost	1.11	2.93	[−4.64, 6.86]	0.38	.69
Low-Cost	−15.57	2.93	[−21.32, −9.82]	−5.32	< .001
High-Cost	−30.91	2.93	[−36.66, −25.16]	−10.56	< .001

Discussion

Study 3 replicated the findings of Study 1. Participants made increasingly asymmetric preference inferences between accepting versus switching away from the default as the cost of switching increased. These findings suggest that our causal reasoning account applies to more typical default formats, in which the default option is presented as a pre-selected option without explicit explanation. In particular, participants appear to spontaneously infer cost avoidance as a plausible alternative explanation for accepting the default, even when this explanation is not explicitly stated. This provides evidence for the generalizability of the causal reasoning account to choice contexts in which the reason why a default option is set is unspecified.

Study 4

Study 4 tests whether the findings from Study 2 generalize to more typical, unexplained default options. As in these studies, participants read about children choosing between broccoli and chocolate, one of which was the default. The default option in this situation can be seen as implying an implicit recommendation by the default-setter (e.g., a recommendation to eat healthy, when broccoli is the default; or licensed indulgence, when chocolate is the default; McKenzie, 2006). Unlike in Studies 2a and 2b, the recommendation is not explicitly stated, but is fully implicit: The default is simply presented as a pre-selected option, and it is unspecified

who the default-setter was or how the default was determined. In this way, the alternative explanation for accepting the default—following the default-setter's recommendations—is implied rather than explicitly described.

As in Study 3, we included a manipulation check to assess whether participants shared our prior assumptions about the DM's prior food preferences. We expected that participants would *a priori* believe that for most children, chocolate is preferred over broccoli; having this expectation is crucial for the effectiveness of the task in testing the predictions of our model (which in this case depends on starting with a high prior of liking one of the two available choices). We again analyzed our data both with the full sample, and also with a restricted sample of only participants who passed this manipulation check.

If the findings from Study 2 generalize, then when the chosen item is the a priori dispreferred option (broccoli), participants should infer stronger preferences when the DM switches from the default than when they accept it. This asymmetry should decrease when the chosen option is the a priori preferred option (chocolate), resulting in decreased asymmetry in inferred preferences across the two choice conditions.

Methods

Participants

As preregistered, $N = 120$ adults from the U.S. participated, recruited from www.prolific.com ($M_{Age} = 40.97$, $SD_{Age} = 13.64$, Range = 18-74; 57 men, 61 women, 1 non-binary gender). One additional participant was also tested, but excluded and replaced due to the preregistered exclusion criterion of not passing the LLM or bot authenticity checks on Prolific.

Design

The design was the same as in Study 2b.

Procedure and Stimuli

In each trial, participants read and viewed illustrated vignettes depicting first-grade child characters choosing between two snacks at school. Illustrations were created using ChatGPT. In these vignettes and illustrations, a child stood near an iPad, which displayed images of broccoli and chocolate side by side, with broccoli always presented on the left and chocolate on the right.

One of the two snacks was pre-selected as the default. In the Broccoli Default condition, broccoli was pre-selected with a red circle (see Figure 2.6); in the Chocolate Default condition, chocolate was pre-selected with a red circle. In both conditions, the default option was accompanied by a green checkmark underneath, and the alternative option was accompanied by a blue arrow underneath. Participants read that "the kids can either click the green checkmark button to confirm the current selection, or they can click the blue arrow button to switch to the chocolate". They also learned that "the kids are free to choose whichever snack they want".

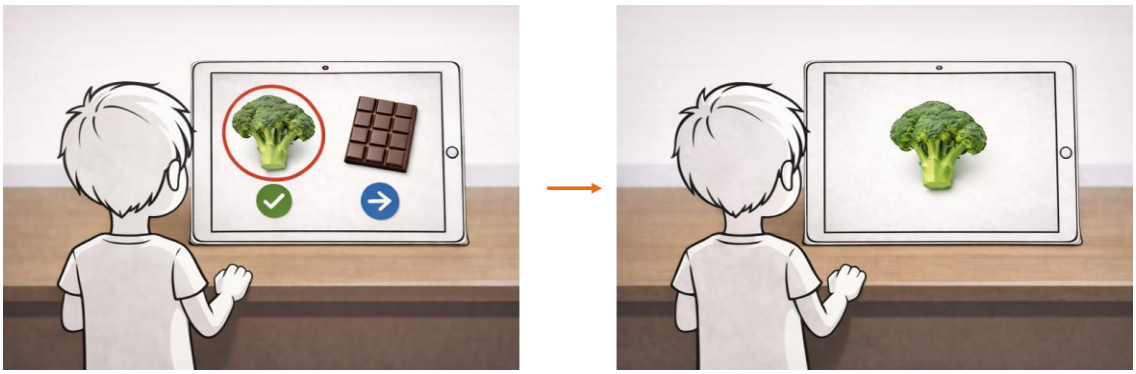


Figure 2.6: Illustration of the default and the character's choice in Study 4. The snack pre-selected with a red circle was the default option in Study 4. On each trial, one character chose the default option (pictured); the other chose the non-default option. Images generated in part by ChatGPT.

Within each Default condition, participants saw each character in two situations: one in which they chose the default option (Accept trial), and one in which they chose the alternative

option (Switch trial). The character's choice was illustrated alongside the display of the default selection. The order of Accept and Switch trials was counterbalanced across participants.

After each trial, participants inferred the extent to which the character preferred one snack over the other, on a scale of -100 to 100. Based on pilot findings that suggested some participants had a rightward response bias (see Supplemental Materials & OSF), the endpoints of the scale were counterbalanced across participants, as in Study 3: For half of the participants, -100 indicated that the character "likes broccoli a lot more than chocolate", and 100 indicated that the character "likes chocolate a lot more than broccoli"; for the other half, this mapping was reversed. The slider handle was initially positioned at 0, indicating that the character 'likes both snacks equally', and participants moved the slider to indicate their inferred preference.

After completing both trials in one Default condition, participants proceeded to the other Default condition, which was identical except for depicting new child characters, and the opposite snack marked as the default.

Finally, after completing all trials, participants completed manipulation checks that assessed their beliefs about first-graders' relative preferences for chocolate and broccoli. Specifically, participants rated on a scale of 1 to 5. A score of 1 indicated that "almost all kids prefer chocolate", 2 that "most kids prefer chocolate", 3 that "the same number of kids prefer either", 4 that "most kids prefer broccoli", and 5 that "almost all kids prefer broccoli".

Results

We first examined participants' responses to the manipulation check to assess whether they had the expected a priori expectations about children's typical preferences. The majority of participants (85%, $N = 102$) met the preregistered criteria, defined as selecting either 1 or 2, indicating that they expected most first-graders to prefer chocolate over broccoli. The analyses

below report the results from this restricted sample; results with the full sample were qualitatively similar, and are reported in Supplemental Materials.

As in Study 2b, we fit a linear mixed-effects regression predicting the absolute value of Inferred Preference from Default, Choice (Accept or Switch), and their interaction, with random intercepts for participants. Replicating Study 2b, there was a significant interaction between Default and Choice ($b = -26.29$, $SE = 4.10$, $t(303) = -6.42$, $p < .001$, *95% CI* = [−34.31, −18.28]), indicating that the strength of asymmetry in inferred preference between Accept and Switch trials differed based on which snack was chosen.

To examine the magnitude of asymmetry for each chosen snack, we conducted paired-samples *t*-tests on inferred preferences across the two trials when each snack was chosen. However, contrary to our predictions and previous findings, participants did not make significantly asymmetric preference inferences in either case – either when broccoli was chosen, or when chocolate was chosen (see Figure 2.7). When broccoli was chosen, participants inferred numerically weaker preferences in the Accept trials ($M = -48.45$, $SD = 42.42$) than in the Switch trials ($M = -56.08$, $SD = 56.30$). However, this difference was not significant, $t(101) = 1.52$, $p = .132$, $d = 0.15$, *95% CI* = [− 0.04, 0.38], indicating little to no asymmetry.

When chocolate was chosen, participants again inferred numerically weaker preferences in the Accept trials ($M = 58.00$, $SD = 53.69$) than in the Switch trials ($M = 70.88$, $SD = 48.02$). Although this difference (mean difference = 12.88) was numerically larger than when broccoli was chosen (mean difference = 7.63), it was not statistically significant, $t(101) = -1.76$, $p = .082$, $d = -0.17$, *95% CI* = [−0.38, 0.02], indicating little to no asymmetry.

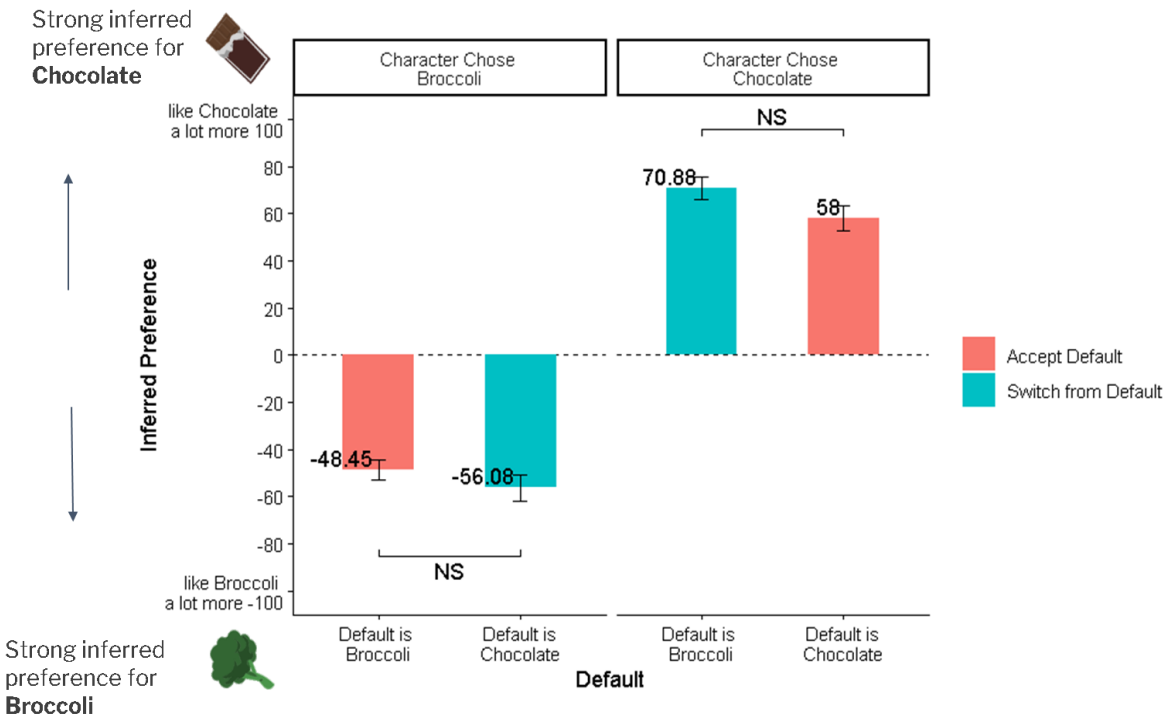


Figure 2.7: Average inferred preference across Default conditions in Study 4. Contrary to our predictions and Studies 2a and 2b, participants did not make asymmetric inferences when broccoli was chosen, nor when chocolate was chosen. Error bars represent ±1 *SE*. The sample includes participants who passed the manipulation check (N = 102).

Discussion

Contrary to our predictions and previous findings, participants in Study 4 did not make asymmetric preference inferences in any case, regardless of which snack was chosen. Notably, they did not infer weaker preference for broccoli for children who accepted the broccoli default than for those who switched from the chocolate default, even though the broccoli default was meant to provide a plausible alternative explanation for choice.

Although these results do not support our predictions, they may help clarify the assumptions that underlie our causal reasoning account. In our account, observers undergo an inverse decision-making process in which they update their beliefs about a DM's preference based on the extent to which the default can explain away preference. In this study, two

components of this reasoning process may have differed from our assumptions, leading to a lack of asymmetry.

One possibility is that participants did not begin with a strong prior that children prefer chocolate to broccoli, but instead might have expectations that others vary in their preferences, or value these choices equally. In this case, preferences should be similarly asymmetric for both items chosen, whether the chosen item was chocolate or broccoli. This seems unlikely to be the case for two reasons: Firstly, our manipulation checks show that most participants do expect that children prefer chocolate to broccoli; and the analyses reported above included only participants who passed these manipulation checks, and thus believe that most children prefer chocolate to broccoli. Second, if participants have a low prior on preference for chocolate, we should see asymmetric preferences in both conditions, but we instead see significant asymmetry in neither condition.

A more likely possibility is that participants perceived our operationalization of the default to have very low strength, δ. That is, people may have understood that one option was pre-selected by marking it with a red circle, but may have not expected that children would feel any pressure to accept this default option, and instead expected most children to ignore this default option entirely. In this case, when delta is low, the situation becomes analogous to a free choice – in which choice reflects preference, because choices are made regardless of the default. If our participants interpreted our stimuli as implying extremely low pressure from the default (low δ), then the predicted asymmetry should diminish across both Choice (Accept and Switch) conditions, as was the case in our results.

It is thus likely that participants in Study 4 interpreted the default as less impactful than we intended. Why would this occur? Our predictions were based on the expectation that the

default-setter would be perceived as a benign authority, so that the broccoli default would be seen as a recommendation worth following (McKenzie, 2006). However, since the default-setter was unspecified in this study, it is possible that participants interpreted the default-setter as a gesture towards some general health norms rather than a specific recommendation from a trusted authority. If so, then participants may have had little reason to believe that first graders (the characters in the vignettes) would care much about such generic normative endorsement. In this case, endorsement would not explain away preference as a reason accepting the broccoli default, which may explain why participants did not infer weaker preferences for broccoli for children who accepted the broccoli default than for those who switched from the chocolate default. Thus, while the results of Study 4 differ from our predictions, the framework of the causal reasoning model remains helpful in explaining potential reasons for reduced asymmetry of preference inferences in the current task.

General Discussion

Why do people infer weaker preferences when they observe others accept, rather than switch away from, a default option? Although prior work has documented this asymmetric preference inference (Davidai et al., 2012; Leong et al., 2020; Lin et al., 2018), existing accounts do not explain when and why this asymmetry occurs, and importantly, when and why it diminishes. In this paper, we find evidence for a causal reasoning account: Asymmetric preference inferences appear to reflect a process of causal inference and Bayesian inverse decision-making (Tenenbaum et al., 2006, Jern et al., 2017). In line with this account, we find that asymmetric preference inferences occur only when defaults provide strong plausible alternative explanations for choice other than preference, rendering the choice less diagnostic of preference. Crucially, the asymmetry diminished when no such explanations are available and the choice remains diagnostic of preference.

Across five preregistered studies, we found support for this causal reasoning account of asymmetric preferences, and also identified limits on when this account generalizes. We tested the predictions of our account using default options that provided two kinds of alternative explanations: avoiding switching cost and following the default-setter's recommendations. We also tested these predictions across two default formats: one in which the alternative explanation was directly stated, and a more typical format in which the alternative explanation was implicit, such as a pre-selected option.

Studies 1 and 3 provided support for the causal reasoning account using ease-based defaults. In Study 1, participants observed characters choosing between different flavored snacks that varied in physical accessibility, with the default flavor presented as the easiest one to obtain. As predicted, participants inferred weaker preferences when the character accepted rather than switched from the default, and the strength of this asymmetry increased parametrically with switching cost. Study 3 replicated this pattern of results in a different choice context, using a more typical default – where the default was simply a pre-selected option. Together, these findings show that people can incorporate physical features of the choice environment to infer the most likely explanation for a DM's choice.

Studies 2 and 4 provided mixed evidence for our predictions using endorsement-based defaults, but the findings in both studies are consistent with the causal reasoning account. In Study 2, participants observed child characters choosing between broccoli and chocolate, with the default presented as the option the parents' chose for the children. As predicted, participants inferred weaker preference when the character accepted rather than switched from the default, but only when broccoli was chosen. However, in Study 4, when defaults were presented as unexplained pre-selections, participants did not make asymmetric inferences regardless of which

snack was chosen. One possibility is that participants did not see the default in Study 4 as impactful (a default strength, δ, that is close to zero), and therefore the manipulation did not work as intended. If this is the case, the default would have a negligible impact on choice, making the situation resemble a free choice in which the child's choice simply reflects their preference. Thus, the findings in Studies 2 and 4 are consistent with the causal reasoning account. They suggest that people can integrate social features of the choice environment into their reasoning about others' preferences, and part of that reasoning process is inferring the strength of a default's impact on choice.

Overall, our findings provide evidence that asymmetric preference inferences from default options reflect a process of rational causal inference.

When asymmetric inferences occur and diminish

Our results broadly support the causal reasoning account in two key ways. First, we clarified when the classic asymmetric preference inference occurs (Davidai et al., 2012; Leong et al., 2020; Lin et al., 2018), replicating this effect in situations where the default explicitly provided a plausible alternative explanation for choice. In line with prior work on defaults, we found this effect when the alternative explanations were based on both physical ease and social endorsement (Jachimowicz et al., 2019).

Second, and more crucially, we showed when asymmetric inferences diminish. In situations where defaults provide weak alternative explanations for choice, the asymmetry diminished or disappeared. This novel finding supports our prediction that people engage in structured causal reasoning about how defaults relate to preferences and choices: Accepting a default only "explains away" preference as the cause of a choice when the default provides an alternative explanation for that choice (as in Pearl, 2000; Tenenbaum et al., 2006).

Generalization from explicit to implicit explanations

Participants perceived strong default strength when we explicitly specified the alternative explanations that a default could provide, but they did not do so consistently when these alternative explanations were implicit. This contrast is important because in many real-world settings, default options are unaccompanied by an explicit rationale. Instead, observers must infer whether the default reflects cost avoidance, a desire to follow social norms, or some other cause. Our findings highlight some potential factors that may determine whether participants perceive a default as having strong or weak explanatory force over a DM's choice.

When defaults provide ease-based explanations, it seems relatively straightforward to infer ease of choice from a pre-selected option. Our findings supported the causal reasoning account both when switching costs were explicit (Study 1) and implicit (Study 3). This may be because switching costs can be relatively easy to infer from the structure of the choice itself: if switching requires obvious effort (e.g., calling to reschedule an appointment), accepting the default can be explained by effort avoidance.

However, when defaults provide endorsement-based explanations, it appears to be more complicated to infer the default-setter's recommendations from a pre-selected option. We found support for the casual reasoning account when the recommendation of the default option was made explicit (Study 2); but there was no effect of default on preference inferences when the recommendation was implicit (Study 4). Contrary to prior research suggesting that defaults often work because they are seen as implicit recommendations (McKenzie et al., 2006), participants did not seem to perceive the default in Study 4 as a recommendation, or at least not one that exerted strong pressure for choosing it. This may be because inferring implicit recommendation involves additional inferences about the default-setters' intentions and their relationship to the DM, which may be difficult when the default is simply an unspecified pre-selected option.

Together, our findings provide strong evidence for the causal reasoning account across different kinds of defaults and default formats. They also reveal social and contextual factors that may shape this inverse-decision making process, in particular, whether observers perceive the default as providing strong alternative reasons for a DM's choice.

More broadly, our results provide a theoretical framework for understanding when default options will or will not shape social inferences. They also generate novel predictions about when the asymmetry may occur, diminish, or even reverse in contexts beyond the ones tested here. For instance, defaults may provide alternative explanations *not* for accepting the default, but for switching away from the default. Consider a case in which the default-setter is untrustworthy, such as a large pharmaceutical company. In this case, the default may provide a strong reason to switch away regardless of preference. Here, switching from the default becomes less diagnostic of preference, whereas accepting the default remains diagnostic, leading to an asymmetry in the opposite direction from the asymmetry examined here and in prior work. Furthermore, this reversal may depend on whether distrust of the default-setter is made explicit, or whether observers can infer it from the choice context.

On the other hand, there are many other contexts in which the default does not provide a strong alternative explanation for choice. For instance, when a default reflects a status quo, such as continuing a prescription, Jachimowicz et al., 2019), or when the default-setter's recommendations is irrelevant to the DM, such as accepting the default seat selection on a flight, the default strength may be low and may therefore not explain away preference.

Future work could investigate how different contexts shape whether asymmetric preference inferences occur, diminish, or reverse — and whether these patterns depend on

alternative explanations being explicitly stated or implicitly inferred from the choice environment.

Developmental Relevance

The causal reasoning account also raises intriguing questions about the development of social inference from defaults. Prior work has shown that by ages 5-6, young children can reason about how physical and epistemic constraints influence others' choices when inferring preferences (Pesowski et al., 2016). An open question is whether children can similarly reason about how social constraints, like default options, provide alternative explanations for others' choices. Default options have been shown to impact young children's behavior (Zhao et al., 2023), but less is known about whether children take into account the social meaning of defaults to infer others' preferences. We are currently examining whether children also consider default options when inferring preferences, and whether they show adult-like patterns of asymmetric inferences. In a preregistered study with 7-to-9-year-old children (N = 120), children showed the opposite pattern of responses than adults in Study 2: they made strongly asymmetric inferences when chocolate was the chosen snack, but not when the broccoli was the chosen snack (reproducible code and results available here: https://osf.io/qrmfv/?view_only=c134bf1f25eb4d00a5450294146a0a04). Interestingly, this pattern of results resembles the adult pattern in Study 4, where defaults were presented as unspecified pre-selected options. One possibility is that children have different priors about other children's food preferences. Alternatively, they may perceive default strength differently than adults do, even when the explanation provided by the default is made explicit. We are exploring these possibilities in ongoing work.

Future directions: Other domains and other default-setters

Another direction for future work is to test the causal reasoning account across different domains. While most of our studies focused on food choices (Studies 1, 2 and 4), Study 3 explored appointment scheduling, suggesting that the causal reasoning framework may generalize to other domains than food. Future research can explore domains such as clothing, experiences, payment plans, or other consumer decisions that may involve defaults in more complex social situations. Different domains may involve different priors, as well as different kinds of alternative explanations for choice that a default can provide. For instance, a default clothing pairing may suggest recommendations based on popularity or expertise, rather than an injunctive social norm.

In addition to different choice domains, future research can also examine different kinds of default-setters. The identity and assumed intention of the default-setter is key for interpreting default strength, which in turn impacts reasoning about endorsement-based defaults. While we explicitly described the default-setter as benign individuals (i.e., parents) in Study 2, many real-world defaults are set by large institutions. For instance, governments set organ-donation policies, energy companies set energy-use defaults, and companies set employees' retirement-savings defaults (Ebeling & Lotz, 2015; Madrian & Shea, 2001). Defaults set by institutions may not feel as intentional, personal, or trustworthy as defaults set by known individuals, and may therefore have lower perceived default strength.

Study 4 begins to explore this distinction by introducing defaults that are likely set by institutions – food choices in the cafeteria are likely set by the school or the catering company, versus defaults set by directly-known individual social partners (such as parents). The lack of asymmetric inferences in Study 4 is consistent with the possibility that institutional defaults may provide weaker endorsement-based explanations for choice compared to defaults set by known

individuals. Future research should extend this comparison to children, who slowly develop an understanding of the nature of institutions over middle childhood (Jara-Ettinger & Dunham, 2025; Noyes et al., 2020, 2023).

Conclusion

In conclusion, this work suggests that a classic effect in judgment and decision-making — asymmetric preference inferences from defaults — reflects a process of Bayesian causal inference based on inverse decision-making. Across five preregistered studies, people appeared to reason about the extent to which a choice was diagnostic of preference based on the extent to which defaults provide explanations for choice other than preference. As predicted by our account, they inferred asymmetric only when the default provides a plausible alternative explanation, and the asymmetry diminished or changed when such alternative explanations were weak or ambiguous. Our findings were consistent with the causal reasoning account across different kinds of defaults and default formats, and also revealed contextual factors that may influence people's perceived default strength when the rationale behind a default is implicit rather than explicit. This distinction highlights the complexity of drawing inferences from the choice environment in making such causal inferences.

In sum, this framework provides new insights into the cognitive processes underlying social inference from choice architecture, allowing for novel predictions about when and where default options will result in asymmetric preference inferences across a large range of contexts.

Acknowledgements

Chapter 2, is currently being prepared for submission for publication of the material. Liu, Xingyu S.; McKenzie, Craig R.M.; and Schachner, Adena. The dissertation author was the primary researcher and author of this material. The data in Study 2a were presented at the 2025

Annual Meeting of the Cognitive Science Society and published in the conference proceedings (see Liu et al., 2025).

REFERENCES

Davidai, S., Gilovich, T., & Ross, L. D. (2012). The meaning of default options for potential organ donors. *Proceedings of the National Academy of Sciences*, *109*(38), 15201–15205. https://doi.org/10.1073/pnas.1211695109

Dinner, I., Johnson, E. J., Goldstein, D. G., & Liu, K. (2011). Partitioning default effects: Why people choose not to choose. *Journal of Experimental Psychology. Applied*, *17*(4), 332–341. https://doi.org/10.1037/a0024354

Ebeling, F., & Lotz, S. (2015). Domestic uptake of green energy promoted by opt-out tariffs. *Nature Climate Change*, *5*(9), Article 9. https://doi.org/10.1038/nclimate2681

Gopnik, A., & Sobel, D. M. (2000). Detecting blickets: How young children use information about novel causal powers in categorization and induction. *Child Development*, *71*(5), 1205–1222. https://doi.org/10.1111/1467-8624.00224

Jachimowicz, J. M., Duncan, S., Weber, E. U., & Johnson, E. J. (2019). When and why defaults influence decisions: A meta-analysis of default effects. *Behavioural Public Policy*, *3*(2), 159–186. https://doi.org/10.1017/bpp.2018.43

Jara-Ettinger, J., & Dunham, Y. (2025). The institutional stance. *The Behavioral and Brain Sciences*, 1–62. https://doi.org/10.1017/S0140525X25103427

Jern, A., Lucas, C. G., & Kemp, C. (2017). People learn other people's preferences through inverse decision-making. *Cognition*, *168*, 46–64. https://doi.org/10.1016/j.cognition.2017.06.017

Leong, L. M., Yin, Y., & McKenzie, C. R. M. (2020). Exploiting asymmetric signals from choices through default selection. *Psychonomic Bulletin & Review*, *27*(1), 162–169. https://doi.org/10.3758/s13423-019-01699-2

Lin, Y., Osman, M., Harris, A. J. L., & Read, D. (2018). Underlying wishes and nudged choices. *Journal of Experimental Psychology: Applied*, *24*(4), 459–475. https://doi.org/10.1037/xap0000183

Liu X.S., & McKenzie, C.R., & Schachner, A.(2025). When default options explain away preferences: a causal reasoning account of mental state reasoning from default options. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 47. https://escholarship.org/uc/item/9bq3d5ch

Madrian, B. C., & Shea, D. F. (2001). The Power of Suggestion: Inertia in 401(k) Participation and savings behavior. *The Quarterly Journal of Economics*, *116*(4), 1149–1187.

McKenzie, C. R. M., Liersch, M. J., & Finkelstein, S. R. (2006). Recommendations implicit in policy defaults. *Psychological Science*, *17*(5), 414–420. https://doi.org/10.1111/j.1467-9280.2006.01721.x

Noyes, A., Dunham, Y., & Keil, F. C. (2023). Subjectivity and social constitution: Contrasting conceptions of institutions, artifacts, and animals. *Developmental Psychology*, *59*(2), 377–389. https://doi.org/10.1037/dev0001499

Noyes, A., Keil, F. C., & Dunham, Y. (2020). Institutional actors: Children's emerging beliefs about the causal structure of social roles. *Developmental Psychology*, *56*(1), 70–80. https://doi.org/10.1037/dev0000847

Pearl, J. (2000). *Causality: Models, reasoning, and inference* (pp. xvi, 384). Cambridge University Press.

Pesowski, M. L., Denison, S., & Friedman, O. (2016). Young children infer preferences from a single action, but not if it is constrained. *Cognition*, *155*, 168–175. https://doi.org/10.1016/j.cognition.2016.07.004

Sunstein, C. R., & Thaler, R. H. (2003). Libertarian paternalism is not an oxymoron. *The University of Chicago Law Review*, *70*(4), 1159–1202. https://doi.org/10.2307/1600573

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, *10*(7), 309–318. https://doi.org/10.1016/j.tics.2006.05.009

Zhao, L., Mao, H., Zheng, J., Fu, G., Compton, B. J., Heyman, G. D., & Lee, K. (2023). Default settings affect children's decisions about whether to be honest. *Cognition*, *235*, 105390. https://doi.org/10.1016/j.cognition.2023.105390

Chapter 2 Supplemental Materials

Study 1 Stimuli randomization

Supplemental Table 2.1 shows the Latin square design for counterbalancing the correspondence between food items and Cost conditions, as well as the order in which Cost conditions were presented in Study 1. This resulted in four between-subject presentation orders which were randomly assigned to participants (A, B, C, D). Each participant saw the food items in the same order (potato chips, chewing gum, chocolate bars, juice), and the corresponding Cost condition varied according to the Latin square.

Supplemental Table 2.1: Food items across cost conditions in Study 1. Each row in the table is one presentation order that a participant sees, in order from left to right.

	Potato Chips	**Chewing Gum**	**Chocolate Bars**	**Juice**
A	No Cost	High Cost	Medium Cost	Low Cost
B	Low Cost	No Cost	High Cost	Medium Cost
C	Medium Cost	Low Cost	No Cost	High Cost
D	High Cost	Medium Cost	Low Cost	No Cost

Study 3 Stimuli randomization

Supplemental Table 2.2 shows the Latin square design for counterbalancing the correspondence between food items and Cost conditions, as well as the order in which Cost conditions were presented in Study 3. This resulted in three between-subject presentation orders which were randomly assigned to participants (A, B, C). Each participant saw the event type in the same order (dentist's appointment, doctor's appointment, class), but the corresponding Cost condition varied according to the Latin square.

Supplemental Table 2.2: Event types across cost conditions in Study 3. Each row in the table is one presentation order that a participant sees, in order from left to right.

	Dentist's Appointment	**Doctor's appointment**	**Class**
A	No Cost	High Cost	Low Cost
B	Low Cost	No Cost	High Cost
C	High Cost	Low Cost	No Cost

Study 3 Results with full sample

We repeated the same analyses in the main text with the full sample, including participants who did not pass the manipulation checks (N = 120).

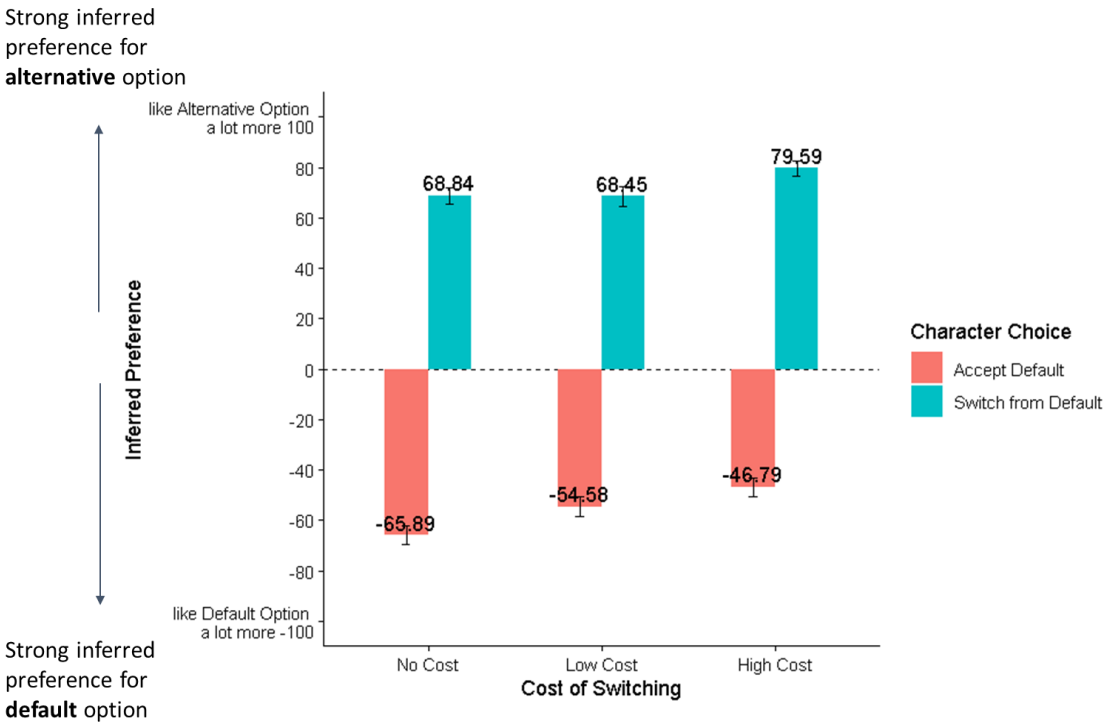
As in Study 1, we fit a linear mixed-effects regression predicting the absolute value of the Inferred Preference from Cost, Choice, and their interaction, with random intercepts for participants. Cost was coded as a linearly increasing ordinal variable (No Cost = 0, Low Cost = 1, High Cost = 2). As predicted, there was a significant interaction between Cost and Choice (b = 14.69, SE = 2.00, t (597) = 7.34, $p < .001$, *95% CI* = [10.77, 18.61]), indicating that the asymmetry in the strength of inferred preference between Accept and Switch trials increased parametrically with switching cost.

To examine the asymmetry within each cost condition without assuming linearity, we conducted follow-up analyses treating Cost as a categorical variable. Comparisons based on estimated marginal means revealed that the difference in inferred preference between Accept and Switch trials increased across cost conditions (see Supplemental Figure 2.1). As expected, there was no significant difference in the No Cost condition ($b = -1.13$, $p = .69$). In contrast, participants inferred significantly stronger preferences in the Switch trials than in the Accept trials in the other two Cost conditions (Low: $b = -16.32$; High: $b = -30.52$; $ps < .001$; see Supplemental Table 2.3), indicating a parametric increase in asymmetry as switching cost increased.

Overall, these findings qualitatively replicate the findings with the sample that passed the manipulation check.

Supplemental Table 2.3: Difference in inferred preference strength (stay − switch) across cost conditions in Study 3. Negative values indicate stronger inferred preference in the Switch trials than in the Accept trials.

Cost Condition	Difference (b)	SE	95% CI	t	p
No Cost	−1.13	2.83	[−6.69, 4.43]	−0.40	.69
Low Cost	−16.32	2.83	[−21.88, −10.76]	−5.76	< .001
High Cost	−30.52	2.83	[−36.08, −24.96]	−10.78	< .001



Supplemental Figure 2.1: Participants' average inferred preference across Cost conditions in Study 3. As in the sample who passed the manipulation check, the asymmetry in inferred preferences increased parametrically with switching cost. Error bars represent ±1 SE. The sample includes all participants, including those who did not pass the manipulation check (N = 120).

Study 4 Results with full sample (N = 120)

We repeated the same analyses in the main text with the full sample, including participants who did not pass the manipulation checks (N = 120).

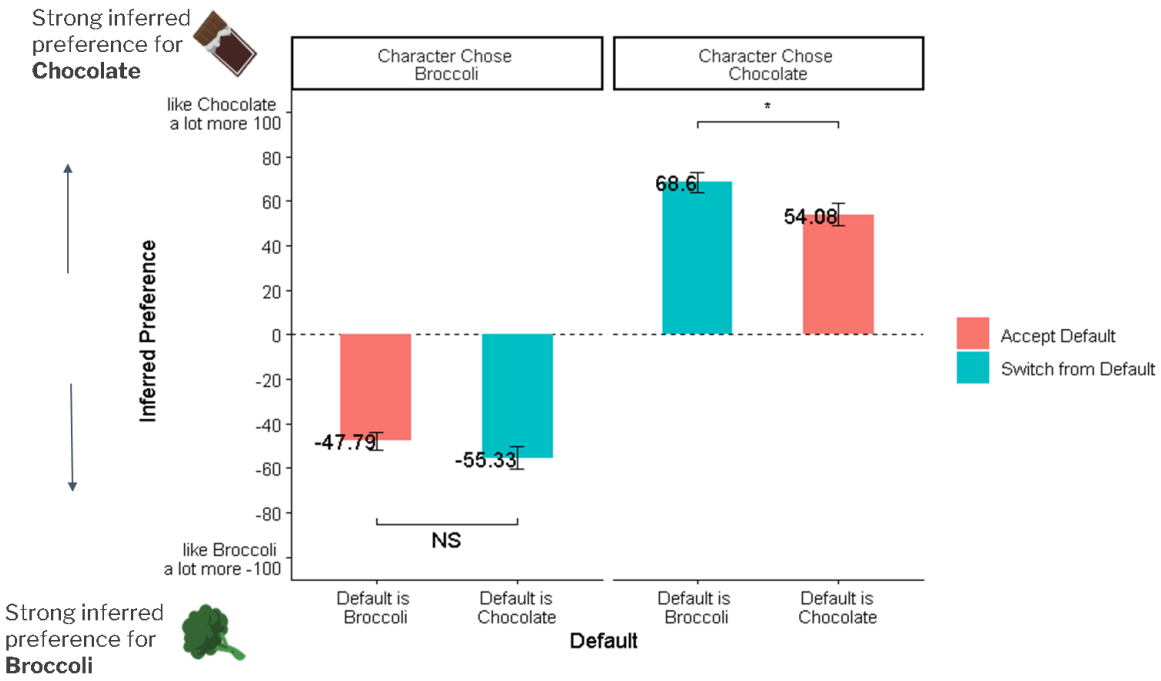
As in Study 2b, we fit a linear mixed-effects regression predicting the absolute value of Inferred Preference from Default, Choice, and their interaction, with random intercepts for participants. Replicating Study 2b, there was a significant interaction between Default and Choice, $b = -23.16$, $SE = 3.80$, $t(357) = -6.08$, $p < .001$, *95% CI* = [−30.61, −15.70], indicating that the strength of asymmetry in inferred preference between Accept and Switch trials differed based on which snack was the default.

To examine the magnitude of asymmetry for each chosen snack, we conducted paired-samples *t*-tests on inferred preferences across the two trials when each snack was chosen. However, contrary to our predictions and previous findings, participants did not make asymmetric preference inferences when broccoli was chosen, but they did when chocolate was chosen.

When broccoli was chosen, participants inferred numerically weaker preferences in the Accept trials ($M = -47.79$, $SD = 44.62$) than in the Switch trials ($M = -55.33$, $SD = 57.27$). However, this difference was not significant, $t(119) = 1.44$, $p = .151$, $d = 0.13$, *95% CI* = -0.04, 0.32], indicating little to no asymmetry.

When chocolate was chosen, participants now inferred significantly weaker preferences in the Accept trials ($M = 54.08$, $SD = 56.28$) than in the Switch trials ($M = 68.60$, $SD = 50.64$), $t(119) = -2.08$, $p = .040$, $d = -0.19$, *95% CI* = [-0.39, 0.00], indicating stronger asymmetry.

Overall, these findings qualitatively replicated the findings with the sample that passed the manipulation check.



Supplemental Figure 2.2: Average inferred preference across Default conditions in Study 4. Contrary to our predictions and Studies 2a and 2b, participants did not make asymmetric inferences when broccoli was chosen, and this (lack of) asymmetry increased when chocolate was chosen. Error bars represent ±1 SE. The sample includes all participants, including those who did not pass the manipulation check (N = 120).

Study 2c

The stimuli used in this study were informed by several pilot studies that were not themselves preregistered. Below, we briefly describe how these pilot studies shaped the final stimuli. In the pre-registration for this study (https://osf.io/5njau/overview?view_only=4961f9b67a7c4bea96947828211c0d00), we initially specified the stimuli in the study without prior piloting. After running a small pilot sample (N = 10), we found that the stimuli did not seem to effectively manipulate the variables we intended. We describe these issues in more detail in the updated pre-registration.

To address this, we ran a second pilot sample (N = 25), where we tested alternative stimuli and added a manipulation check to assess whether participants shared our prior assumptions about people's preferences (e.g., for specific foods or behaviors). These data

suggested that participants' assumptions were quite variable. Based on this, we chose to use stimuli that have produced more consistent patterns in prior studies (e.g., children choosing between broccoli and chocolate; rescheduling an appointment by calling vs. texting).

We then ran the full sample as specified in the update to the preregistration (N = 120) using this updated design. Methods and results are reported below:

Methods

Participants

N = 120 adults from the U.S. participated, recruited from www.prolific.com (M_{Age} = 41.39, SD_{Age} = 12.89, Range = 18-73; 70 men, 49women, 1 non-binary gender). 9 additional participants were also tested, but excluded and replaced due to the exclusion criterion of not passing the bot check.

Design

The study included two scenarios (Cost and Food scenarios), with a total of 10 within-subject trials. The Cost scenario used a 3(Cost: No, Low, High) by 2 (Choice: Accept vs. Switch) within-subject design, where each participant participated in six trials, and each trial was a unique condition. The trials were grouped into three blocks by Cost condition, with each block containing an Accept trial and a Switch trial. The order of the Cost conditions was counterbalanced across participants, as was the order of Choice conditions within each Cost condition.

The Food scenario used a 2 (Default: Broccoli vs. Chocolate) by 2 (Choice: Accept vs. Switch) within-subject design, where each participant participated in four trials, and each trial

was a unique condition. The four trials were grouped into two blocks by Default condition, with each block containing an Accept trial and a Switch trial.

Procedure and Stimuli

Participants read the Food Scenario and the Cost Scenario in counterbalanced order. In the Cost scenario, in each trial, participants read about a character choosing between two time slots for an event. Each character was automatically assigned a time slot – which serves as a default option in its typical format. The character can switch to the other time slot with some effort, and the level of effort varies by Cost condition: they can switch to the other time slot by sending a one-word text (Low-Cost condition), or by calling (High-Cost condition). We also included a No-Cost condition in which there was no default option – the character simply chose a time lot. This condition served as a baseline to test the assumption that, in the absence of an alternative explanation based on ease, there should be no asymmetric preference inference – choice likely reflects preference.

Across all trials, participants saw three types of events: a dentist appointment, a doctor's appointment, and a class. The character chose between a morning and an afternoon time slot for each event. We used a Latin square design to counterbalance the correspondence between event type and Cost conditions, as well as the order in which Cost conditions were presented (see Supplemental Table 2.4). This resulted in three between-subject presentation orders (A, B, C), which were randomly assigned to participants. Each participant saw the event type in the same order (dentist's appointment, doctor's appointment, class), but the corresponding Cost condition varied according to the Latin square.

Supplemental Table 2.4: Event types across cost conditions in Study 3.Each row in the table is one presentation order that a participant sees, in order from left to right.

	Dentist's Appointment	**Doctor's appointment**	**Class**
A	No Cost	High Cost	Low Cost
B	Low Cost	No Cost	High Cost
C	High Cost	Low Cost	No Cost

In the Food scenario, in each trial, participants read vignettes depicting first-grade child characters choosing between two snacks at school. Participants learned that the child can "choose their snack on an iPad where the two snacks are pictured on the screen". Currently, one of the two snacks was pre-selected as the default. In the Broccoli Default condition, broccoli was pre-selected; in the Chocolate Default condition, chocolate was pre-selected. Participants read that "The kids can either tap a button to confirm the current selection, or they can tap on the chocolate to switch. The kids are free to choose whichever snack they want."

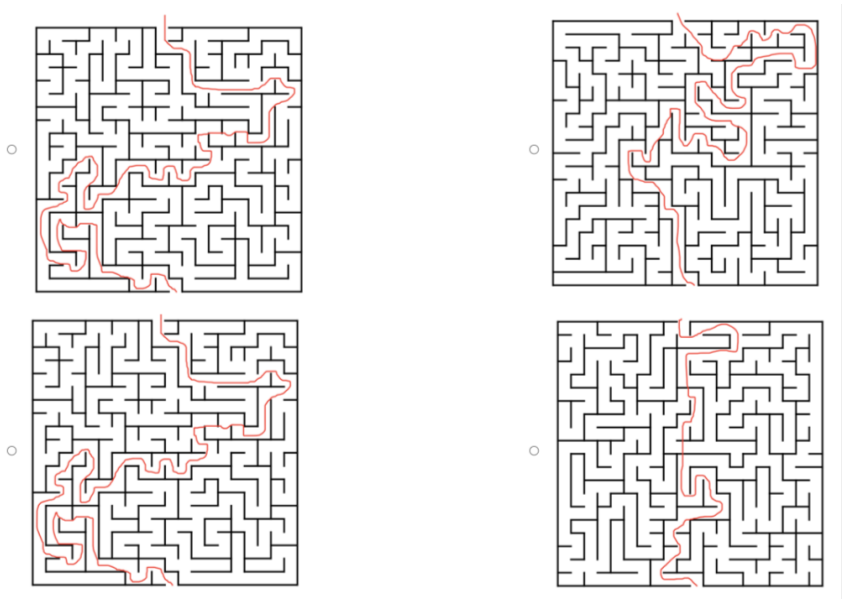
Within each Cost condition in the Cost scenario and within each Default condition in the Food scenario, participants were asked to imagine the character in two situations: one in which they chose the default option (Accept trial), and one in which they chose the alternative option (Switch trial). The order of these two trials was counterbalanced across participants. After each trial, participants inferred the extent to which the character preferred one time slot over the other on a scale of -100 to 100, where -100 indicated that the character 'likes [the default time slot] a lot more' and 100 indicated that the character 'likes [the alternative time slot] a lot more". The slider handle was initially positioned at 0, indicating that the character 'likes both equally', and participants moved the slider to indicate their inferred preference.

Within each scenario, after completing all trials in one condition, participants proceeded to the remaining conditions, which were identical in structure, but described new characters, different event types and different switching costs in the Cost scenario), and different default

snacks in the Food scenario. After completing all trials in one scenario, participants then proceeded to the other scenario.

After completing all trials, participants completed two manipulation checks. One assessed their beliefs about relative effort required to call versus text as a means of rescheduling appointments on a scale of 1 to 5. A score of 1 indicated that "almost all people find making a phone call more effortful", 2 that "most people find making a phone call more effortful", 3 that "the same number of people find either one more effortful", 4 that "most people find sending a text message more effortful", and 5 that "almost all people find sending a text message more effortful". The other assessed their beliefs about first-graders' relative preferences for chocolate and broccoli. Specifically, participants rated on a scale of 1 to 5. A score of 1 indicated that "almost all kids prefer chocolate", 2 that "most kids prefer chocolate", 3 that "the same number of kids prefer either", 4 that "most kids prefer broccoli", and 5 that "almost all kids prefer broccoli".

Finally, to identify and screen out any potential bots, we asked participants to complete a task developed by Roderiguez & Oppenheimer (2024) to distinguish between human and bot responses. We chose this task because GPT4's performance was much worse than humans'. More specifically, participants were asked to identify the only incomplete maze out of the four mazes shown in Supplemental Figure 2.3.



Supplemental Figure 2.3: Mazes for Bot check in Study 2c. The incomplete maze is the bottom-right one These mazes were generated on "mazegenerator.net". The mazes were filled out by hand on a laptop.

Results

Cost scenario

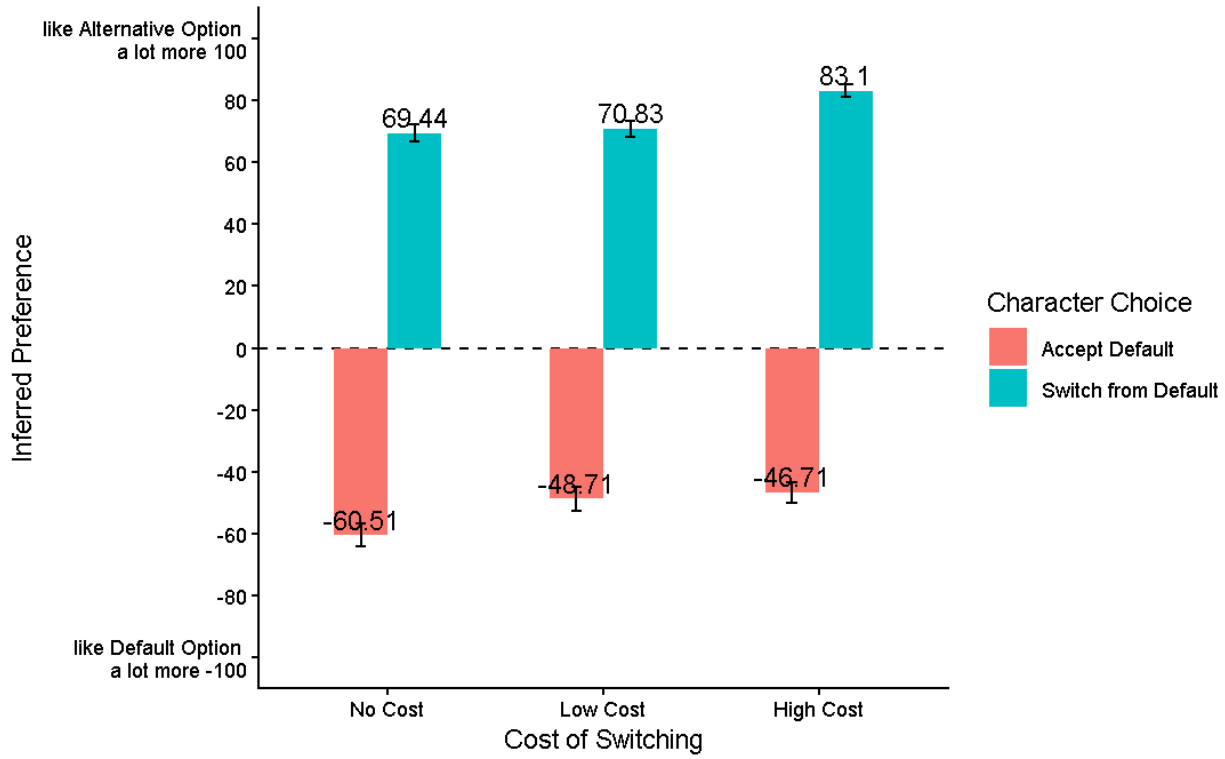
We first examined participants' responses to the manipulation check to assess whether they perceived the intended differences in switching costs. The majority of participants (90%, N = 108) selected either 1 or 2 – indicating that they expect most people to find calling more effortful than texting. Here, we report the results with the restricted sample of participants since results with the full sample were qualitatively similar.

We fit a linear mixed-effects regression predicting the absolute value of the Inferred Preference from Cost, Choice, and their interaction, with random intercepts for participants. Cost was coded as a linearly increasing ordinal variable (No-Cost = 0, Low-Cost = 1, High-Cost = 2). As predicted, there was a significant interaction between Cost and Choice (b = 10.89, SE = 4.07, t (306) = 2.68, p < .001, *95% CI* = [2.94, 18.85]), indicating that the asymmetry in the strength of inferred preference between Accept and Switch trials increased parametrically with switching cost.

To examine the asymmetry within each cost condition without assuming linearity, we conducted follow-up analyses treating Cost as a categorical variable. Comparisons based on estimated marginal means revealed that the difference in inferred preference between Accept and Switch trials increased across cost conditions (see Supplemental Figure 2.4). As expected, there was no significant difference in the No-Cost condition (b = -5.43, p = .060). In contrast, participants inferred significantly stronger preferences in the Switch trials than in the Accept trials in the other two Cost conditions (Low: b = -16.32; High: b = -27.21; both ps < .001; see Supplemental Table 2.5), indicating a parametric increase in asymmetry as switching cost increased.

Supplemental Table 2.5: Difference in inferred preference strength (stay − switch) across cost conditions in Study 2c. Negative values indicate stronger inferred preference in the Switch trials than in the Accept trials.

Cost Condition	**Difference (b)**	**SE**	**95% CI**	**t**	**p**
No Cost	-5.43	2.87	[−11.10, 0.23]	-1.89	.060
Low Cost	-16.32	2.87	[−22.00, −10.66]	-5.68	< .001
High Cost	- 27.21	6.43	[−39.90, −14.57]	-4.23	< .001



Supplemental Figure 2.4: Replicating Study 1 and 3, the asymmetry in inferred preferences increased parametrically with switching cost. Error bars represent ±1 *SE*. The sample includes participants who passed the manipulation check (N = 108).

Because the difference between the Accept and Switch trials in the No-Cost condition almost reached significance ($p = .060$), we ran an exploratory follow-up analysis to probe whether order effects predicted inferred preference above and beyond the Cost × Choice effect. We fit a linear mixed-effects model predicting the absolute value of Inferred Preference from Cost, Choice, their interaction, and Order with which the three conditions were presented, with random intercepts for participants. Low-Cost condition first was the reference level for order. The Cost × Choice interaction remained significant, $b = 10.89$, SE = 4.07, $t(306) = 2.68$, $p = .008$, indicating that the asymmetry between Accept and Switch trials increased as switching cost increased. In addition, there was a significant effect of presentation order: participants who saw the No-Cost condition first made weaker absolute preference inferences than participants who saw the Low-Cost condition first, $b = -13.64$, SE = 5.54, $t(100) = -2.46$, $p = .016$. Participants

who saw the High-Cost condition first did not significantly differ from those who saw the Low-Cost condition first (p = .292).

Thus, although the predicted Cost × Choice effect remained significant when accounting for the order in which the conditions were presented, there was also evidence that participants who saw the No-Cost condition first made generally less extreme preference inferences than those who saw the Low-Cost condition first.

Food scenario

We first examined participants' responses to the manipulation check to assess whether they perceived the intended differences in default strength. The majority of participants (91.6%, N = 110) selected either 1 or 2, indicating that they expected most first-graders to prefer chocolate over broccoli. The results from this restricted sample and the full samples differed qualitatively. Therefore, we report the results from both samples below:

Restricted sample

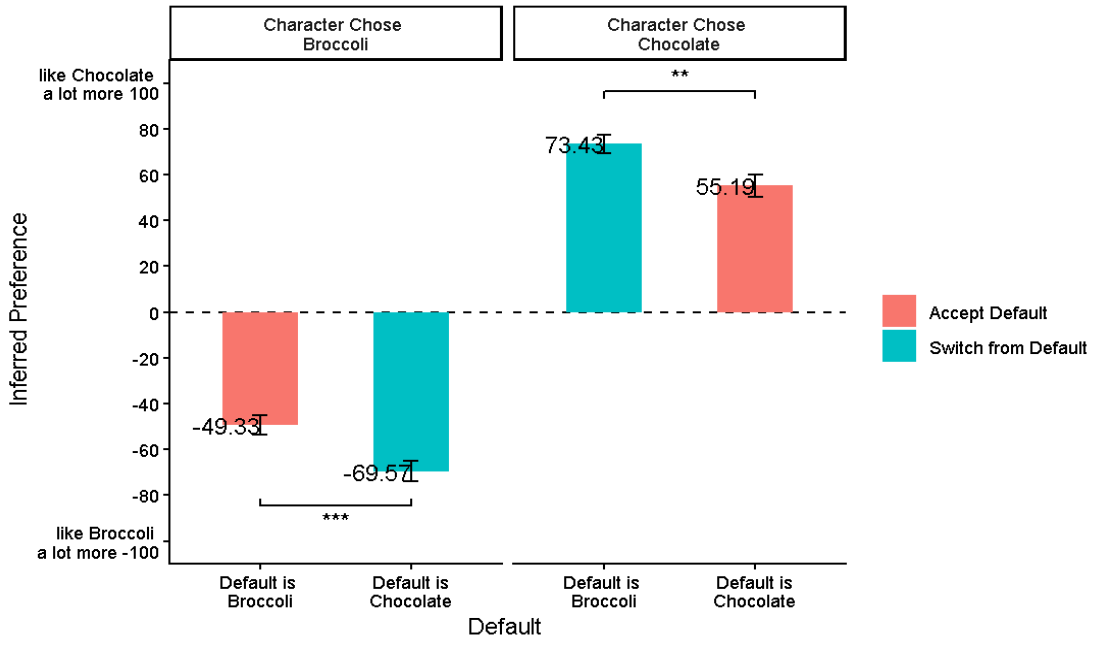
We fit a linear mixed-effects regression predicting the absolute value of Inferred Preference from Default, Choice (Accept or Switch), and their interaction, with random intercepts for participants. There was a significant interaction between Default and Choice (b = −13.06, SE = 4.01, $t(306) = -3.26$, p = .001, *95% CI* = [−20.90, −5.22]), indicating that the strength of asymmetry in inferred preference between Accept and Switch trials differed based on which snack was the default.

To examine the magnitude of asymmetry for each chosen snack, we conducted paired-samples t-tests on inferred preferences across the two trials when each snack was chosen. As in Study 2a, participants inferred weaker preferences in Accept trials than in the Switch trials when broccoli was chosen, and this asymmetry diminished when chocolate was chosen (see Supplemental Figure 2.5).

When broccoli was chosen, participants inferred significantly weaker preferences in the Accept trials ($M = -49.33$, $SD = 41.97$) than in the Switch trials ($M = -69.57$, $SD = 45.42$), $t(102) = 3.94$, $p < .001$, $d = .39$, *95% CI* = [-0.19, 0.71], indicating strong asymmetry.

By contrast, this asymmetry was diminished when chocolate was chosen. Although participants again inferred significantly weaker preferences in the Accept trials ($M = 55.19$, $SD = 49.35$) than in the Switch trials ($M = 73.43$, $SD = 41.01$), this difference was smaller in magnitude, $t(102) = -3.03$, $p = .003$, $d = -0.30$, *95% CI* = [-0.51, -0.12], indicating weaker asymmetry.

An exploratory analysis confirmed that the asymmetry diminished when chocolate was chosen. We fit a linear mixed effects model predicting the absolute value of Inferred Preference from Choice, *Chosen Snack* (broccoli vs. chocolate), and their interaction with random intercepts for participants. We found a significant interaction between Choice and Chosen Snack, $b = -7.99$, SE = 4.01, $t(306) = -1.99$, $p = .047$, *95% CI* = [-15.83, -0.15], indicating that the magnitude of asymmetry differed depending on which snack was chosen.



Supplemental Figure 2.5: As in Study 2a, participants made asymmetric inferences when broccoli was chosen, but this asymmetry was larger when chocolate was chosen. Error bars represent ±1 *SE*. The sample includes participants who passed the manipulation check (N = 108).

Full sample

We fit a linear mixed-effects regression predicting the absolute value of Inferred Preference from Default, Choice (Accept or Switch), and their interaction, with random intercepts for participants. There was a significant interaction between Default and Choice, $b = -14.30$, SE = 3.88, $t(357) = -3.69$, $p < .001$, indicating that the strength of asymmetry in inferred preference between Accept and Switch trials differed based on which snack was the default.

To examine the magnitude of asymmetry for each chosen snack, we conducted paired-samples *t*-tests on inferred preferences across the two trials when each snack was chosen. When broccoli was chosen, participants inferred significantly weaker preferences in the Accept trials ($M = -50.56$, SD = 41.96) than in the Switch trials ($M = -67.30$, SD = 47.49), $t(119) = 3.46$, $p < .001$, $d = 0.32$, *95% CI* = [0.11, 0.59], indicating significant asymmetry.

When chocolate was chosen, participants also inferred significantly weaker preferences in the Accept trials (M = 54.80, SD = 51.15) than in the Switch trials (M = 73.28, SD = 42.24), $t(119) = -3.19$, $p = .002$, $d = -0.29$, *95% CI* = [−0.51, −0.10], again indicating significant asymmetry. However, contrary to our prediction, the difference in inferred preference ratings did not diminish when chocolate was chosen. In fact, the difference was numerically larger when chocolate was chosen (mean difference = 18.48) than when broccoli was chosen (mean difference = 16.74).

An exploratory analysis further tested whether the magnitude of asymmetry differed by chosen snack. We fit a linear mixed-effects model predicting the absolute value of Inferred Preference from Choice, Chosen Snack (broccoli vs. chocolate), and their interaction, with random intercepts for participants. The Choice × Chosen Snack interaction was not significant, $b = -6.48$, SE = 3.88, $t(357) = -1.67$, $p = .096$, *95% CI* = [−14.08, 1.11]. Thus, in the full sample, we did not find clear evidence that asymmetry diminished when chocolate was chosen.

Discussion

In Study 2c, participants made asymmetric preference inferences when defaults provided alternative explanations for choice, but the evidence was clearer in the Cost scenario than in the Food scenario. In the Cost scenario, where ease served as the alternative explanation for accepting the default, participants made increasingly asymmetric preference inferences between accepting and switching as the cost of switching increased. Replicating findings in Studies 1 and 3, participants inferred weaker preferences when characters accepted the default than when they switched away from it, and this difference increased parametrically with switching cost.

In the Food scenario, participants who passed the assumption check and the full sample showed opposite patterns of results. Among participants who passed the assumption check, the

results replicated Study 2: when the chosen snack was broccoli, participants inferred weaker preferences from accepting the default than from switching away from it, but this asymmetry diminished when chocolate was chosen. However, this pattern reversed in the full sample. When all participants were included, the asymmetry was slightly larger when chocolate was chosen than when broccoli was chosen. Thus, although the restricted sample supported our prediction, the full-sample results suggested that when defaults provide endorsement-based alternative explanations, participants are less homogenous in their prior beliefs about the DMs' existing preferences.

One possibility is the stimuli in Study 2c, compared to those in Study 2, did not highlight the fact that the DMs were young children – who, therefore, likely prefers chocolate over broccoli. This is because Study 2 included illustrations of children making the choices, whereas Study 2c did not. The absence of illustrations may have made it less salient that the characters were children, which could have weakened the shared assumption that children generally prefer chocolate to broccoli. Without these illustrations, it may have been difficult for participants to keep in mind that the characters were children. This possibility is consistent with the finding that participants who failed the assumption check showed the opposite pattern of results.

We also observed two additional methodological issues in the Cost scenario. First, participants' responses were somewhat skewed toward the right side of the scale, raising the possibility of a dominant-hand motor-spatial bias. Second, participants' response patterns suggested an order effect. To address these issues, we collected a third pilot sample ($N = 25$) in which we made three changes: we added illustrations of children making the choice, counterbalanced the endpoints of the dependent measure, and added instructions giving participants an overview of the Cost trials. The pilot suggested that the additional instructions

were unnecessary and somewhat confusing. Therefore, in Studies 3 and 4, we retained only the first two changes: adding illustrations and counterbalancing scale endpoints.

Finally, because the Food and Cost scenarios appeared to introduce different sources of variability, we separated them into two preregistered studies. Study 3 tested the Cost scenario with counterbalanced scale endpoints, and Study 4 tested the Food scenario with illustrations and counterbalanced scale endpoints.

Chapter 3 WATCH ME BE SELF-CONTROLLED: HOW SOCIAL EVALUATION SHAPES REPORTED SELF-CONTROL AND COOPERATION

Introduction

People care about whether others are self-controlled, and whether they appear self-controlled to others, even in mundane and private matters (e.g., choosing salad over fries). While self-control in interpersonal settings can be consequential for others, self-control over harmless private indulgences does not seem to have much social relevance. Why, then, do people care about others' self-control over private indulgences?

A functional reason is that self-control signals cooperativeness (Fitouchi et al., 2022). Self-control, defined through various terms such as "delay of gratification, effortful control, willpower, executive control, time preference, self-discipline, self-regulation, and ego strength," encapsulates the concept of effortful self-regulation (Duckworth, 2011, p.2639). This ability to self-regulate is essential to cooperation, which requires individuals to prioritize prosocial benefits over selfish temptations – this is why some researchers define cooperation as interpersonal self-control (Dewitte & Cremer, 2001; Kocher et al., 2017; Rachlin, 2016). It follows that self-control over private indulgences may signal one's capability for self-regulation in interpersonal settings, and therefore signal cooperativeness. Indeed, individuals who are more self-controlled are often perceived as more cooperative (Buyukcan-Tetik et al., 2015; Buyukcan-Tetik & Pronk, 2023; Gai & Bhattacharjee, 2022; Gomillion et al., 2014; Koval et al., 2015; Peetz & Kammrath, 2013; Pronk et al., 20110117; Righetti & Finkenauer, 2011). Furthermore, future-oriented time preference–a signature of self-control– has been shown to explain variability in trust and cooperation with non-kin (Lie-Panis & André, 2022).

Does the association between self-control and cooperation suggest a causal relationship between the two? The answer to this question is unclear for two main reasons: 1) self-control and

cooperation can both be driven by an epiphenomenal variable: concern for social evaluation, and 2) while social evaluation causally affects cooperation, few studies examined the causal effect of social evaluation on self-control. Social evaluation influences people's reputation and outcomes in cooperation settings, and can therefore motivate people to behave in ways that enhance their reputation for cooperativeness, such as increasing cooperation behavior or expressing higher self-control. Therefore, social evaluation might causally impact both expressed self-control and cooperation behavior. If so, this covariation may explain away the potential causal relationship between self-control and cooperation.

Extensive research has documented the impact of social evaluation on cooperation (Bateson et al., 2006; Bradley et al., 2018; Cain et al., 2014; Franzen & Pointner, 2012; McAuliffe et al., 2018; Thielmann et al., 2016; Winking & Mizer, 2013). Moreover, many of these findings suggest that social evaluation has a robust causal effect on increasing cooperation (Bateson et al., 2006; Franzen & Pointner, 2012; McAuliffe et al., 2018; Thielmann et al., 2016, Winking & Mizer, 2013). In particular, when social evaluation is experimentally manipulated, people behave less cooperatively towards others when evaluation is low, both in economic games (Bradley et al., 2018; Cain et al., 2014; Franzen & Pointner, 2012; McAuliffe et al., 2018; Thielmann et al., 2016) and in charitable donations (Bateson et al., 2006; McAuliffe et al., 2018; Winking & Mizer, 2013).

However, it is unclear whether social evaluation also causally affects self-control, partly because few experiments have experimentally manipulated social evaluation on expressed self-control, especially when the evaluation is consequential for future cooperation. In contrast to the extensive empirical evidence showing that social evaluation affects cooperation, only a few studies have documented self-control's association with social evaluation, most of which has

focused on correlational associations. For instance, individuals with higher trait self-control were shown to score higher on impression management scales (Stavrova & Kokkoris, 2019). In addition, religious primes, which are thought to activate self-control (Marcus & McCullough, 2021), were found to predict self-control behaviors in delayed gratification and impossible puzzle tasks (Rounding et al., 2012), but these findings do not replicate after controlling for social desirability as a predictor (Harrison & McKay, 2013). At least one study has experimentally manipulated social evaluation to test its effect on self-control (Ma et al., 2023). However, this study was conducted with children, whose concern for social evaluation seemed to be confounded with motivation to gain social rewards. More specifically, Ma et al. (2023) implemented a modified Marshmallow task, where children learned that their self-control behavior (i.e., waiting for a larger reward) was either observed by the experimenter or not. Children waited longer when their behavior was observed by the experimenter, and the authors attributed these findings to children's desire to seek social rewards: the fact that self-control (i.e., waiting) is associated with a larger reward signals the experimenter's approval of self-control, and this approval confers social rewards to children. In other words, these findings likely reflect children's motivation to gain social rewards, rather than a causal relationship between social evaluation and self-control *per se*.

In sum, there is empirical evidence that self-control is correlated with cooperation and that social evaluation causally increases cooperation. However, in order to clarify the causal relationship between self-control and cooperation, it is crucial to empirically assess the causal relationship between social evaluation and self-control, and its downstream consequences for cooperation.

The present experiments

We reason that social evaluation may cause people to increase their expressed self-control in order to appear as good cooperation partners. Our logic is as follows: people want to seek out high-quality cooperation partners (Barclay, 2013); to do so, they rely on cues that signal whether others are high- (or low-) quality partners, such as direct observation of their social interactions, or third-party information derived from gossip (Nowak, 2006). Self-control can serve as one such cue for high quality partners because it signals one's ability to prioritize prosocial benefits over selfish ones in cooperative settings. On this view, when people believe that they are being evaluated by others, then they ought to display more self-control in order to appear like a higher-quality cooperation partner. We broadly predict, then, that social evaluation ought to have a causal effect on people's expression of self-control.

A second related question is: If people express more self-control under social evaluation in order to appear more cooperative, do they do so only when social evaluation is directly consequential for cooperation, or whenever they are socially evaluated? We make two further competing predictions about *when* concern for social evaluation might influence the desire to appear self-controlled:

On one view, people may try to appear self-controlled only when they are observed by potential cooperation partners because in these situations their reputation for cooperativeness is at risk (Bateson et al, 2006; Bradley et al, 2018; McAuliffe et al., 2018). We term this the *reputation-maintenance hypothesis*. Based on the reputation-maintenance hypothesis, we make the following prediction:

Prediction 1.1: Social evaluation will cause people to inflate their self-control, but only when their reputation is at risk.

On another view, people may try to appear self-controlled anytime they are observed, regardless of whether the observers will be their future cooperation partners, because assuming one's reputation isn't at risk when it really is can cause massive reputational damage (McAuliffe et al., 2018; Yamagishi et al., 2008). We term this the *misfiring hypothesis*. Based on the misfiring hypothesis, we make the following competing prediction:

Prediction 1.2: Social evaluation will cause people to inflate their self-control, regardless of reputational risk.

If social evaluation indeed increases expressed self-control, then the correlation between self-control and cooperation may or may not hold after controlling for social evaluation. We make two further hypotheses about how social evaluation on self-control might affect actual cooperation behavior:

From one standpoint, people may inflate their self-control so that they *appear* like a high-value cooperation partner without actually being one. In this case, self-control will be a poor predictor of cooperation when social evaluation is high, that is, when people artificially increase their self-control to appear like a good cooperative partner. By contrast, self-control will positively predict cooperation when social evaluation is low, that is, when the relationship between self-control and cooperation is not obscured by the desire to appear cooperative. We term this hypothesis the *cheap-talk hypothesis.* Based on the cheap-talk hypothesis*,* we make the following prediction:

Prediction 2.1: Self-control will predict cooperation to a lesser degree when social evaluation for self-control is high versus low.

From another standpoint, people may be motivated to remain consistent in their expression of self-control and cooperation behavior, regardless of whether or not people inflate

their self-control to appear more cooperative. In this case, self-control can be an honest signal of one's cooperation behavior even when the former was driven by concern for social evaluation. We term this hypothesis the *honest-signal hypothesis.* Based on the honest-signal hypothesis, we make the following prediction:

Prediction 2.2: Self-control will be equally predictive of cooperation regardless of the degree of social evaluation for self-control.

To address our hypotheses, we conducted two experiments to test four predictions about the causal relationship between social evaluation, self-control, and cooperation. In both experiments, participants learned that they would complete a self-control questionnaire before playing a one-shot Trust Game with another randomly selected participant.

We used a self-control questionnaire with face-valid items – rather than behavioral measures of self-control – in order to minimize the possibility that different participants interpret self-control differently, and make sure that participants are aware that their responses can signal how self-controlled they appear to others. For instance, the questionnaire item "I am good at resisting temptation" presumably introduces less room for alternative interpretations of self-control than a behavioral measure such as the amount of time one waits for a larger reward or the amount of time one spends solving an impossible puzzle.

The Trust Game provides an ideal framework to explore the impact of social evaluation and reputational risk on reported self-control. More specifically, it creates a situation in which participants are incentivized to build a good reputation in order to attract a cooperation partner who will send them money; however, after they receive the money, participants can make a unilateral financial allocation – that is, they can behave selfishly without fear of consequence.

Before starting the self-control questionnaire, participants either learned that 1) their self-control scores would be shared with their Trust Game partner (social evaluation by cooperation

partner), or 2) their self-control score would be shared with another randomly selected participant who is not their game partner (social evaluation by an irrelevant stranger), or 3) no additional information (no evaluation). If reported self-control is driven by the motivation to appear cooperative only when reputation is at risk, as predicted by the reputational-maintenance hypothesis, then self-control scores will be higher under social evaluation by a cooperation partner compared to under social evaluation by an irrelevant stranger or under no social evaluation. However, if reported self-control is driven by any social evaluation, regardless of reputational risk, as predicted by the misfiring hypothesis, then the self-control scores will be similar under social evaluation by either a cooperation partner or an irrelevant stranger, and these scores will be higher than the scores under no social evaluation.

Participants' cooperation was measured by the proportion of money participants returned in a one-shot Trust Game after they reported their self-control. If reported self-control is driven by the desire to appear as a good cooperation partner, but does not causually predict actual cooperation behavior, as predicted by the cheap-talk hypothesis, then the correlation between reported self-control and cooperation will be weaker under social evaluation than under no social evaluation. Alternatively, if reported self-control causally predicts actual cooperation behavior regardless of social evaluation, as predicted by the honest-signal hypothesis, then the correlation between self-control and cooperation will be equal across the three social evaluation conditions.

Transparency and Openness

Experiment materials, data, and analysis code for all studies can be found at https://osf.io/9fktb/overview?view_only=a3b22030ae5542ed95b481d9d2ef0321. The pre-registration for Experiment 1 can be viewed online at https://osf.io/kpfte/overview?view_only=4f0cf5e25a884456a7d213efa609edf7

The pre-registration for Experiment 2 can be viewed online at https://osf.io/vdp2y/overview?view_only=1ce6c0bba69647aebd1f028cafe44037

Data were analyzed using R version 4.3.0 (R Core Team, 2023), and plots were made using the package ggplot2, version 3.5.1. All materials and procedures were approved by the institutional review board of the University of California, San Diego. We report all measures, manipulations, and exclusions in these experiments (additional measures not included in analyses are described in the Supplemental Materials).

Experiment 1

Methods

Participants

Participants were N = 367 UC San Diego undergraduate students (M_{Age} = 20.95, SD_{Age} = 2.86; 60 male, 294 female, 12 non-binary, 1 participant failed to provide sex information). We estimated that we would need a minimum of N = 300 in order to detect a medium effect size of d = 0.4 with 80% power at a significance level of $\alpha = .05$ for any two-group comparison. We aimed to collect data from at least N = 300. We oversampled to ensure a sufficient sample size in case of unforeseen reasons participants might need to be excluded, although no participants needed to be removed.

Participants completed the experiment remotely on Qualtrics via their computers. They received course credit for participating, and were told they could earn up to $6, depending on their performance in the task. They received the actual payoff they earned in the Trust Game. Data collection took place from January to February 2025.

Procedure

Partner pairing and chat. After providing informed consent, participants were first randomly paired with another participant who was completing the study at the same time, and

engaged in a 6-minute guided chat using questions selected from the Relationship Closeness Induction Task (RCIT; Sedikide et al., 1999). Then, participants learned that they and their chat partner would be paired with two other participants named Jordan and Desmond in a group of four, and that they would play an economic game with a randomly selected member of their group. In reality, all remaining interaction between the participant and their partner was a pre-programmed computer script. Similarly, Jordan and Desmond were not real participants in the study, and we named them as such so that they would be gender-neutral.

Trust Game Instructions. Participants were then told they would play a game with Desmond, who was (ostensibly) randomly selected from their group. In this game (the Trust Game), one player is assigned the role of the "sender", and is given the option to send money to another player assigned the role of the "receiver". The receiver then has the option to return money to the sender. All participants were told that they were randomly assigned as the receiver in the game. Consequently, participants were in a position to unilaterally control the monetary fate of Desmond, and believed that Desmond was always in a position to judge whether or not the participant would reciprocate.

Social Evaluation Manipulation. Before playing the Trust Game, participants were instructed to complete a questionnaire. At this point, participants were randomly assigned to one of three conditions: In the Recipient-partner condition, they were told that their questionnaire response was randomly selected to be shared with Desmond, their game partner; in the Recipient-stranger condition, they were told that their questionnaire response was randomly selected be shared with, Jordan, another participant from their group; and in the No-recipient condition, they were not told that their questionnaire response would be shared with anyone. Thus, participants in the Recipient-partner and Recipient-stranger conditions were in a position

to signal their self-control to their partner before the Trust Game began by broadcasting their self-reported self-control.

Comprehension Checks. Before starting the questionnaire, participants completed some comprehension checks. All participants indicated their role in the game, the possible actions they could take (i.e., the receiver, receive money from the sender and return money back), their partner's name (Desmond), and when they will play the game (after the questionnaire). Participants in the Recipient-partner and the Recipient-stranger conditions also indicated with whom their questionnaire responses would be shared (Desmon and Jordan, respectively).

Reported Self-control. Participants completed Tangney et al.'s (2004) Brief Self-Control Scale, which included 13 items rated on a scale of 1 to 5.

Trust Game. Upon completing the questionnaire, participants then played the Trust Game with Desmond. Participants were told that both they and Desmond received $2 windfall before the game began. We also fixed the amount of money that Desmond sent to $0.50, which was doubled, so that participants always received $1. Participants then indicated how much of the $1 they wanted to return to Desmond on a 5-point scale ranging from $0.00 to $1.00 with $0.25 intervals.

Manipulation check. After the Trust Game, participants indicated how likely they thought they would interact with their game partners after the experiment ended on a 5-point scale ranging from extremely unlikely (1) to extremely likely (5).

Debriefing and dismissal. Finally, participants reported their demographic information, and completed a funnel questionnaire where they answered open-ended questions about the experiment to determine whether they were aware of the experimental hypothesis, and if they were suspicious about the veracity of the experiment. Participants also completed a few other

measures that were outside the scope of this article. Please see the Supplemental Materials for information about these other measures. We thought participants might be skeptical about aspects of the experimental procedure (e.g., whether the other "participants" in their group were real, or if they would actually earn money in the Trust Game), which could influence their responses.

Results

Did participants understand their task?

We first examined participants' responses in the comprehension checks to see if they understood the task. The majority of participants (79%, $N = 290$) correctly answered all comprehension checks. Here, we report the results with this restricted sample of participants. Results with the full sample differed slightly, and are available in Supplemental Materials.

Did participants think they would see their Trust Game partner after the experiment?

To ensure that participants' cooperative behavior in the Trust Game was not inflated by the perceived reputational risk of future interaction, we compared participants' beliefs that they would encounter their partner in the Trust Game sometime in the future ($M = 1.69$, $SD = 1.00$). Pairwise comparisons across conditions showed no significant effect of condition on participants' responses ($ps > .786$). As expected, regardless of social evaluation, participants didn't think they would re-encounter their partner after the game.

Prediction 1.1 vs. 1.2: Does social evaluation positively predict reported self-control scores only when reputation is at risk (1.1), or under any social evaluation (1.2)?

Only when reputation is at risk (Prediction 1.1). For each participant, we first calculated a composite of self-control scores ($M = 3.16$, $SD = 0.60$; *McDonald's* $\Omega = .86$) by taking the average score of the items in the Brief Self-control Scale (Tangney et al., 2014). Figure 3.1 shows participants' average self-control scores in each condition.

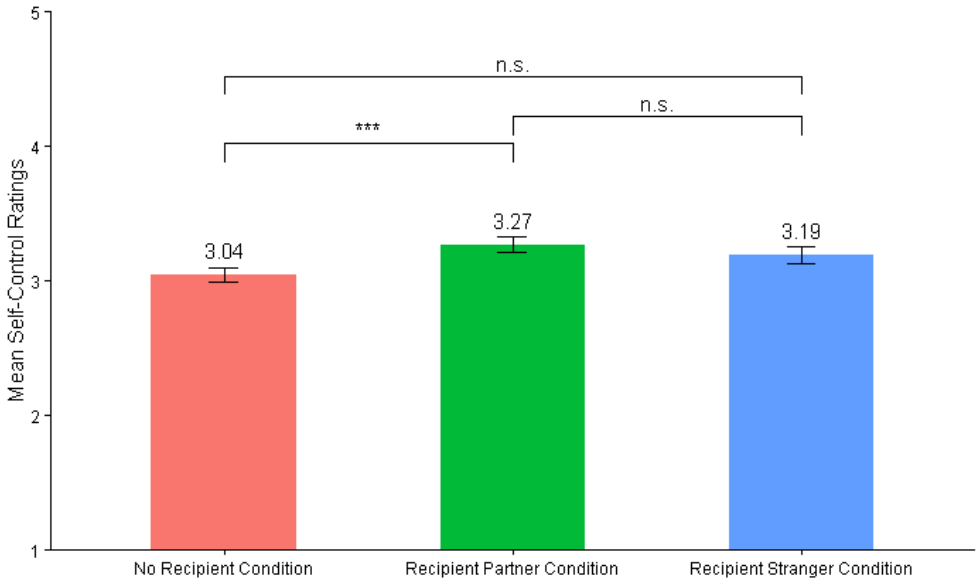


Figure 3.1: Self-control scores and standard errors across the three conditions in Experiment 1. Error bars represent ±1 *SE*. The sample includes participants who passed all attention checks (*N* = 290).

We regressed self-control scores on: (1) A Helmert-coded predictor variable representing the effects of the Recipient-partner vs. Recipient-stranger and No-recipient conditions, and (2) A Helmert-coded predictor variable representing the effects of the Recipient-stranger vs. the No-recipient conditions. Table 1 lists the coding scheme that we used to test our prediction.

Table 3.1: Coding scheme for the model corresponding to Prediction 1.

Condition	**Helmert Code 1**	**Helmert Code 2**
Recipient-partner	2/3	0
Recipient-stranger	-1/3	1/2
No-recipient	-1/3	-1/2

In this analysis, the coefficient for the first Helmert-coded predictor variable corresponds to the difference in self-control scores between the Recipient-partner condition and the combined mean of the Recipient-stranger and No-recipient conditions. The coefficient for the second

Helmert-coded predictor variable corresponds to the difference in self-control scores for the Recipient-stranger and No-recipient conditions.

In support of Prediction 1.1, there was a main effect of reputational risk. More specifically, the coefficient for Helmert Code 1 was positive and statistically significant (b = 0.15, SE = 0.07, *95% CI* = [0.02, 0.29], β = .13, p = .027), and the coefficient for Helmert Code 2 was non-significant (b = 0.15, SE = 0.08, *95% CI* = [-0.01, 0.31], β = .106, p = .071).

These findings support the reputation-maintenance hypothesis, such that participants in the Recipient-partner condition increased their self-control prior to the Trust Game, compared to participants in the Recipient-stranger and No-recipient conditions.

Prediction 2.1 vs. 2.2: Is there an interaction between social evaluation and the correlation between reported self-control and cooperation (2.1) or not (2.2)?

Yes (Prediction 2.1). We first examined the correlations between self-control scores and the amount of money returned (M = 0.71, SD = 0.25) in each condition. Self-control scores were significantly and negatively correlated with the amount of money returned in the Recipient-partner condition, r (99) = -0.24, p = .016. The correlation was non-significant for both the Recipient-stranger (p = .754) and the No-recipient conditions (p = .067). Figure 3.2 shows correlations between the self-control scores and the amount of money returned in each condition.

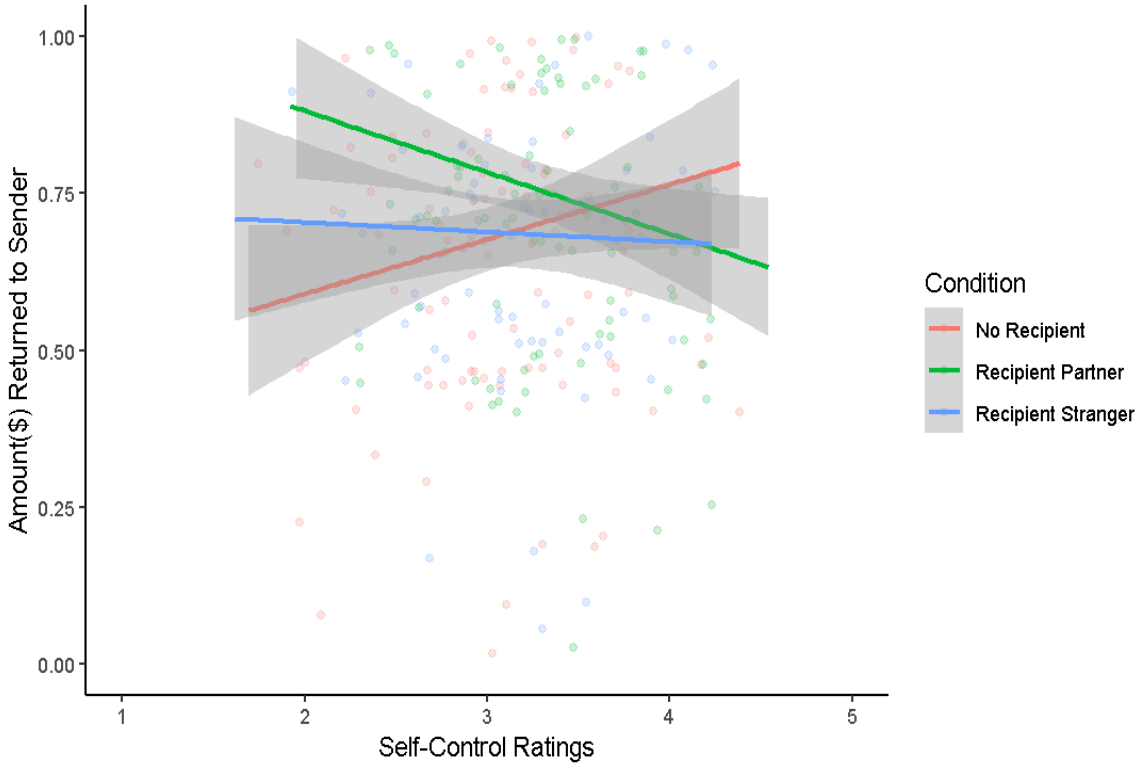


Figure 3.2: Interaction between the conditions and the correlation between self-control and cooperation in Experiment 1. Lines represent linear model fits. Shaded bands represent 95% CI around the fitted regression lines. The sample includes participants who passed all attention checks (N = 290).

To statistically test whether the correlation between self-control and money returned differed across the three conditions, we regressed the amount of money participants returned in the Trust Game on five predictors using the same Helmert coding scheme shown in Table 1: (1) Helmert Code 1, (2) Helmert Code 2, (3) self-control scores; (4) the interaction between Helmert Code 1 and self-control scores, and (5) the interaction between Helmert Code 2 and self-control scores.

In support of Prediction 2.1, the interaction between the coefficient for Helmert Code 1 and self-control scores was negative and significant (b = -0.13, SE = 0.06, *95% CI* = [-0.24, -0.03], β = -.84, p = .015). However, the interaction between coefficient for Helmert Code 2 and self-control was non-significant (b = -0.10, SE = 0.07, *95% CI* = [-0.23, 0.03], β = -.52, p =

.119). The coefficient for Helmert Code 1 was positive and significant (b = 0.50, SE = 0.18, *95% CI* = [0.15, 0.86], t = 2.81, β = .96, p = .005). The coefficients for Self-control and for Helmert Code 2 were non-significant (ps > .127).

These findings partially support the cheap-talk hypothesis, such that participants with higher self-control scores in the Recipient-partner condition were less likely to send money to their partner, compared to participants in the Recipient-stranger and No-recipient conditions. We elaborate on this interpretation below.

Discussion

In Experiment 1, we found evidence that reported self-control is causally influenced by social evaluation, but only when one's reputation is at risk. In support of the reputation-maintenance hypothesis, we found that participants reported higher self-control after learning that their self-control scores would be shared with their partner in the Trust Game, compared to those who thought their scores would be shared with a stranger, or no one at all. In other words, participants inflated their self-control when they were evaluated by future cooperative partners, using self-control as a signal of cooperativeness.

When examining the relationship between self-control and cooperation, we found partial support for the cheap talk hypothesis such that self-control has a weaker association with cooperation for participants in the Recipient-partner condition versus the other two conditions, but we found no difference in the strength of the association between self-control and cooperation when comparing the Recipient-stranger and No-recipient conditions. These results suggest that participants inflated their self-control when their reputation was at risk, but did not behave more cooperatively, and are therefore consistent with the logic of the cheap-talk hypothesis, but they did not match the exact contrast pattern we preregistered.

This difference reflects a limitation in how we translated the cheap-talk hypothesis into preregistered contrasts. For the effect of social evaluation on reported self-control, we preregistered two distinct sub-predictions: the sub-prediction (1.1) that people inflate their reported self-control when their reputation is at risk (reputation-maintenance), and the sub-prediction (1.2) that people inflate their reported self-control under any social evaluation (misfiring hypothesis). These sub-predictions mapped cleanly onto our Helmert contrasts: the reputation-maintenance hypothesis required only a significant contrast comparing the Recipient-partner condition to the other two conditions, whereas the misfiring hypothesis required both Helmert contrasts to be significant.

However, for the moderation of the association between self-control and cooperation, our preregistered criterion for the cheap-talk hypothesis required both Helmert contrasts to be significant. In hindsight, this criterion was stricter than the conceptually required. The central prediction of the cheap-talk hypothesis was that when evaluated by a future cooperative partner, people may appear more self-controlled without acting more cooperatively. Therefore, this prediction could be supported by a significant Recipient-partner contrast, even if the Recipient-stranger and No-recipient conditions did not differ. With this in mind, we interpret these results as in partial support for the cheap-talk hypothesis.

The observed interaction was also driven by a different pattern of correlations than we predicted. We predicted that self-control would positively predict cooperation when social evaluation is low, and that this correlation would diminish why social evaluation is high. However, we found that self-control did not predict cooperation when social evaluation was low – although it was in the predicted direction, and it negatively predicted cooperation when social evaluation was high.

To better understand this pattern of results, we conducted a follow-up experiment using a new sample of Prolific workers. We directly replicated the method from Experiment 1, with an additional measure of individual differences as part of another project (see Supplemental Materials for more information about this measure).

Experiment 2

In Experiment 2, we attempted to replicate our results in Experiment 1 in a sample of Prolific workers.

Methods

Participants

As preregistered, $N = 301$ adults from the U.S. participated, recruited from www.prolific.com. We removed two participants who had duplicate Prolific IDs, and two participants who were not matched with another participant for the live chat. Overall, data from $N = 297$ participants were included in the analyses ($M_{Age} = 37.68$, $SD_{Age} = 12.37$; 137 male, 156 female, 4 non-binary). As in Experiment 1, we selected our sample size based on our preregistered goal of sampling $N = 300$.

Participants completed the experiment remotely on Qualtrics via their computers. They received $12/hr for their participation, and were told they could earn up to $6, depending on their performance in the task. They received their actual payoff earned in the Trust Game. Data collection took place on May 28th, 2025.

Procedure

The procedure for Experiment 2 was identical to that of Experiment 1.

Results

Did participants understand their task?

We first examined participants' responses in the comprehension checks to see if participants understood the task. As in Experiment 1, the majority of participants (71%, $N = 211$)

correctly answered all comprehension checks, although this percentage was lower than in Experiment 1. Again, we report the results with this restricted sample of participants. Results with the full sample show similar patterns, and are available in Supplemental Materials.

Did participants think they would see their Trust Game partner after the experiment?

We conducted pairwise comparisons for participants' perceptions of the likelihood that they would interact with their Trust Game partner after the experiment ended (M = 2.63, SD = 1.42) across the three social evaluation conditions. All pairwise comparisons were non-significant (ps > .667), indicating that the social evaluation manipulations did not affect participants' beliefs about future interaction with their game partner.

Prediction 1.1 vs. 1.2: Does social evaluation positively predict reported self-control scores only when reputation is at risk (1.1), or under any social evaluation (1.2)?

Neither. Social evaluation did not predict reported self-control (M = 3.74, SD = 0.59; *McDonald's* Ω = .86). For each participant, we first calculated a composite of self-control scores by taking the average score of the items in the Brief Self-control Scale (Tangney et al., 2014). Figure 3.3 shows participants' average self-control scores in each condition.

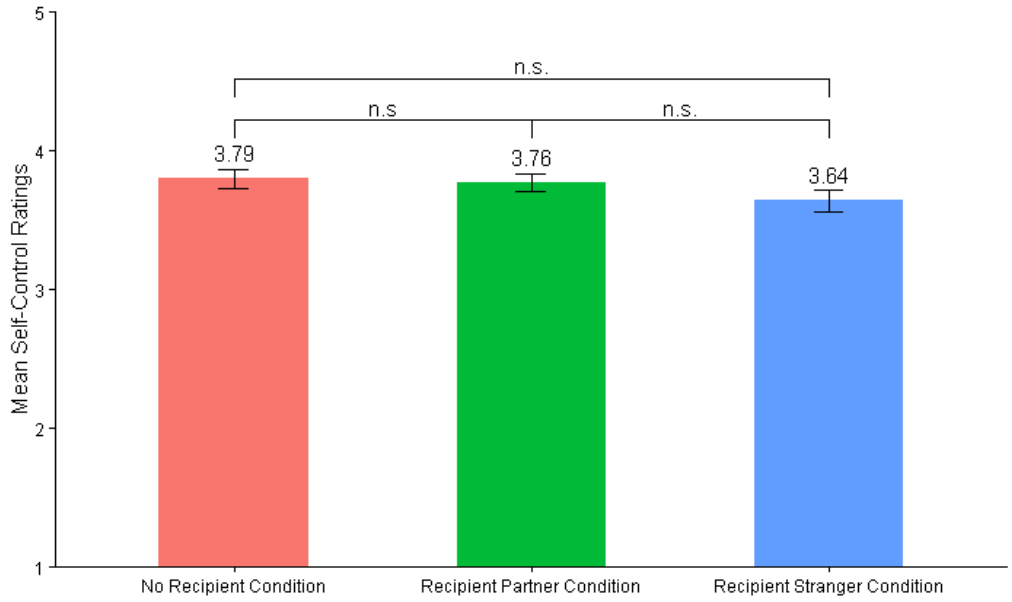


Figure 3.3: Self-control scores across all three conditions in Experiment 2. Error bars represent ±1 *SE*. The sample includes participants who passed all attention checks ($N = 211$).

Using the same Helmert coding scheme as above, we regressed self-control scores on: (1) A Helmert-coded predictor variable representing the effects of the Recipient-partner vs. Recipient-stranger and No-recipient conditions, and (2) A Helmert-coded predictor variable representing the effects of the Recipient-stranger vs. the No-recipient conditions.

Contrary to our predictions, there was no main effect of social evaluation on reported self-control. More specifically, the coefficient for Helmert Code 1 was non-significant ($b = 0.04$, $SE = 0.08$, *95% CI* = [-0.11, 0.21], $\beta = .04$, $p = .557$), and the coefficient for Helmert Code 2 was also non-significant ($b = -0.16$, $SE = 0.10$, *95% CI* = [-0.36, 0.47], $\beta = -.11$, $p = .130$).

Unlike in Experiment 1, these findings did not support our hypothesis that self-control is driven by social evaluation.

Prediction 2.1 vs. 2.2: Is there an interaction between social evaluation and the correlation between reported self-control and cooperation (2.1) or not (2.2)?

No (Prediction 2.2). We first examined the correlations between self-control scores and the amount of money returned ($M = 0.55$, $SD = 0.27$) in each condition. Self-control scores were

not significantly correlated with the amount of money returned in any condition (*p*s > .166). Figure 3.4 shows correlations between the self-control scores and the amount of money returned in each condition.

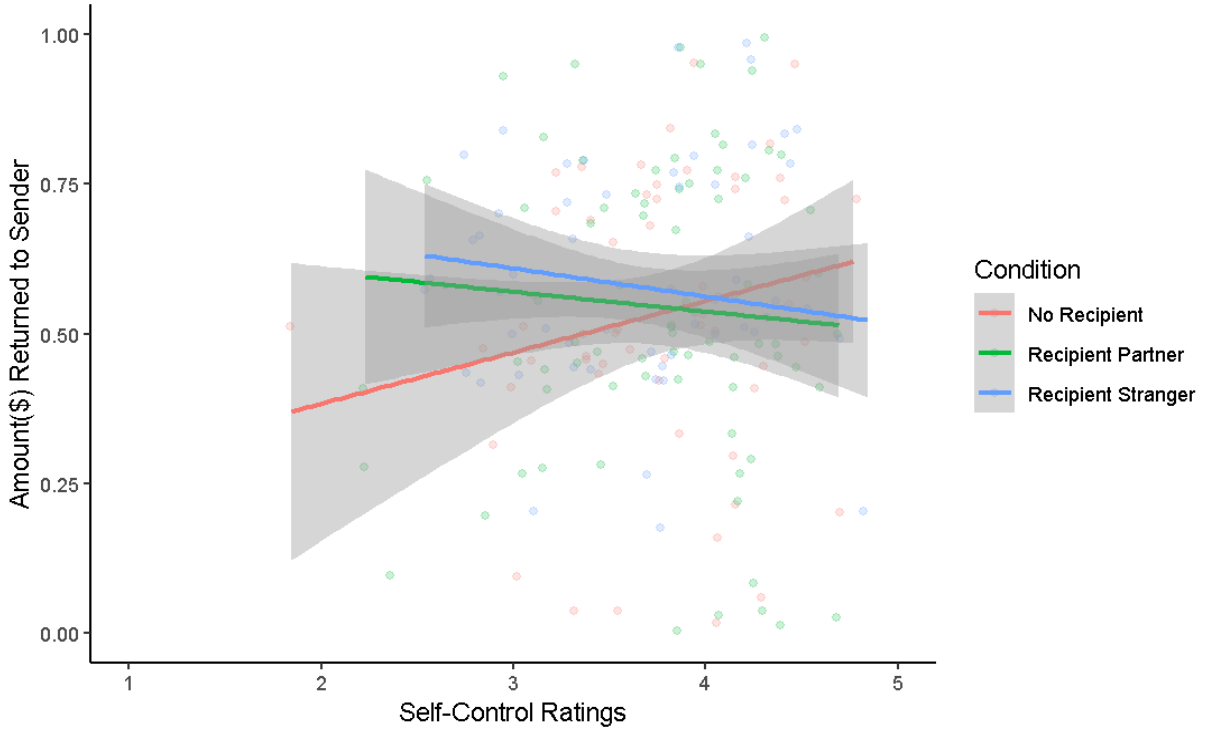


Figure 3.4: Interaction between the conditions and the correlation between self-control and cooperation in Experiment 2. Lines represent linear model fits. Shaded bands represent *95% CI* around the fitted regression lines. The sample includes participants who passed all attention checks ($N = 211$).

Using the same Helmert coding scheme, we regressed the amount of money participants returned in the Trust Game on five predictors: (1) Helmert Code 1, (2) Helmert Code 2, (3) self-control scores; (4) the interaction between Helmert Code 1 and self-control scores, and (5) the interaction between Helmert Code 2 and self-control scores.

In support of Prediction 2.2, there was no significant interaction between the conditions and the correlation between self-control scores and the amount of money returned. More specifically, the interaction between the coefficient for Helmert Code 1 and self-control scores was non-significant ($b = -0.05$, $SE = 0.07$, *95% CI* = [-0.18, 0.08], $\beta = -.36$, $p = .425$), neither

was the interaction between coefficient for Helmert Code 2 and self-control scores (b = -0.13, SE = 0.08, *95% CI* = [-0.30, 0.03], β = -.72, p = .108). The coefficients for Helmert Code 1, Self-control, and Helmert Code 2 were all non-significant (ps > .084).

Unlike in Experiment 1, by the letter of our pre-registration, these findings support the honest-signal hypothesis such that the association between self-control and cooperation did not depend on the experimental condition. We elaborate on this interpretation below.

Discussion

In Experiment 2, we attempted to replicate Experiment 1 in a new sample of Prolific workers. In contrast to the results in Experiment 1, we did not find a causal effect of social evaluation upon self-control, such that there was no difference in reported self-control across the different social evaluation conditions. In other words, contrary to our hypotheses, participants did not inflate their self-control regardless of whether they thought they were evaluated by a future partner, a stranger, or no one.

We also found no evidence that reported self-control predicted cooperation. Reported self-control did not significantly predict the amount of money returned, and neither of the Helmert-coded interactions significantly predicted cooperation. By the letter of our pre-registration, this would indicate support for the honest-signal hypothesis, because the association between self-control and cooperation did not depend on the experimental condition. However, it's worth noting that there was no overall association between self-control and cooperation

To better understand the mixed pattern of results from Experiments 1 and 2, we combined the data from both experiments and conducted a Mega-analysis to obtain a more robust test of our hypotheses.

Mega-analysis

We combined the data from both Experiments 1 and 2, and analyzed them as a single dataset (N = 664). Again, we report here data from the restricted sample of participants who have passed all comprehension checks (75%, N = 501). Results with the full sample differed, and are reported in detail in the Supplemental Materials.

We re-ran the same regression models used to test predictions 1 and 2. For each model, we also included a dummy coded predictor variable representing the experiment number to control for between-experiment differences, with Experiment 1 serving as the reference group (0 = *Experiment 1*, 1 = *Experiment 2*).

Prediction 1.1 vs. 1.2: Does social evaluation positively predict reported self-control scores only when reputation is at risk (1.1), or under any social evaluation (1.2)?

Only when reputation is at risk (Prediction 1.1). For each participant, we first calculated a composite of self-control scores (M = 3.40, SD = 0.64; McDonald's Ω = .88) by taking the average score of the items in the Brief Self-control Scale (Tangney et al., 2014). Figure 3.5 shows participants' average self-control scores in each condition.

The coefficient for Helmert Code 1 was positive and statistically significant (b = 0.11, SE = 0.05, 95% CI = [0.07, 0.21], β = .083, p = .038); and the coefficient for Helmert Code 2 was non-significant (b = 0.03, SE = 0.07, 95% CI = [-0.10, 0.15], β =.015, p = .070). Figure 3.5 shows participants' average self-control scores in each condition.

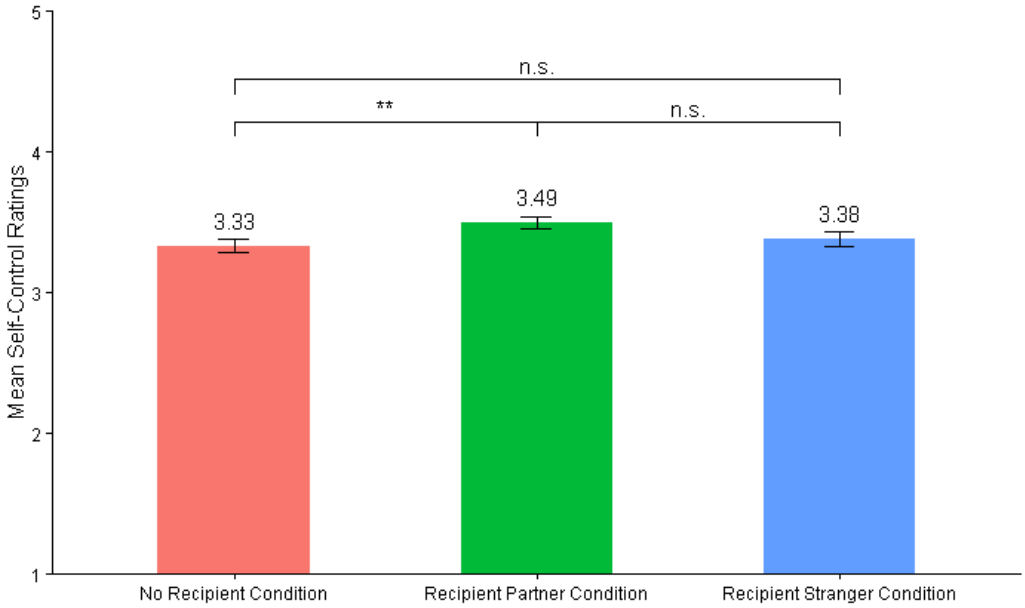


Figure 3.5: Self-control scores across all three conditions in the Mega-analysis. Error bars represent ±1 *SE*. The sample includes participants who passed all attention checks (N = 501).

The coefficient for the dummy code was positive and significant ($b = 0.57$, $SE = 0.05$, *95% CI* = [0.47, 0.67], $\beta = .44$, $p < .001$), indicating that participants in Experiment 2 reported higher self-control than those in Experiment 1.

These findings replicated those in Experiment 1, supporting the reputation-maintenance hypothesis, such that participants in the Recipient-partner condition increased their reported self-control prior to the Trust Game, compared to participants in the Recipient-stranger and No-recipient conditions.

Prediction 2.1 vs. 2.2: Is there an interaction between social evaluation and the correlation between reported self-control and cooperation (2.1) or not (2.2)?

Yes (Prediction 2.1). We first examined the correlations between self-control scores and the amount of money returned ($M = 0.64$, $SD = 0.27$) in each condition. Self-control scores significantly and negatively correlated with the amount of money returned in the Recipient-partner condition, $r(183) = -.278$, $p < .001$. The correlation was non-significant for both the

Recipient-stranger (p = .083) and the No-recipient conditions (p = .976). Figure 3.6 shows the correlations between self-control scores and the amount of money returned in each condition.

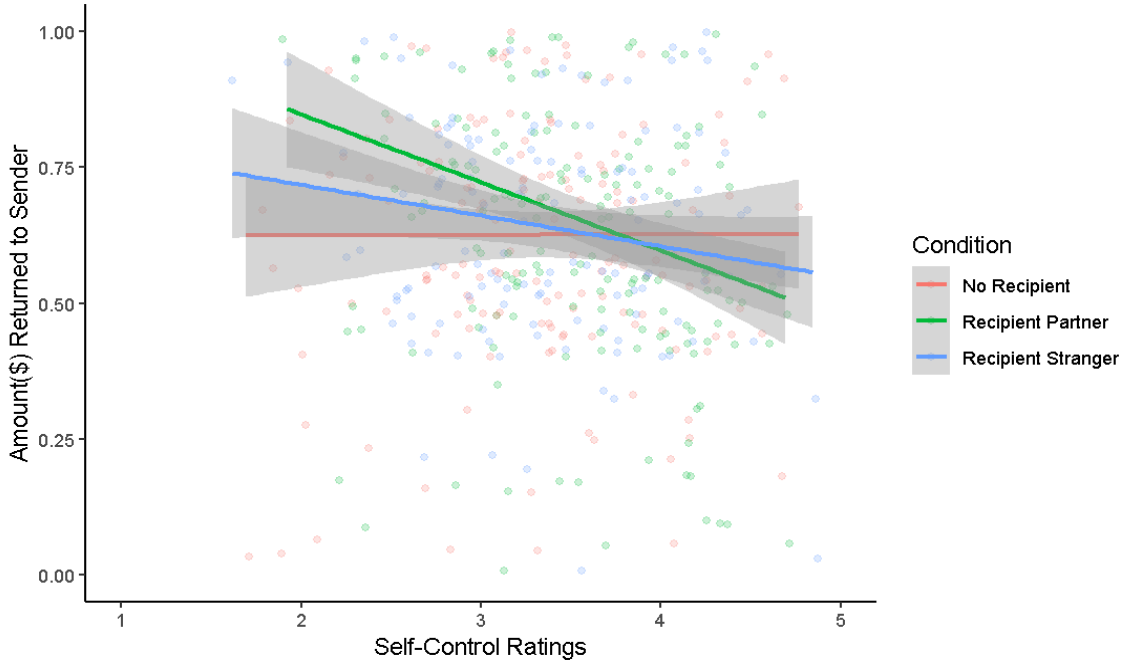


Figure 3.6: Interaction between the conditions and the correlation between self-control and cooperation in the Mega-analysis. Lines represent linear model fits. Shaded bands represent *95% CI* around the fitted regression lines. The sample includes participants who passed all attention checks (N = 501).

In support of Prediction 2.1, there was a significant interaction between the conditions and the correlation between self-control scores and the amount of money returned (M = 0.64, SD = 0.27), such that the interaction between the coefficient for Helmert Code 1 and self-control was negative and statistically significant (b = -0.10, SE = 0.034, *95% CI* [-0.18, -0.03], β = -.635, p = .008). However, the interaction between coefficient for Helmert Code 2 and self-control was non-significant (b = -0.08, SE = 0.05, *95% CI* = [-0.17, 0.01], β = -.39, p = .086). The coefficient for Helmert Code 1 was positive and significant (b = 0.38, SE = 0.13, *95% CI* = [0.12, 0.65], t = 2.89, β = .69, p = .004). The coefficients for Self-control and for Helmert Code 2 were non-significant (ps > .070).

The coefficient for the dummy code was negative and significant (b = -0.16, SE = 0.03, *95% CI* = [-0.21, -0.11], β = -.30, $p < .001$), indicating that participants in Experiment 2 shared less money than participants in Experiment 1.

These findings replicated those in Experiment1, partially supporting the cheap-talk hypothesis, such that participants with higher self-control scores in the Recipient-partner condition were less likely to send money to their partner, compared to participants in the Recipient-stranger and No-recipient conditions.

General Discussion

People who are more self-controlled in their behavior are typically also more cooperative, but the causal association between self-control and cooperation remains unclear because the two frequently covary with social evaluation. Moreover, few studies have simultaneously examined the causal effect of social evaluation on self-control, and its downstream effect on cooperation. To address these issues, we conducted two preregistered experiments in which participants ostensibly shared their reported self-control scores with other people, who then decided whether or not to initiate a cooperative exchange in the Trust Game. Overall, the two experiments and a Mega-analysis combining both datasets provided mixed evidence that social evaluation causally affects reported self-control, and moderates the association between reported self-control and cooperation.

In Experiment 1 and the Mega-analysis, participants who thought that their self-control scores would be shared with someone who would judge their trustworthiness in the future self-reported higher self-control scores than those who thought their scores would be shared with a stranger, or no one. Correspondingly, we found little evidence that participants inflated their self-control scores simply because they thought their self-control scores would be shared with any other person. Thus, we found some support for the reputation-maintenance hypothesis of self-

control, such that people try to appear more self-controlled in front of people with whom they might share a cooperative relationship.

When we turned to the question about whether self-control is an honest signal of cooperation or mere cheap talk, we found partial support for the cheap-talk hypothesis. In Experiment 1 and the Mega-analysis, reported self-control was less predictive of the amount of money returned when participants thought their self-control scores were shared with their game partner. This pattern provides partial support for the cheap-talk hypothesis, suggesting that participants who inflated their self-control scores did not necessarily act more cooperatively.

In sum, our results suggest that people inflate their self-control scores to appear like better cooperative partners, but these inflated self-control scores may not translate into greater cooperation.

Reported self-control is sensitive reputational risk

To summarize, our primary findings were: first, participants failed to inflate their self-control scores when their scores would be shared with someone that participants did not expect to interact with again, but did inflate their scores when the other person would judge their trustworthiness in an ensuing economic game. Second, participants who inflated their self-control scores did not act more cooperatively in the actual game. This suggests that our psychological system is not only blindly sensitive to any kind of social evaluation, but sensitive to whether the social evaluation is consequential for future reciprocity.

Why did reported self-control fail to predict cooperation?

A key caveat to our findings is that reported self-control did not show the canonical positive association with cooperation. In both experiments, reported self-control did not significantly predict the amount of money returned in the Trust Game. Most surprisingly, in Experiment 1 and the Mega-analysis, reported self-control was significantly *negatively*

correlated with cooperation in the Recipient-partner condition. This pattern complicates the established link between expressions of self-control and cooperativeness and provides a more nuanced interpretation of the cheap-talk hypothesis: people may use reported self-control as cheap signals of cooperativeness even when there is no association between reported self-control and cooperation.

Visual inspection of the interaction between self-control and social evaluation in Figure 3.2 suggested that the difference in the amount of money returned across social evaluation conditions was largest among participants who reported low self-control. Thus, to further investigate this pattern of results, we explored whether the unexpected negative association between self-control and cooperation in the Recipient-partner condition was driven by participants who reported low self-control. We divided participants into tertiles based on their self-control ratings and tested whether the amount of money returned differed across social evaluation conditions within each tertile. Analyses and results are reported in Supplemental Materials.

Notably, the effect of social evaluation on the amount of money returned differed systematically as a function of reported self-control. Participants whose self-control scores were in the low tertile returned significantly more money in the Recipient-partner condition than in the No-recipient condition. Participants whose self-control scores were in the mid tertile returned significantly more money in the Recipient-partner condition than in the Recipient-stranger condition. By contrast, participants whose self-control scores were in the highest tertile returned similar amounts of money across conditions, regardless of social evaluation contexts.

These exploratory findings suggest that the relationship between self-control and cooperation may be more sensitive to social evaluation contexts than prior work has assumed.

Participants who reported lower self-control seemed much more sensitive to reputational concern in their cooperation behavior than those who reported higher self-control. One possibility is that participants who reported lower self-control were motivated to cooperate in order to compensate for the unfavorable reputational signals sent by their low self-control scores. This possibility aligns with the fact that they only increased cooperation behavior with a partner who they thought had seen their self-control scores, and is consistent with a reputation-management account. By contrast, participants who reported high self-control scores may not have felt this need to compensate. Future research should systematically explore individual differences that contribute to the relationship between self-control and cooperation.

Mixed findings & demographic differences

We do not know why the predicted effects emerged in Experiment 1 and the Mega-analysis, but not in Experiment 2, because the two experiments were conducted with markedly different samples. Experiment 1 recruited undergraduate participants through SONA, whereas Experiment 2 recruited online Prolific workers. These samples likely differed in many ways, including age, socioeconomic context, familiarity with online experiments, and how socially meaningful or reputationally consequential the interaction seemed. For instance, Prolific workers may have been less concerned about their reputation because the game may have felt more abstract when it is through an online platform among strangers than on campus among peers.

At the same time, our sample source was confounded with gender composition. Participants in Experiment 1 were mostly undergraduate women (80%, $N = 294$), whereas participants in Experiment 2 had a more balanced gender distribution (52.5% women, $N = 156$). Because prior work suggests that women tend to score higher on traits relevant to the present paradigm, including empathy (McCauley et al., under review), cooperativeness (Balliet et al., 2011), and Honesty-Humility (Lee & Ashton, 2020), gender composition may also have

contributed to the different findings across experiments. However, the present experiments cannot determine whether the discrepancy between Experiment 1 and Experiment 2 was due to sample source, gender composition, or some other difference between the samples. Future research should disentangle these possibilities by comparing more closely matched samples or directly testing these factors as moderators.

Comprehension of the social-evaluation manipulation

While our primary analyses focused on participants who passed all attention checks, intent-to-treat analyses using the full sample produced different patterns in Experiment 1 and the Mega-analysis. These differences are consistent with the idea that sensitivity to the cooperative consequences of social evaluation was crucial to our findings supporting the reputation-maintenance hypothesis.

The clearest difference between the full sample and the restricted sample emerged in Experiment 1. In the restricted sample, participants reported higher self-control only when their scores were shared with a future cooperative partner. In the full sample, however, participants reported higher self-control whenever their scores were shared, whether it is with a future cooperative partner or a stranger. Therefore, contrary to the restricted-sample analyses, the full-sample analyses supported the misfiring hypothesis: participants inflated their self-control whenever they were observed, regardless of whether that observation was consequential for cooperation.

We also found some differences in the association between reported self-control and cooperation. In Experiment 1 and the Mega-analysis, unlike in the restricted sample, social evaluation did not moderate the relationship between reported self-control and cooperation. By the letter of pre-registration, these results provided support for the honest-signaling hypothesis.

Although it is noteworthy that in Experiment 1, reported self-control was not significantly associated with cooperation, and in Mega-analysis, this association was negative.

One possibility is that participants who failed attention checks were less engaged in the game and responded primarily to the presence of observation, with little motivation to cooperate regardless of the social signals conveyed by their self-control scores. Overall, these differences highlight that the perception of and sensitivity to the cooperative consequences of social evaluation matters for strategic signaling behaviors.

Constraints on generalizability

Finally, there are some constraints on the generalizability of our findings. First, cultural norms may differ in how self-control relates to cooperation. Although we replicated our study among two relatively diverse populations, our participants are all living in the US, and it is unclear whether our findings would diverge in other cultures.

Second, while the Trust Game is canonically used to measure cooperation in the laboratory, it may not capture the complexity of cooperation in the real world, such as participants' relationship to their cooperation partners. The Trust Game involves a situation in which two people need to mutually cooperate in order to maximize monetary benefits. In the real world, situations like this do not typically involve strangers–as is the case in a Trust Game – but often involve people who already share social relationships, such as friends or business partners,

Finally, real-world self-control signaling often occurs behaviorally, rather than through direct claims. We used reported self-control as a measure of self-control signaling in order to make the signal explicit and reduce ambiguity about what was being evaluated. However, strategic signaling in everyday life rarely involves people directly themselves as "I have iron discipline". Instead, people often signal self-control through behavior such as choosing salad over fries or resisting other forms of temptation. Future research should examine the effect of

social evaluation on self-control through observable self-control behavior, and measure whether these behavioral signals are more predictive of cooperation than self-reported self-control.

Conclusion

Across two preregistered experiments, we tested the causal effect or social evaluation on reported self-control and its relation to cooperation. Our findings complicate the established link between self-control and cooperation: reported self-control did not reliably predict cooperative behavior in a one-shot Trust Game. That said, in Experiment 1 and the Mega-analysis, we found some evidence that people inflated their reported self-control when social evaluation is directly consequential without actually behaving more cooperatively. These findings suggest that reported self-control can sometimes serve as cheap signals for cooperativeness. However, this effect emerged among undergraduate students in Experiment 1 and in the Mega-analysis, but not among Prolific workers in Experiment 2. These mixed findings suggest that the effect of social evaluation may vary with individual differences and demographic differences. Future work should examine more diverse populations with the use of more naturalistic behavioral measures of self-control signaling.

Acknowledgements

Chapter 3, is currently being prepared for submission for publication of the material. Liu, Xingyu S.; McCauley, Thomas G.; and McCullough, Michael E. The dissertation author was the primary researcher and author of this material.

REFERENCES

Balliet, D., Li, N. P., Macfarlan, S. J., & Van Vugt, M. (2011). Sex differences in cooperation: A meta-analytic review of social dilemmas. *Psychological Bulletin*, *137*(6), 881–909. https://doi.org/10.1037/a0025354

Barclay, P. (2013). Strategies for cooperation in biological markets, especially for humans. *Evolution and Human Behavior*, *34*(3), 164–175. https://doi.org/10.1016/j.evolhumbehav.2013.02.002

Bateson, M., Nettle, D., & Roberts, G. (2006). Cues of being watched enhance cooperation in a real-world setting. *Biology Letters*, *2*(3), 412–414. https://doi.org/10.1098/rsbl.2006.0509

Bradley, A., Lawrence, C., & Ferguson, E. (2018). Does observability affect prosociality? *Proceedings of the Royal Society B: Biological Sciences*, *285*(1875), 20180116. https://doi.org/10.1098/rspb.2018.0116

Buyukcan-Tetik, A., Finkenauer, C., Siersema, M., Heyden, K. V., & Krabbendam, L. (2015). Social relations model analyses of perceived self-control and trust in families. *Journal of Marriage and Family*, *77*(1), 209–223. https://doi.org/10.1111/jomf.12154

Buyukcan-Tetik, A., & Pronk, T. M. (2023). Partner self-control and intrusive behaviors: A gender-specific examination of the mediating role of trust. *Current Psychology: A Journal for Diverse Perspectives on Diverse Psychological Issues*, *42*(14), 11782–11792. https://doi.org/10.1007/s12144-021-02462-4

Cain, D. M., Dana, J., & Newman, G. E. (2014). Giving versus giving in. *Academy of Management Annals*, *8*(1), 505–533. https://doi.org/10.5465/19416520.2014.911576

Dewitte, S., & Cremer, D. D. (2001). Self-control and cooperation: different concepts, similar decisions? A question of the right perspective. *The Journal of Psychology*, *135*(2), 133–153. https://doi.org/10.1080/00223980109603686

Duckworth, A. L. (2011). The significance of self-control. *Proceedings of the National Academy of Sciences*, *108*(7), 2639–2640. https://doi.org/10.1073/pnas.1019725108

Fitouchi, L., André, J.-B., & Baumard, N. (2022). Moral disciplining: The cognitive and evolutionary foundations of puritanical morality. *Behavioral and Brain Sciences*, *46*, 1–71. https://doi.org/10.1017/S0140525X22002047

Franzen, A., & Pointner, S. (2012). Anonymity in the dictator game revisited. *Journal of Economic Behavior & Organization*, *81*(1), 74–81. https://doi.org/10.1016/j.jebo.2011.09.005

Gai, P. J., & Bhattacharjee, A. (2022). Willpower as moral ability. *Journal of Experimental Psychology: General*, *151*(8), 1999–2006. https://doi.org/10.1037/xge0001169

Gomillion, S., Lamarche, V. M., Murray, S. L., & Harris, B. (2014). Protected by your self-control: The influence of partners' self-control on actors' responses to interpersonal risk.

Social Psychological and Personality Science, *5*(8), 873–882. https://doi.org/10.1177/1948550614538462

Harrison, J. M. D., & McKay, R. (2013). Give me strength or give me a reason: Self-control, religion, and the currency of reputation. *The Behavioral and Brain Sciences*, *36*(6), 688–689; discussion 707-726. https://doi.org/10.1017/S0140525X13001003

Kocher, M. G., Martinsson, P., Myrseth, K. O. R., & Wollbrant, C. E. (2017). Strong, bold, and kind: Self-control and cooperation in social dilemmas. *Experimental Economics*, *20*(1), 44–69. https://doi.org/10.1007/s10683-015-9475-7

Koval, C. Z., vanDellen, M. R., Fitzsimons, G. M., & Ranby, K. W. (2015). The burden of responsibility: Interpersonal costs of high self-control. *Journal of Personality and Social Psychology*, *108*(5), 750–766. https://doi.org/10.1037/pspi0000015

Lee, K., & Ashton, M. C. (2020). Sex differences in HEXACO personality characteristics across countries and ethnicities. *Journal of Personality*, *88*(6), 1075–1090. https://doi.org/10.1111/jopy.12551

Lie-Panis, J., & André, J.-B. (2022). Cooperation as a signal of time preferences. *Proceedings. Biological Sciences*, *289*(1973), 20212266. https://doi.org/10.1098/rspb.2021.2266

Ma, F., Gu, X., Tang, L., Luo, X., Compton, B. J., & Heyman, G. D. (2023). If they won't know, i won't wait: anticipated social consequences drive children's performance on self-control tasks. *Psychological Science*, *34*(11), 1220–1228. https://doi.org/10.1177/09567976231198194

Marcus, Z. J., & McCullough, M. E. (2021). Does religion make people more self-controlled? A review of research from the lab and life. *Current Opinion in Psychology, Religion*, *40*, 167–170. https://doi.org/10.1016/j.copsyc.2020.12.001

McAuliffe, W. H. B., Forster, D. E., Pedersen, E. J., & McCullough, M. E. (2018). Experience with anonymous interactions reduces intuitive cooperation. *Nature Human Behaviour*, *2*(12), 909–914. https://doi.org/10.1038/s41562-018-0454-9

McCauley, T. G., McAuliffe, W. H., Mulhinch, M., Pedersen, E. J., Schug, J., Ohtsubo, Y., & McCullough, M. E. (under review). *Does Measurement Bias Explain Sex Differences in Self-Reported Empathy?* (6ps2r_v1). PsyArXiv. https://doi.org/10.31234/osf.io/6ps2r

Nowak, M. A. (2006). Five rules for the evolution of cooperation. *Science*, *314*(5805), 1560–1563. https://doi.org/10.1126/science.1133755

Peetz, J., & Kammrath, L. (2013). Folk understandings of self-regulation in relationships: Recognizing the importance of self-regulatory ability for others, but not the self. *Journal of Experimental Social Psychology*, *49*(4), 712–718. https://doi.org/10.1016/j.jesp.2013.02.007

Pronk, T. M., Karremans, J. C., & Wigboldus, D. H. J. (20110117). How can you resist? Executive control helps romantically involved individuals to stay faithful. *Journal of Personality and Social Psychology*, *100*(5), 827. https://doi.org/10.1037/a0021993

Rachlin, H. (2016). Social cooperation and self-control. *Managerial and Decision Economics*, *37*(4–5), 249–260. https://doi.org/10.1002/mde.2714

Righetti, F., & Finkenauer, C. (2011). If you are able to control yourself, I will trust you: The role of perceived self-control in interpersonal trust. *Journal of Personality and Social Psychology*, *100*(5), 874–886. https://doi.org/10.1037/a0021827

Rounding, K., Lee, A., Jacobson, J. A., & Ji, L.-J. (2012). Religion replenishes self-control. *Psychological Science*, *23*(6), 635–642. https://doi.org/10.1177/0956797611431987

Sedikides, C., Campbell, W. K., Reeder, G. D., & Elliot, A. J. (1999). The relationship closeness induction task. *Representative Research in Social Psychology, 23*, 1–4.

Stavrova, O., & Kokkoris, M. D. (2019). Struggling to be liked: The prospective effect of trait self-control on social desirability and the moderating role of agreeableness. *International Journal of Psychology: Journal International De Psychologie*, *54*(2), 232–236. https://doi.org/10.1002/ijop.12444

Tangney, J. P., Baumeister, R. F., & Boone, A. L. (2004). High self-control predicts good adjustment, less pathology, better grades, and interpersonal success. *Journal of Personality*, *72*(2), 271–324. https://doi.org/10.1111/j.0022-3506.2004.00263.x

Thielmann, I., Heck, D. W., & Hilbig, B. E. (2016). Anonymity and incentives: An investigation of techniques to reduce socially desirable responding in the Trust Game. *Judgment and Decision Making*, *11*(5), 527–536. https://doi.org/10.1017/S1930297500004605

Winking, J., & Mizer, N. (2013). Natural-field dictator game shows no altruistic giving. *Evolution and Human Behavior*, *34*(4), 288–293. https://doi.org/10.1016/j.evolhumbehav.2013.04.002

Yamagishi, T., Hashimoto, H., & Schug, J. (2008). Preferences versus strategies as explanations for culture-specific behavior. *Psychological Science*, *19*(6), 579–584. https://doi.org/10.1111/j.1467-9280.2008.02126.x

Chapter 3 Supplemental Materials

List of all measures included in Experiments 1 and 2, but not reported in analyses

Across both experiments, we included some additional measures. We did not analyze data from those additional measures in the present article, because they are part of a broader project, and thus outside the scope of the current research. They were as follows:

Balanced inventory of desirable responding (BIDR). Paulhus, D. L. (1988). Balanced inventory of desirable responding (BIDR). Acceptance and Commitment Therapy. Measures Package, 41, 79586-79587.

Prosocial and Intrinsic Motivations scale. Grant, A. M. (2008). Does intrinsic motivation fuel the prosocial fire? Motivational synergy in predicting persistence, performance, and productivity. Journal of Applied Psychology, 93(1), 48. We adapted items from this scale to measure participants' motives for sending money to their partner in the Trust Game.

Honesty-Humility (H). Ashton, M. C., & Lee, K. (2008). The HEXACO model of personality structure and the importance of the H factor. Social and Personality Psychology Compass, 2(5), 1952-1962.

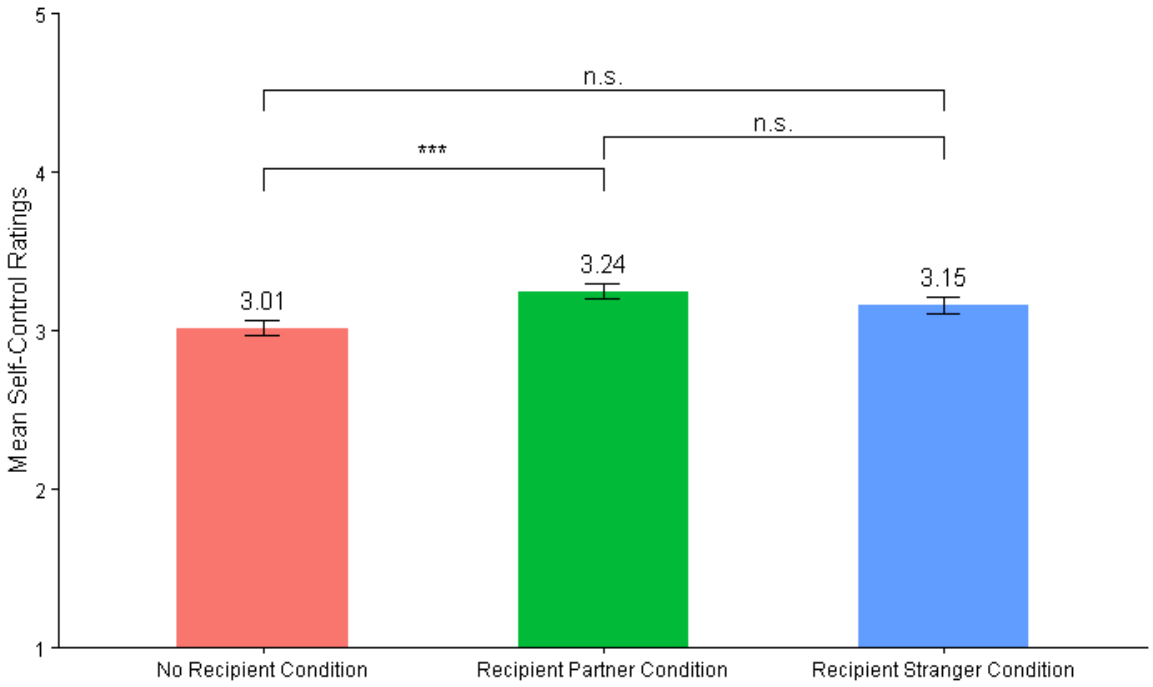
Experiment 1 Result with Full Sample (N = 367)

Did participants think they would see their Trust Game partner after the experiment?

We wanted to ensure that participants' cooperation behavior in the Trust Game was not inflated by the perceived reputational risk of interacting with the Trust Game partner after the experiment. As expected, participants on average thought it would be relatively unlikely that they would see their Trust Game partner after the experiment ended (M = 1.67, SD = 1.00).

Prediction 1.1 vs. 1.2: Does social evaluation positively predict reported self-control scores only when reputation is at risk (1.1), or under any social evaluation (1.2)?

Under any social evaluation (Prediction 1.2). For each participant, we first calculated a composite of self-control scores (M = 3.13, SD = 0.57; *McDonald's* Ω = .85) by taking the average score of the items in the Brief Self-control Scale (Tangney et al., 2014). Supplemental Figure 3.1 shows participants' average self-control scores in each condition.



Supplemental Figure 3.1: Self-control scores across the three conditions with full sample in Experiment 1. Error bars represent ±1 *SE*. The sample includes all participants (N = 367).

Using the same Helmert coding scheme as in the main analyses, we regressed self-control composite scores on: (1) A Helmert-coded predictor variable representing the effects of the Recipient-Partner vs. Recipient-Stranger and No-Recipient conditions, and (2) A Helmert-coded predictor variable representing the effects of the Recipient-Stranger vs. the No-Recipient conditions.

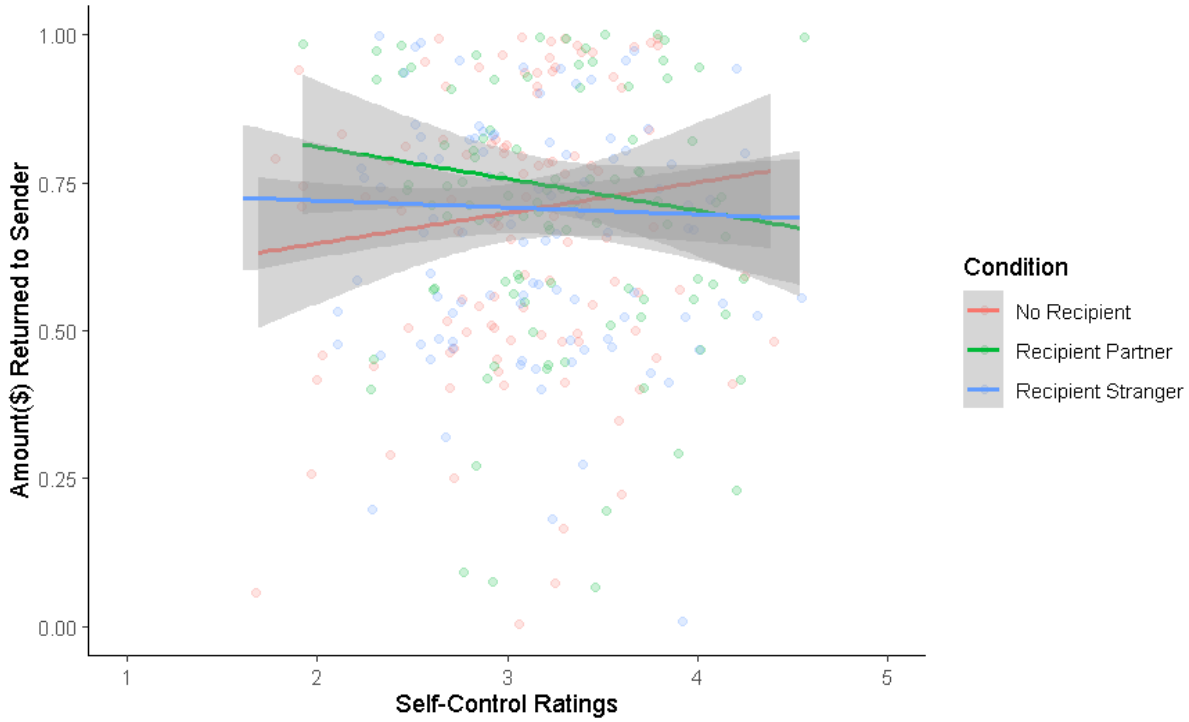
In support of Prediction 1.2, there was a main effect of observation. More specifically, the coefficient for Helmert Code 1 was positive and statistically significant (b = 0.16, SE = 0.06,

95% CI = [0.04, 0.29], β = .29, *p* = .010); and the coefficient for Helmert Code 2 was also positive and significant (*b* = 0.14, *SE* =0.07, *95% CI* = [0.00, 0.28], β = .10, *p* = .048).

Contrary to the restricted sample, these findings support the misfiring hypothesis, such that participants in both the Recipient-partner condition and in the Recipient-stranger condition increased their self-control prior to the Trust Game, compared to participants in the No-recipient condition.

Prediction 2.1 vs. 2.2: Is there an interaction between social evaluation and the correlation between reported self-control and cooperation (2.1) or not (2.2)?

No (Prediction 2.2). Self-control scores were not significantly correlated with the amount of money returned (*M* = 0.72, *SD* = 0.26) in any condition (*p*s > .192). Supplemental Figure 3.2 shows correlations between the self-control composite scores and the amount of money returned in each condition.



Supplemental Figure 3.2: Interaction between the conditions and the correlation between self-control and cooperation in Experiment 1. Lines represent linear model fits. Shaded bands represent *95% CI* around the fitted regression lines. The sample includes all participants (*N* = 367).

Using the same Helmert coding scheme as above, we regressed the amount of money participants returned in the Trust Game on five predictors: (1) Helmert Code 1, (2) Helmert Code 2, (3) self-control scores; (4) the interaction between Helmert Code 1 and self-control scores, and (5) the interaction between Helmert Code 2 and self-control scores.

In support of Prediction 2.2, there was no significant interaction between the conditions and the correlation between self-control composite scores and the amount of money returned. More specifically, the interaction between the coefficient for Helmert Code 1 and self-control was non-significant (b = -0.07, SE = 0.05, *95% CI* = [-0.18, 0.03], β = -.43, p = .159); neither was the interaction between coefficient for Helmert Code 2 and self-control (b = -0.06, SE = 0.06, *95% CI* = [-0.18, 0.05], β = -.32, p = .275). The coefficients for Helmert Code 1, Self-control, and Helmert Code 2 were all non-significant (ps > .107).

Unlike in the restricted sample, by the letter of our pre-registration, these results provided support for the honest-signaling hypothesis, such that social evaluation did not moderate the effect of reported self-control on cooperation. However, it is noteworthy that there was no main effect of reported self-control on cooperation.

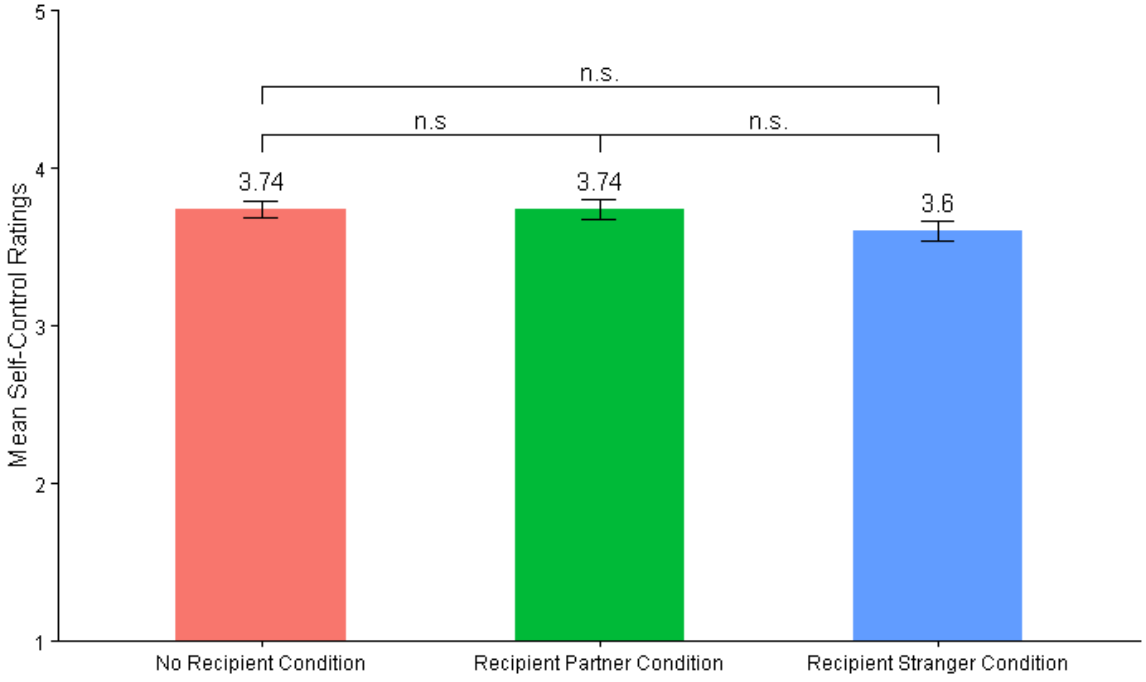
Experiment 2 Result with Full Sample (N = 297)

Did participants think they would see their Trust Game partner after the experiment?

We wanted to ensure that participants' cooperation behavior in the Trust Game was not inflated by the perceived reputational risk of interacting with the Trust Game partner after the experiment. Participants on average thought it would be somewhat unlikely that they would see their Trust Game partner after the experiment ended (M = 2.68, SD = 1.41).

Prediction 1.1 vs. 1.2: Does social evaluation positively predict reported self-control scores only when reputation is at risk (1.1), or under any social evaluation (1.2)?

Neither. Social evaluation did not predict reported self-control. For each participant, we first calculated a composite of self-control scores (M = 0.53, SD = 0.29; *McDonald's* Ω = .86) by taking the average score of the items in the Brief Self-control Scale (Tangney et al., 2014). Supplemental Figure 3.3 shows participants' average self-control scores in each condition.



Supplemental Figure 3.3: Self-control scores across the three conditions with full sample in Experiment 2. Error bars represent ±1 *SE*. The sample includes all participants (N = 297).

Using the same Helmert coding scheme as above, we regressed self-control composite scores on: (1) A Helmert-coded predictor variable representing the effects of the Recipient-Partner vs. Recipient-Stranger and No-Recipient conditions, and (2) A Helmert-coded predictor variable representing the effects of the Recipient-Stranger vs. the No-Recipient conditions.

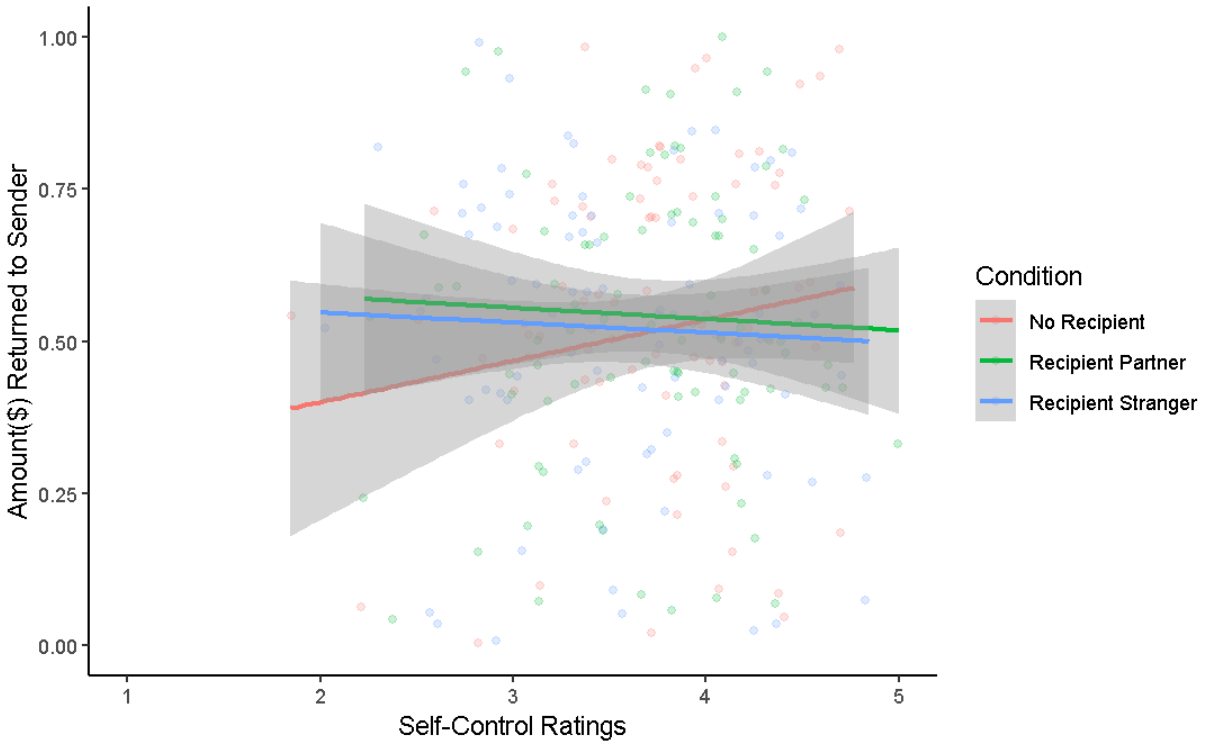
Contrary to our predictions, there was no main effect of social evaluation on reported self-control. More specifically, the coefficient for Helmert Code 1 was non-significant (b = 0.07,

SE = 0.07, *95% CI* = [-0.08, 0.22], β = .05, *p* = .357), and the coefficient for Helmert Code 2 was also non-significant (*b* = -0.14, *SE* = 0.09, *95% CI* = [-0.31, 0.04], β = -.09, *p* = .121).

Like in the restricted sample, these findings did not support our hypothesis that self-control is driven by social evaluation.

Prediction 2.1 vs. 2.2: Is there an interaction between social evaluation and the correlation between reported self-control and cooperation (2.1) or not (2.2)?

No (Prediction 2.2). Self-control scores were not significantly correlated with the amount of money returned in any condition (*p*s > .208). Supplemental Figure 3.4 shows correlations between the self-control composite scores and the amount of money returned in each condition.



Supplemental Figure 3.4: Interaction between the conditions and the correlation between self-control and cooperation in Experiment 2. Lines represent linear model fits. Shaded bands represent *95% CI* around the fitted regression lines. The sample includes all participants (*N* = 297).

Using the same Helmert coding scheme as above, we regressed the amount of money participants returned in the Trust Game on five predictors: (1) Helmert Code 1, (2) Helmert Code

2, (3) self-control composite scores; (4) the interaction between Helmert Code 1 and self-control composite scores, and (5) the interaction between Helmert Code 2 and self-control composite scores.

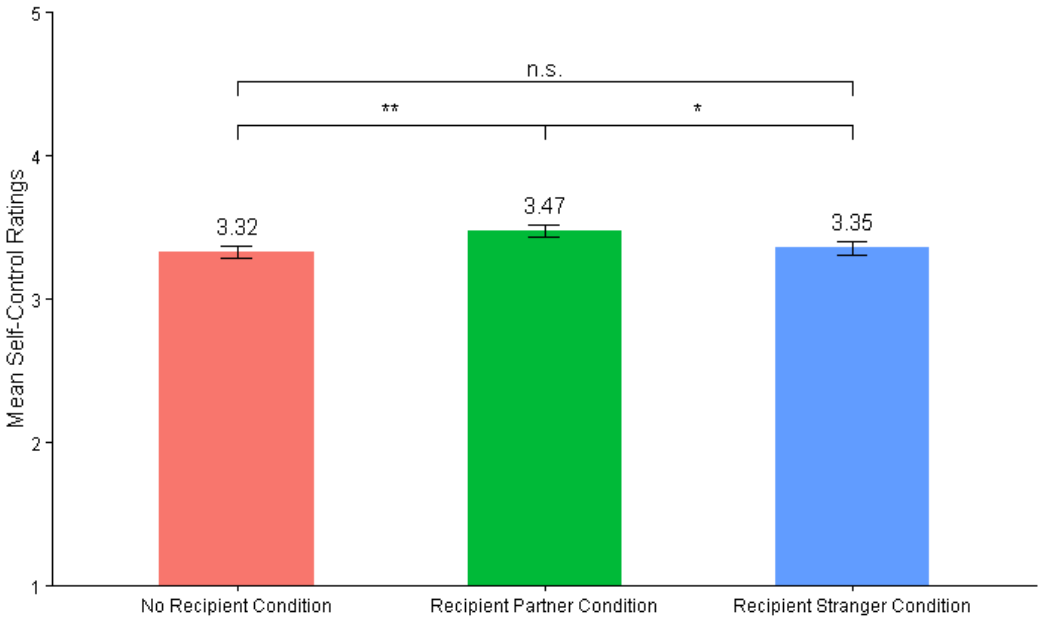
In support of Prediction 2.2, there was no significant interaction between the conditions and the correlation between self-control scores and the amount of money returned. More specifically, the interaction between the coefficient for Helmert Code 1 and the self-control composite was non-significant (b = -0.04, SE = 0.06, *95% CI* = [-0.16, 0.07], β = -.27, p = .453), and neither was the interaction between coefficient for Helmert Code 2 and the self-control composite (b = -0.08, SE = 0.07, *95% CI* = [-0.22, 0.05], β = -.44, p = .228).

Like in the restricted sample, by the letter of our pre-registration these results provided support for the honest-signaling hypothesis, such that social evaluation did not moderate the effect of reported self-control on cooperation. However, it is noteworthy that there was no main effect of reported self-control on cooperation.

Mega-analysis Results with Full Sample (N = 664)

Prediction 1.1 vs. 1.2: Does social evaluation positively predict reported self-control scores only when reputation is at risk (1.1), or under any social evaluation (1.2)?

Only when reputation is at risk (Prediction 1.1). In support of Prediction 1.1, the coefficient for Helmert Code 1 was positive and statistically significant (b = 0.12, SE = 0.05, *95% CI* = [0.03, 0.22], β =.09, p =.013), and the coefficient for Helmert Code 2 was non-significant (b = 0.02, SE = 0.06, *95% CI* = [-0.09, 0.13], β = .01, p = .723). Supplemental Figure 3.5 shows participants' average self-control scores *(M* = 3.38, SD = 0.65; *McDonald's Ω* = .87) in each condition.



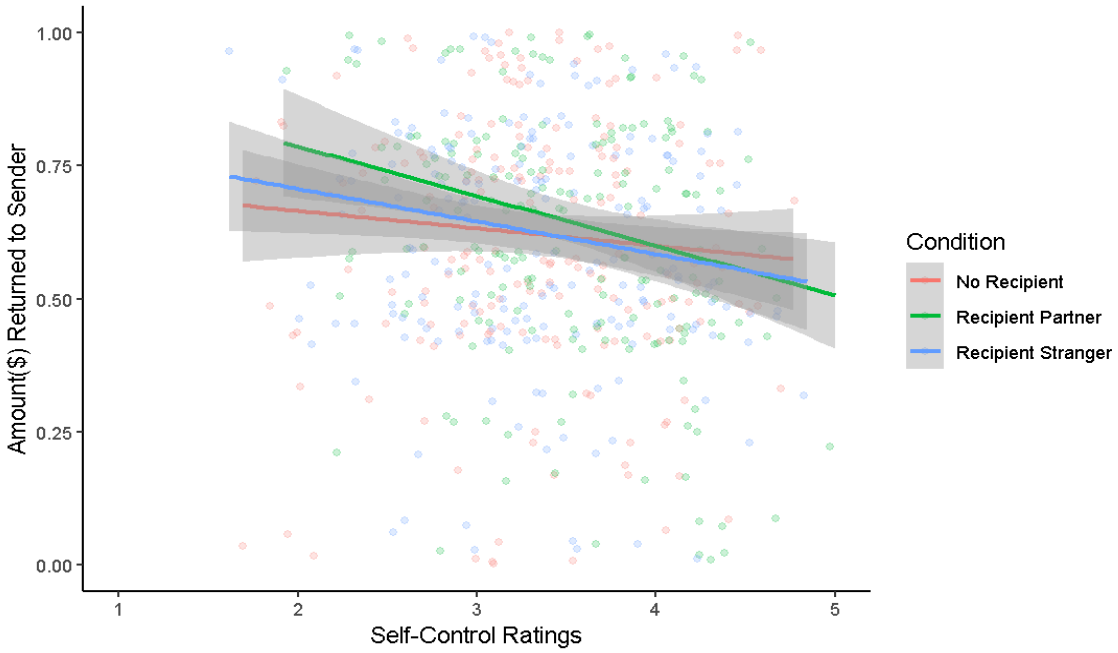
Supplemental Figure 3.5: Self-control scores across the three conditions with full sample in the Mega-analysis. Error bars represent ±1 *SE*. The sample includes all participants (N = 664).

The coefficient for the dummy code was positive and significant ($b = 0.56$, $SE = 0.05$, *95% CI* = [0.466, 0.645], $\beta = .43$, $p < .001$), indicating that participants in Experiment 2 reported higher self-control than those in Experiment 1.

Like in the restricted sample, these results provide support for the reputation-maintenance hypothesis.

Prediction 2.1 vs. 2.2: Is there an interaction between social evaluation and the correlation between reported self-control and cooperation (2.1) or not (2.2)?

No (Prediction 2.2). Self-control scores significantly and negatively correlated with the amount of money returned ($M = 0.63$, $SD = 0.29$) in the Recipient-partner condition ($r\,(218) = -.20$, *95% CI* = [−.32, −.07], $p = .003$), as well as in the Recipient-stranger condition ($r\,(216) = -.15$, *95% CI* = [−.27, −.01], $p = .031$). The correlation was non-significant in the No-recipient condition ($p = .282$). Supplemental Figure 3.6 shows correlations between the self-control scores and the amount of money returned in each condition.



Supplemental Figure 3.6: Interaction between the conditions and the correlation between self-control and cooperation in the Mega-analysis. Lines represent linear model fits. Shaded bands represent *95% CI* around the fitted regression lines. The sample includes all participants (N = 664).

In support of Prediction 2.2, there was no significant interaction between the conditions and the correlation between self-control scores and the amount of money returned. More specifically, the interaction between the coefficient for Helmert Code 1 and the self-control composite was non- significant (b = -0.05, SE = 0.04, *95% CI* = [-0.12, -0.02], β = -.29, p = .153), neither was the interaction between coefficient for Helmert Code 2 (b = -0.057, SE = 0.04, *95% CI* = [-0.14, 0.02], β = -.28, p = .152). The coefficients for Helmert Code 1, Self-control, and Helmert Code 2 were all non-significant (ps > .098).

The coefficient for the dummy code was negative and significant (b = -0.20, SE = 0.02, *95% CI* = [-0.24, -0.15], β = -.34, p < .001), indicating that participants in Experiment 2 shared less money than participants in Experiment 1.

Unlike in the restricted sample, by the letter of our pre-registration, these results provided support for the honest-signaling hypothesis, such that social evaluation did not moderate the effect of reported self-control on cooperation.

Individual differences in self-control

To examine whether the effect of social evaluation on the amount of money returned varied as a function of reported self-control, we conducted a 3 (Social evaluation: Recipient-Partner, Recipient-Stranger, No Recipient) × 3 (Self-Control: Low, Mid, High) between-subjects ANOVA. Participants were divided into self-control groups by binning their self-control scores into tertiles (Low: ≤ 2.92, n = 100; Mid: 2.93–3.38, n = 99; High: > 3.38, n = 91).

There was a significant main effect of social evaluation on the amount of money returned, $F(2, 281) = 3.09$, $p = .047$, and a significant difference between social evaluation and self-control group, $F(4, 281) = 2.98$, $p = .020$. The main effect of the self-control group was not significant ($p = .551$).

To probe this interaction, we tested the effect of social evaluation separately within each self-control group. Social evaluation positively and significantly predicted the amount of money returned among participants in the low self-control group, $F(2, 97) = 6.43$, $p = .002$, and those in the mid self-control group, $F(2, 96) = 3.51$, $p = .034$. However, social evaluation did not predict the amount of money returned among participants in the high self-control group, $F(2, 88) = 0.12$, $p = .885$.

We then conducted follow-up Tukey comparisons within the low and mid self-control group. Among participants in the low self-control group, the amount of money returned was significantly higher in the Recipient-partner condition compared to the No-recipient condition (difference = 0.20, $p = .002$, *95% CI* = [0.07, 0.33]), with no other significant pairwise differences (all ps> .196). Within the mid self-control group, cooperation was significantly

higher in the Recipient-partner condition compared to the Recipient-stranger condition (difference = 0.17, $p = .026$, *95% CI* = [0.02, 0.32]), with no other significant pairwise differences (all ps > .205).

CONCLUSION

When Alice finally stepped back through the looking glass, she didn't return to a simpler world, but to the same old world with a new set of eyes. In similar ways, this dissertation offers a richer understanding of choices – not as transparent windows into preferences, but a looking glass that is multifaceted, tricky, but informative.

Across the three lines of research in this dissertation, I traced a common thread of meaning-making. Making sense of choices and preferences — both our own and others' — is rarely straightforward, and rarely passive. Whether we are navigating conflicts between our current and aspirational selves, inferring the preferences of others from the choices they make, or managing the impressions we make to enhance future reciprocity, the relationship between choice and preference is actively constructed, interpreted, and sometimes strategically managed.

The three lines of research each offered distinct but complementary insights. Chapter 1 brought P2 conflicts from the conceptual realm to the psychological realm: people with stronger P2 conflicts showed a greater preference for changing their desires rather than their behaviors, providing initial evidence that P2 conflicts are meaningfully preference-directed rather than merely behavior-directed. Moreover, the subtypes of P2 conflicts varied predictably across domains — food-related P2 conflicts were more strongly tied to instrumental concerns, whereas sex-related P2 conflicts were more strongly tied to moral and identity concerns, suggesting that the psychology of P2 conflicts is sensitive to the nature of the desire in question. Chapter 2 showed that people's inferences about others' preferences are shaped by a causal reasoning process that takes into account the choice context. Because default options affect choices through reasons other than preferences, they can explain away preference as the main reason for choice, leading to asymmetric inferences when people accept versus switch from a default. In support of

our account, we provide novel evidence that the asymmetry diminished when defaults no longer provided plausible alternative explanations for choice. Chapter 3 sought to clarify the causal relationship between self-control and cooperation by experimentally manipulating the role of social evaluation, which may covary with both reported self-control and cooperative behavior. We found that people strategically inflate their reported self-control under social evaluation, but only when that evaluation carries real consequences for future cooperation. This inflation did not translate into more cooperative behavior, suggesting that the association between self-control and cooperation is partly driven by reputational concerns rather than genuine disposition.

Each line of research also opens as many questions as it closes. Chapter 1 is merely a first step towards a larger project of integrating P2 conflicts into the psychological realm. Variations in these conflicts across many different domains remain unexplored, and the link between P2 conflicts and metacognition, the true self literature, and internal conflicts beg to be explored. A natural next step is to examine whether P2 conflicts and self-control conflicts can be further distinguished empirically, across the domains suggested in our pilot survey. This distinction, in turn, raises an interesting question about clinical applications: if P2 conflicts are psychologically distinct from self-control conflicts, but behaviorally conflated with them, then interventions designed to help people change their behavior may be missing the point for people who don't only want to change what they do, but also what they want. Therapy, coaching, and behavioral interventions rarely target preferences directly, and this work suggests they might benefit from doing so. More broadly, P2s are one instance of a wider class of metacognitive phenomena, ranging from akrasia, to epistemic emotions, to world beliefs – that have been richly theorized in philosophy but remain empirically underexplored in psychological domains. These questions are not confined to human cognition — as artificial intelligence systems become increasingly

capable of modeling their own goals and outputs, questions about whether they might develop something analogous to P2 conflicts become increasingly meaningful and worth exploring. Probing these metacognitive phenomena in human and in AI can be illuminating for broader questions on agency and humanhood.

We cannot ask questions about who we are without asking questions about who others are. The causal reasoning framework developed in Chapter 2 explores how we understand others from single shot choices, integrating the judgment and decision-making phenomenon related to default effects into a larger Bayesian reasoning framework. However, the robustness of this framework requires testing it across different domains, default-setters, and choice environments. For instance, in some cases, the asymmetry may not just diminish but reverse, when the default itself provides a reason to switch rather than stay. These predictions remain largely untested, and doing so would both stress-test the framework and deepen our understanding of how people read meaning into the choices of others. On a broader note, this work also forges a connection between the judgment and decision-making literature and the causal reasoning literature — suggesting that how people make choices and interpret others' choices may be governed by the same inferential principles that guide how people explain events in the physical and social world. The natural next question that falls from this connection is how do these inferential principles develop, which we explore in ongoing work.

Finally, understanding who we are and who others are leads us to shape how we appear to others, and Chapter 3 provided new insights into one facet of the large literature on signaling and functional explanations of social behavior. Most directly, strategic signaling in everyday life rarely involves people explicitly claiming self-control — it more often happens through behavior, like choosing the salad over the fries or resisting other visible temptations. Whether

observable self-control behavior is subject to the same strategic inflation as self-reported self-control, and whether it is a more or less honest signal of cooperative intent, remains an open question. It is also unclear whether these patterns generalize beyond Western, US-based samples, where cultural norms around self-control and cooperation may look quite different. More broadly, this work suggests that people care about appearing self-controlled – even in mundane matters — because they are sensitive to the reputational stakes of self-control. Future research should examine whether the same pattern emerges for other traits associated with cooperation, such as generosity or patience, and bring the same experimental rigor to disentangling reputation from disposition.

Finally, these questions suggest that the looking glass of choice has many more rooms to explore. Like Alice, the further in you go, the more questions you have. Lewis Carroll used poetry and absurdity to find some order and truth. I hope to do the same, but with psychological theories and causal manipulations.