



Flexibility is the Key to Stability: An Investigation of the Malleability of
Personality Judgements with a Focus on the Moral Domain.

Elitza Zaharieva Ambrus

Submitted in partial fulfilment of the requirements for the degree of Doctor of
Philosophy in the Department of Psychology

Royal Holloway, University of London

2023

Declaration of Authorship for Co-authored Work

If you are presenting partly co-authored work, please indicate below your individual contribution to the thesis.

Name of Candidate: Elitza Zaharieva Ambrus.....

Thesis Title: Flexibility is the Key to Stability: An Investigation of the Malleability of Personality Judgements with a Focus on the Moral Domain.....

I confirm that the thesis that I am presenting has been co-authored with: Ryan McKay, Bjoern Hartig and Petter Johansson
Within this partly co-authored work, I declare that the following contributions are entirely my own work:

(Here you should indicate, in précis style, the datasets that you gathered, interpreted and discussed; methods that you developed; complete first drafts that you wrote; content that is entirely your own work; etc. It is often appropriate to organise this statement by chapter)

Chapter 1. Introduction

I wrote the manuscript with comments from Ryan McKay.

Chapter 2. Methods

I wrote the manuscript with comments from Ryan McKay.

Chapter 3. The Effect of Socially Imbued Anchors on Choices with and without Moral Implications

The idea for the experimental design of Study 1 and the “wheel of fortune” task was developed in a discussion with Bjoern Hartig and Ryan McKay. I designed the pilot and the experiment in Qualtrics; Bjoern Hartig wrote the code on the “wheel of fortune” task mechanism. I gathered the data for the pilot (via Amazon Mechanical Turk) and the experiment (via Prolific). I wrote the R code for analysing the data. I interpreted the results with comments from Bjoern Hartig and Ryan McKay. I wrote the manuscript with comments from Ryan McKay.

Chapter 4. Self-serving Anchoring of Personality Judgements

I formulated the idea with assistance from Bjoern Hartig and Ryan McKay. I designed the experiment in Qualtrics; I gathered the data for the experiment (via Prolific). Bjoern Hartig and I wrote the R code for analysing the data. I interpreted the results with comments from Bjoern Hartig and Ryan McKay. I wrote the manuscript with comments from Bjoern Hartig and Ryan McKay.

Chapter 5. Anchoring of Personality Judgements and Its Impact on Subsequent Personality Judgements

I formulated the idea with assistance from Bjoern Hartig and Ryan McKay. I designed the pilot for Study 3 and Study 4, as well as the experiment in Qualtrics; I gathered the data for the pilot and the experiment (via Prolific). I wrote the R code for analysing the data. I interpreted the results with comments from Bjoern Hartig and Ryan McKay. I wrote the manuscript with comments from Ryan McKay.

Chapter 6. Anchoring of Personality Judgements and Its impact on Subsequent Prosocial Choices

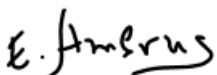
I formulated the idea with assistance from Bjoern Hartig and Ryan McKay. I designed the experiment in Qualtrics; I collected the data for the experiment (via Prolific). I wrote the R code for analysing the data. I interpreted the results with comments from Bjoern Hartig and Ryan McKay. I wrote the manuscript with comments from Ryan McKay.

Chapter 7. A Blind Spot for Flattery: People are Receptive to Enhancing Manipulations of their Personal Qualities

The idea for the experimental design was developed in a discussion with Bjoern Hartig, Ryan McKay and Petter Johansson. I designed the experiment in Qualtrics, Bjoern Hartig wrote the code for the choice blindness mechanism; I gathered the data for the experiment (via Prolific). I wrote the R code for analysing the data. I interpreted the results with comments from Bjoern Hartig and Ryan McKay. I wrote the manuscript with comments from Bjoern Hartig, Ryan McKay and Petter Johansson.

Chapter 8. Discussion

I wrote the manuscript with comments from Ryan McKay.

Signed: 

Date: 20 July 2023

(Candidate)

Signed: 

Date: 20 July 2023

(Supervisor)

This form should be signed by the candidate and the candidate's supervisor and returned to the [Doctoral School](#) when the electronic copy of the thesis is submitted Student Administration, Royal Holloway, Egham, Surrey TW20 0EX with the copies of the thesis.

Acknowledgments

I feel so grateful and truly blessed to have the support of so many wonderful people, who shared this exciting and challenging journey with me. First, I would like to thank my brilliant supervisors, Ryan McKay, Bjoern Hartig and Petter Johansson. Ryan's kindness, patience and help are beyond words – he always believed in me, letting me develop my own ideas while fully supporting my work and reminding me how fascinating doing research is. Bjoern's unique ability to see to the core of each concept and question it has been eye-opening. I am so thankful for Bjoern's challenging questions and comments as they made me contemplate each step of our work. Petter's help and expertise have been invaluable. I am so grateful to Ryan, Bjoern and Petter – I will miss our inspiring, intellectually stimulating, and fun discussions.

I am also so grateful to my wonderful family – Gabor and our lovely children, Emil, Ekaterina and Nia Ambrus – I truly could not have made it without them - they believed in me when I doubted myself. An endless source of love, inspiration, and support. Always happy to listen to and discuss my ideas, providing unique perspectives and ideas themselves. And thankfully, always happy to eat junk food if they have to.

My gratitude also goes to my kind and wise parents, Yordanka and Zahari Moravski – they have been my rock, I have still so much to learn from them. I am also so thankful to my lovely friends, Andrea Kurucz, Ayan Gorhan, Danail Slavov, Gergana Mateeva and Maria Hristova, who share the joys as well as the disappointments with me, helping me see the light and find my own voice. And special thanks to Jonathan who has always been next to me.

A big thank you to all my wonderful colleagues at the Psychology Department who make the atmosphere so warm, welcoming, and nurturing. They always provided me with helpful, constructive comments and kept reassuring me that every PhD has its ups and downs. I am especially thankful to my advisor, Jeanne Shinsky, and Shiri Lev-Ari, who chaired my annual reviews and upgrade, for their helpful and insightful suggestions.

Abstract

This thesis investigates the malleability of individuals' self-concept by extending the anchoring and choice blindness paradigms to the domain of the self. In a series of online experiments, I explore how others' behaviour and one's own (alleged) previous behaviour influence current personality judgements and decisions. Study 1 investigates whether moral choices are *more malleable* than choices in other domains in response to social anchors. Study 2 asks whether participants are especially vulnerable to *self-serving* anchors, i.e., anchors heightening participant's qualities. Studies 3 and 4 explore the potential self-serving *aftereffect* of anchoring on subsequent personality judgements (Study 3) and prosocial choices (Study 4). Study 5 investigates whether personality judgements are susceptible to choice blindness manipulations, especially when the manipulations elevate the self-view. Throughout the studies, I contrast moral and non-moral attitudes to explore whether moral behaviours and personality judgements are more susceptible to cognitive influences.

The main conclusion from the present thesis is that personality judgements are flexible in response to cognitive influences in a self-serving manner: personality judgements seem flexible enough to accommodate adjustments elevating the self-image, however they remain relatively stable in the face of diminishing manipulations. Although, there was no unanimous evidence that self-serving manipulations of personality judgements influence the general self-image, enhancing anchors led to nearly 15% more generous donations in a subsequent Dictator Game. The analysis did not support magnified anchoring or choice blindness effects for moral traits, rather morality had a general elevating effect with individuals ranking themselves more positively on moral than on non-moral traits. The data also provided evidence for a "phrasing effect" with participants ranking themselves higher, on average, for negatively than positively phrased traits. These findings suggest that personality judgements are constructed and adjusted in a somewhat different way than previously thought. Implications for the anchoring and choice blindness frameworks are also discussed.

Table of Contents

Table of Contents

Flexibility is the Key to Stability: An Investigation of the Malleability of Personality Judgements with a Focus on the Moral Domain.	1
Declaration of Authorship for Co-authored Work.....	2
Acknowledgments.....	5
Abstract.....	6
Table of Contents.....	7
Table of Tables.....	12
Table of Figures.....	14
Chapter 1. Introduction.....	16
Introduction.....	1
Trajectory of the Research.....	3
Theoretical Context of the Research.....	4
Flexibility of choices (Study 1).....	4
Flexibility of personality judgements (Studies 2 to 5).....	5
Research aims for each study.....	8
Chapter 2. Methodology.....	17
Methodology.....	1
Flexibility of choices and contrasting moral and non-moral choices. Study 1.	1
Flexibility of personality judgements. Contrasting moral and non-moral personality judgements. Studies 2 to 5.....	4
Psychological paradigm employed to test the malleability of choices and personality judgements: anchoring (Studies 1 to 4).....	5
Psychological paradigm employed to test the malleability of personality judgements: choice blindness (Study 5).....	6
Experimental design and sampling.....	9
Measuring choices and personality judgements.....	13
Pre-registration.....	15
Methods summary.....	16

Chapter 3. Study 1. The Effect of Socially Imbued Anchors on Choices with and without Moral Implications	18
Study 1. The Effect of Socially Imbued Anchors on Choices with and without Moral Implications	1
Abstract.....	2
The Effect of Socially Imbued Anchors on Choices with and without Moral Implications.....	3
Theoretical framework	3
Method	9
Participants	9
Materials and Procedure	10
Design and Analysis.....	10
Results.....	11
Discussion.....	14
References	17
Appendix 3A. Pilot study.....	24
Method	24
Participants	24
Materials and Procedure	24
Design and Analysis.....	25
Results & Discussion	25
Appendix 3B. Wheel of fortune task.....	27
Instructions for one of the four experimental conditions in Study 1 (risk-for-charity with anchoring)	27
Risk-for-charity task	30
Chapter 4. Study 2. Self-serving anchoring of self-judgements	19
Study 2. Self-serving Anchoring of Personality Judgements	1
Abstract.....	2
1. Introduction	3
2. Method	7
2.1. Overview.....	7
2.2. Participants.....	8
2.3. Materials and Procedure	8
2.4. Design and Analysis	9

3. Results.....	10
Anchoring influenced judgements of participants' own personal qualities.....	10
Anchoring influenced self-judgements in a self-serving manner, though this was not more pronounced for moral traits	11
4. Discussion.....	13
References	19
Appendix A. Pre-registered ANOVA analysis	31
Chapter 5. Study 3. Anchoring of Personality Judgements and Its impact on Subsequent Personality Judgements.....	20
Study 3. Anchoring of Personality Judgements and Its impact on Subsequent Personality Judgements	1
Abstract.....	2
Anchoring of Personality Judgements and Its impact on Subsequent Personality Judgements.....	3
Method	5
Overview	5
Participants	6
Materials and Procedure	6
Design and Analysis.....	7
Results.....	8
Anchoring at the first stage of the experiment	8
Anchoring <i>aftereffect</i> at the second stage of the experiment	9
Exploratory analysis. The effect of the way the personality traits are phrased (positively or negatively).....	11
Discussion.....	13
References	17
Appendix 5A. Pilot of Study 3 and Study 4	24
Method	24
Overview	24
Participants	25
Materials and Procedure	25
Design and Analysis.....	26
Appendix 5B. Anchoring task.....	27
Appendix 5C. Subsequent self-rankings	29

Chapter 6. Study 4. Anchoring of Personality Judgements and Its impact on Subsequent Prosocial Choices	21
Study 4. Anchoring of Personality Judgements and Its impact on Subsequent Prosocial Choices	1
Abstract.....	2
Anchoring of Personality Judgements and Its impact on Subsequent Prosocial Choices	3
Method	5
Overview	5
Participants	5
Materials and Procedure	6
Design and Analysis.....	7
Results.....	8
Anchoring at the first stage of the experiment	8
Anchoring aftereffect at the second stage of the experiment	9
Exploratory analysis. The effect of the way the personality traits are phrased (positively or negatively).....	10
Discussion.....	12
References	17
Appendix 6A. Dictator Game	24
Chapter 7. Study 5. A Blind Spot for Flattery: People are More Receptive to Enhancing than Diminishing Manipulations of their Personal Qualities	22
Study 5. A Blind Spot for Flattery: People are More Receptive to Enhancing than Diminishing Manipulations of their Personal Qualities	1
Abstract.....	2
A Blind Spot for Flattery: People are More Receptive to Enhancing than Diminishing Manipulations of their Personal Qualities	3
<i>Self-consistency and Self-enhancement of Personality Judgements</i>	3
<i>The Choice Blindness Paradigm</i>	5
<i>The Present Study</i>	6
Method	7
Overview	7
Participants	7
Materials and Procedure.....	8
Results.....	11

<i>Design and Analysis</i>	11
<i>Main Hypotheses</i>	12
<i>Result 1: The Manipulation Is Effective</i>	12
<i>Result 2: Enhancing Manipulations Are Stronger Than Diminishing Manipulations.</i>	15
<i>Secondary and Exploratory Hypotheses</i>	17
Discussion.....	18
<i>Summary and conclusions</i>	20
References	21
Supplementary Material	33
A. Screenshots from the experiment.....	33
B. Analysis of the original personality judgements at the first stage of the experiment.....	36
C. Analysis excluding participants, who detected the manipulation.....	38
D. Analysis, including desirability and type of manipulation as fixed effects.	40
E. Result 2, total sample.....	41
F. Analysis, including the initial self-rankings as predictor of the revised self-rankings.	42
G. Secondary and Exploratory Hypotheses	45
Chapter 8. Discussion.....	23
Discussion.....	1
Summary of the main results.....	1
Critical evaluation of findings: strengths and limitations	3
Implications and future directions.....	17
Summary and conclusions	21
References	23

Table of Tables

	Page
Chapter 4 (starts on page 19 of the thesis)	
Table 1. Estimated fixed effects for Model 1 (depicting the effect of anchors on self-rankings).	10
Table 2. Estimated fixed effects for Models 2, 3 and 4 (depicting the effect of <i>enhancing</i> and <i>diminishing</i> anchors, <i>morality</i> and their interaction on self-rankings).	12
Chapter 5 (starts on page 20 of the thesis)	
Table 1. Estimated fixed effects for personality judgements at the first stage of the study.	9
Table 2. Estimated fixed effects of <i>enhancing</i> , <i>diminishing</i> , <i>morality</i> and their interaction on personality judgements at the second stage of the experiment.	10
Table 3. Means and Standard Deviations for self-rankings at the second stage of the study.	11
Table 4. Model 5. Estimated linear model for the self-rankings indicated at the first stage of the study	13
Chapter 6 (starts on page 21 of the thesis)	
Table 1. Estimated model coefficients from the linear regression on the self-rankings measure.	9
Table 2. T-tests comparing Dictator Game donations across anchoring conditions	10

Table 3. Estimated linear model for personality judgements by anchor and desirability.	11
Chapter 7 (starts on page 22 of the thesis)	
Table 1. Estimated fixed effects for the difference between revised and original self-rankings.	14
Table 2. Estimated fixed effects for enhancing, morality, and their interaction with respect to the <i>magnitudinal difference</i> .	17

Table of Figures

	Page
Chapter 3 (starts on page 18 of the thesis)	
<i>Figure 1.</i> Histograms of the number of spins chosen for each type of task in the no anchor (left panel) and anchor (right panel) by condition.	11
<i>Figure 2.</i> Violin plots for the number of spins chosen by condition (anchor value = 92 spins). The point range represents 2SD around the mean.	12
<i>Figure 3.</i> Effect of anchoring on number of spins in the personal risk and risk-for-charity task.	13
Chapter 4 (starts on page 19 of the thesis)	
Figure 1. Violin plots of self-rankings by <i>anchor</i> . Error bars represent 95% confidence intervals (CI).	11
Figure 2. Violin plots of self-rankings by <i>anchor</i> . Error bars represent 95% confidence intervals (CI).	13
Figure 3. Self-rankings by <i>anchor</i> and <i>morality</i> . Error bars represent 95% confidence intervals (CI).	13
Chapter 5 (starts on page 20 of the thesis)	
Figure 1. Violin plots of the self-rankings data by anchor conditions.	9
Figure 2. Bar plot of self-rankings by anchoring conditions	13

Chapter 6 (starts on page 21 of the thesis)	
Figure 1. Violin plots for the self-rankings measure by anchor condition.	9
Figure 2. Bar plot of self-rankings by anchoring conditions	12
Chapter 7 (starts on page 22 of the thesis)	
Figure 1. Histograms of the difference between revised and original self-rankings for non-manipulated and manipulated (enhanced or diminished) traits, total sample (Panel A) and middle zone (Panel B).	15
Figure 2. Histogram for the <i>magnitudinal difference</i> of enhancing and diminishing manipulations, middle zone (Panel A) and bar chart for the <i>magnitudinal difference</i> for enhancing and diminishing manipulations, middle zone (Panel B).	16

Chapter 1. Introduction

Introduction

“Know thyself” is the Socratic dictum. But to what degree do we *know* ourselves? Psychological research indicates that our self-evaluations and decisions are shaped and influenced by a range of external and internal factors and motivations (Alicke & Govorun, 2005). In this thesis, I investigate the malleability of individuals’ self-concept by extending prominent cognitive paradigms – in particular, anchoring and choice blindness – to the domain of the self. Research has shown that various cognitive biases affect human memory, reasoning, judgements and decisions (Tversky & Kahneman, 1974; Acciarini et al., 2021). Investigating cognitive influences in the domain of the self allows me to integrate two essential influences of the world in and around us in the experimental design: individuals’ perceptions of who they are and others’ behaviour. I also test for any differential effects in the moral domain and explore whether manipulations of participants’ personality judgements affect their general self-image and subsequent decisions.

Knowing ourselves has been promoted as a way of achieving life fulfilment since ancient times and intuitively suggests an exploration within ourselves to discover our innate, true essence. Research has shown that individuals believe in the notion of a “true self” (Newman et al., 2014) and use it as explanation in a variety of contexts, ranging from attributions about behaviour (Landau et al., 2011) to decision satisfaction (Schlegel et al., 2013). Such a concept ensures both stable preferences across contexts and time as well as stable personality judgements about ourselves. Indeed, preferences have commonly been considered stable in decision-making (Von Neumann & Morgenstern, 1947). Nevertheless, research has also shown evidence for flexible, reference-dependent preferences that vary over contexts and time (e.g., Ariely et al., 2006). Instead of stemming from stable attitudes, behaviour might be flexible, only resembling stability by consistently complying with social norms and with individuals’ own previous behaviour (Ariely & Norton, 2008; Ellemers et al., 2013; Kimbrough & Vostroknutov, 2016; Krupka & Weber, 2013).

In the domain of the self, personality characteristics are also traditionally considered stable in adulthood and serve as a valid predictor of life outcomes (e.g., Ozer & Benet-Martinez, 2006). Nevertheless, research has shown that personality traits continue to develop throughout the lifespan (e.g., Bleidorn et al., 2022). Recent research also demonstrates flexibility of personality judgements in response to nonclinical psychological interventions (e.g., Stieger et al., 2020). Despite some degree of flexibility, however, personality traits are still considered stable and consistent behavioural patterns across contexts (Roberts, 2009), which should render them resilient in the face of cognitive influences.

At the same time, constructing and maintaining a positive self-image is central for individuals' wellbeing (Alicke et al., 2013; Leary, 2007). Taylor and Brown (1988) argued that positive illusions, such as unrealistically positive self-evaluations, illusion of control and unrealistic optimism are essential for optimal mental health and wellbeing. Indeed, people seem to have an intrinsic desire to enhance their self-image (Tesser, 1988) and are prone to holding overly optimistic beliefs about their performance and abilities (the "better-than-average" effect, Alicke & Govorun, 2005). Individuals employ a wide range of psychological mechanisms to ensure they maintain such a stable and positive self-image (e.g., Möbius et al., 2022). For instance, autobiographical memory seems to function in a self-serving manner with positive (vs. negative) self-relevant information being easier to recall (e.g., Ritchie et al., 2017). Individuals also actively search for positive self-relevant feedback, updating their beliefs accordingly while avoiding and ignoring negative self-relevant feedback (Alicke & Sedikides, 2009; Gaertner et al., 2012; Sedikides & Strube, 1997; Zhang et al., 2018). Therefore, information processing seems to be shaped by self-serving psychological mechanisms when the self-view is affected.

I explore the interplay between intrinsic motivations to maintain a positive self-image and external influences such as anchoring and choice blindness manipulations on individuals' self-evaluations. Specifically, when anchoring personality judgements, I account for the way self-enhancement motives might interfere with the effect of the

anchor. In a similar way I consider the potential influence of positive self-image concerns when investigating the effect of a choice blindness manipulation on personality judgements. Given the importance of maintaining a positive self-image for wellbeing as well as the self-serving construction and updating of self-beliefs (Möbius et al., 2022; Taylor & Brown, 1988), I expect that personality judgements will exhibit a different kind of flexibility than that currently theorised (Roberts & Yoon, 2022). Specifically, I expect that personality judgements will readily adjust in response to manipulations that enhance the self-image but will remain rigid in response to diminishing manipulations.

Trajectory of the Research

To capture the flexibility of attitudes in response to others' behaviour, I first focussed on choices and their susceptibility to anchoring (Tversky & Kahneman, 1974), integrating social information with the anchor value. Based on literature showing the importance of morality to the self (e.g., Strohminger, 2018) and the observed exaggerated cognitive effects in the moral domain (e.g., Meyers et al., 2019), I contrasted tasks in the moral and non-moral domain, utilising the moral task as a lens for exploring the factors determining the flexibility of choices. I asked whether social anchoring (the anchor value was presented as a benchmark of others' behaviour) differentially affects moral and non-moral choices. Designing and conducting Study 1, however, brought home the difficulty of constructing a comparison of choices that differ only in the sense that one is in the moral and the other in the non-moral domain.

Subsequently, I turned to the domain of the self (Studies 2 through 5), which allowed me to: (i) explore the malleability of the self-concept; (ii) distinguish consideration of maintaining a stable and positive self-image from the pure anchoring or choice blindness effects (via enhancing and diminishing manipulations) as well as (iii) contrasting the susceptibility of moral and non-moral personality judgements to

anchoring and choice blindness manipulations. As the importance of constructing and maintaining a positive self-image (e.g., Leary 2007) is an intrinsic feature of the domain of the self, I focus on the potential effect of positive self-image on individuals' susceptibility to cognitive influences, such as anchoring (Studies 2 to 4) and choice blindness (Study 5).

Theoretical Context of the Research

Flexibility of choices (Study 1)

The classical view of human behaviour is that judgements and choices are stable and consistent over contexts and time (Von Neumann & Morgenstern, 1947). Although theoretical models allow for psychological factors such as relative income (Fehr & Schmidt, 1999) or identity (Akerlof & Kranton, 2000) to influence behaviour, the underlying assumption of stable attitudes is retained. However, ample research evidence shows individuals' preferences are reference dependent (Ariely et al. 2003; 2006; Kahneman & Tversky, 1979; Kőszegi & Rabin. 2006; 2007) with judgements and choices varying not only over task contexts, but also over time (Hoffman et al., 1996; List 2007; Mullen & Monin, 2016; Payne et al., 1992; Slovic, 1995).

Judgements and choices in the moral domain also vary over task contexts (e.g., List 2007) and over time, due to the so-called "moral licensing" (Mullen & Monin, 2016). For example, committing to a moral act in the future (e.g., donating blood) is associated with displaying more racial bias in a current decision-making task (Cascio & Plant, 2015). The size of the moral licensing effect depends on the cultural background (Simbrunner & Schlegelmilch, 2017) and is slightly smaller than other effects in social psychology, yet meta-analysis has shown that the effect persists across task contexts (Blanken et al., 2015, cf. Blanken et al., 2014). Contextual social norms (Ellemers et al., 2013; Krupka & Weber, 2013), social recognition and self-esteem (Heintz et al., 2016) also influence behaviour. Thus, rather than evincing stability, behaviour might only resemble stability

while being flexible in response to the task context, perceived social norms, others' behaviour as well as one's own previous or anticipated future behaviour.

Theoreticians have claimed that attitudes in the moral (vs. non-moral) domain are more sensitive to the influence of psychological factors (Greene & Haidt, 2002; Haidt 2001; cf. Cushman & Young, 2011; Rai & Holyoak, 2010). Indeed, moral values are ranked persistently at the top of individuals' value hierarchies cross culturally (Schwartz & Bardi, 2001; Schwartz & Cieciuch, 2022) and participants' attitudes are motivated by the values indicated as important for them (Sagiv & Roccas, 2021; Sagiv et al., 2017). Moral issues are also strongly associated with emotional engagement which in turns influences behaviour (Greene et al., 2001). Moral violations elicit stronger emotional and behavioural responses than violations of non-moral conventions (Rozin et al. 1997; Rozin, 1999) and some cognitive effects are more pronounced in the moral domain (Brown, 2012; Meyers et al., 2019; Tappin & McKay, 2017). For example, although individuals are generally prone to keep on investing in a futile course of action (the so called "sunk cost effect"), the effect is stronger for moral tasks (Meyers et al., 2019). Based on the intrinsic high sensitivity of moral behaviours, in Study 1 I investigated the flexibility of choices through the lens of morality, asking whether moral choices are more susceptible to social anchoring than non-moral choices are.

Flexibility of personality judgements (Studies 2 to 5)

In Study 2 to Study 5, I explored the flexibility of personality judgements. The field of personality psychology is currently undergoing a stage of rapid development (Roberts & Yoon, 2022). The commonly accepted view on personality traits was that they are fixed and stable in adulthood (Costa & McCrae, 1992; McCrae et al., 2000). In the past two decades, however, meta-analytic studies have challenged this view and determined that certain personality traits are subject to change, albeit slowly, in adulthood too (Ferguson et al., 2010; Roberts et al., 2006). A recent meta-analysis (Bleidorn et al., 2022) reinforces this complex pattern of personality traits' development

over time, showing that while some personality traits exhibit relative stability in middle adulthood, others, e.g., “emotional stability”, continue to change with maturity. That personality traits are adaptable during middle adulthood too is important because of the link between personality traits and life outcomes: personality judgements are a valid predictor of life outcomes, such as work performance, health, and well-being (Beck & Jackson, 2022; Roberts et al., 2007; Soto, 2021). If personality traits are flexible, they can be successfully targeted by interventions to improve well-being, which is a goal embraced by both researchers and policymakers (Bleidorn et al., 2019; Chapman et al., 2014; Mroczek, 2014; OECD, 2020). Recent research efforts have followed this line of investigation, providing empirical evidence that personality traits can be changed via nonclinical psychological interventions and that this change may be lasting (Bleidorn et al., 2021; Olaru et al., 2022; Stieger et al., 2020, see Allemand & Flückiger, 2022 for a review).

Yet, even if current theories allow for some degree of malleability, personality traits are considered as stable behaviour patterns over contexts that could be used to predict behaviour (Roberts & Yoon, 2022; Soto et al., 2021). Dweck (2017) suggests a somewhat different theoretical approach to the construction of personality traits: there are three basic needs at the core of personality (competence, predictability and acceptance) and personality traits are constructed and adjusted to achieve these goals across different contexts and over time. Dweck’s (2017) theory implies a certain degree of flexibility of personality traits and resonates with a different line of research placing individuals’ goal of construing their lives in a positive and sense-making manner at the core of attitudes’ formation, which suggests adaptable behaviour over contexts and time (Chater & Loewenstein, 2016).

Furthermore, instead of being stable, behaviour might be constructed on the fly, drawing on contextual cues and aiming at ensuring consistency with perceived previous attitudes (Johansson et al., 2012). Research has shown that individuals are able to provide justification for their behaviour even when they are not aware of the exact reason that triggered it (Gazzaniga, 2000). Similarly, participants were ready to embrace

and justify judgements and choices that they were led to believe they had made (e.g., Hall et al., 2012). Therefore, judgements and choices might be flexible, allowing instant adjustments to maintain consistency with one's own previous behaviour and/or others' behaviour.

In the personality domain, an important goal outlined by research is to construct and maintain a stable and positive self-image (e.g., Leary, 2007). In Study 2 and Study 5, I explore the possibility that personality judgements are constructed and adjusted in line with a goal of maintaining a stable and positive self-image, that is, that personality judgements are flexible enough to be adjusted in response to a cognitive influence as long as the goal of maintaining a positive self-image is achieved. Such a finding would reveal much more flexibility in the construction and adjustment of personality traits than is currently theorised to be the case.

The potential goal of maintaining a positive self-image naturally links the construction and flexibility of personality judgements to literature on the stability of individuals' self-image. Traditionally, philosophical and social psychological research has outlined autobiographical memory as an important factor for self-identity as it provides both a narrative that unifies individuals' sense of self as well as ensures individuals' distinctiveness in social and group identity (Strohming, 2018). Self-image is thus considered stable in adulthood as self-defining memories peak in early adulthood (Rathbone et al., 2008).

Research in the last decade also demonstrates the central role of morality in shaping individuals' self-concept and perceptions of self-continuity (Heiphetz et al., 2016; Molouki & Bartels, 2017; Stanley et al., 2019; Strohming & Nichols, 2014). For instance, participants indicated moral traits as more likely to be carried with a soul when it switches bodies or is reincarnated for another life (Strohming & Nichols, 2014). While individuals are averse to negative changes in moral traits, positive changes in moral traits are accepted as a natural development of the fundamentally good values humans possess (Molouki & Bartels, 2017; Newmann et al., 2014; Newmann et al.,

2015). Moral traits, such as “honest” and “compassionate” are also perceived to be the most important in impression formation (Goodwin et al., 2014). Therefore, constructing and maintaining a positive moral self-concept is of paramount importance to us (Alicke et al., 2013; Leary, 2007).

Indeed, individuals seem to employ a wide range of selective psychological mechanisms to ensure that their self-concept is stable and positive: participants commit to memory and more easily recall positive rather than negative behaviours (Carlson et al., 2020; Ritchie et al., 2017; Sedikides & Green, 2009; Sedikides et al., 2016; Stanley et al., 2019; Zhang et al., 2018) as well as attribute previous unethical behaviour to contextual factors (Malle et al., 2006). Gaertner et al. (2012) showed that participants actively search for positive social feedback and avoid negative social feedback. Even if negative social feedback is received, individuals seem to neglect it while positive feedback is embraced and overweighted (Eil & Rao, 2011; Korn et al., 2012; Möbius et al., 2022). These psychological mechanisms ensure individuals can sustain a positive self-view despite conflicting evidence. Taken together, these lines of evidence point towards a self-serving element in individuals’ susceptibility to cognitive biases that affect the self-image.

Research aims for each study

Study 1 draws on the robustness of the anchoring phenomenon (e.g., Röseler & Schütz, 2022) and the central role of morality to the self-concept (Strohming, 2018) to explore whether risky choices are flexible in response to social anchoring. The study relies on a task especially designed to be devoid of expectations of what constitutes a prosocial/selfish behaviour (“wheel of fortune” task, please see Appendix 3B, Chapter 3) with the anchor value described as the number of spins chosen on average in a previous round of the study. The experiment also contrasts choices with and without moral

implications to test whether choices with moral implications are more flexible in response to social anchoring.

Study 2 asks whether anchoring affects personality judgements and whether individuals are especially vulnerable to self-serving anchors, i.e., anchors that enhance individuals' self-image. Previous research shows that judgements of self-relevant information, such as judgements of recent behaviour (Cheek et al., 2015) and judgements of future prospects (Joel et al., 2017) are susceptible to anchoring. For instance, Joel et al. (2017) found that motivated reasoning renders anchors suggesting a high probability of undesirable outcomes ineffective. In Study 2, I explore whether even the most intimate judgements, judgements about our own personal qualities, are prone to anchoring. Based on the importance of maintaining a positive self-view (Alicke et al., 2013), I also investigate whether personality judgements exhibit more flexibility in response to enhancing rather than diminishing manipulations.

In Study 3 and Study 4, I explore whether anchoring of personality judgements on moral traits would affect the general self-view measured by subsequently indicated personality judgements (Study 3) or prosocial choices (Study 4). The two studies have an identical first stage; participants report their personality judgements on two traits and are anchored either in an enhancing or diminishing direction (or not anchored, the control group). Based on evidence that moral traits are essential for the self-concept (e.g., Strohminger & Nichols, 2014) and that manipulations of moral personality judgements might affect the overall self-image, I chose two moral traits for the first stage of Study 3 and Study 4 (honest and considerate or dishonest and inconsiderate). Subsequently, in Study 3, I collect data on eight more personality traits (moral and non-moral, desirable and undesirable) to assess potential changes in the general self-view. Specifically, I ask whether enhanced moral personality judgements would have an elevating aftereffect on subsequent personality judgements.

Study 4 focuses on the potential aftereffect of anchored moral self-rankings on prosocial choices. Previous research has linked personality judgements on moral traits,

such as the traits in the honesty-humility dimension of the HEXACO model of personality (Ashton & Lee 2020, Thalmayer & Saucier 2014), with prosocial behaviour (Hilbig et al., 2013; Thielmann et al., 2020; Zettler et al., 2020). There is also empirical evidence associating prosociality and honesty (Isler & Gächter, 2022; Soraperra et al., 2019). Extending previous research, in Study 4, I ask whether enhancing moral personality judgements would lead to more generous donations in a Dictator game.

In Study 5, I test whether the motivation to be consistent with one's previous behaviour affects the flexibility of personality judgements. I rely on a different cognitive phenomenon, the choice blindness framework (Johansson et al., 2005), which allows distinguishing between self-consistency and self-enhancing behaviour motivation. As discussed above, participants might adjust their attitudes to align with their (alleged) previous attitudes (Johansson et al., 2012), i.e., individuals might be flexible to achieve their goal of behaving in a "sense-making" manner (Chater & Loewenstein, 2016), which would result in flexible personality judgements, adjusted in a way to ensure individuals' self-concept is kept stable and consistent. As a drive for self-consistency might also interact with the desire to maintain a positive self-image (Alicke et al., 2013), in Study 5, I ask whether enhancing choice blindness manipulations will be accepted to a higher degree than diminishing manipulations.

Chapter 2. Methodology

Methodology

Studies 1 to 4 draw on key paradigms from decision theory (anchoring, Tversky & Kahneman, 1974), social psychology (the illusion of moral superiority; Tappin & McKay, 2017) and experimental economics (Dictator Game; Kahneman et al., 1986) while Study 5 utilises another key decision theory framework, the choice blindness paradigm (Johansson et al., 2005). All the studies in this thesis share an overarching framework: in each study, I expose choices (Study 1) or personality judgements (Studies 2 to 5) to a certain cognitive influence (anchoring or choice blindness) and explore whether: (i) personality judgements or choices are susceptible to the respective cognitive influence; (ii) the employed cognitive paradigm influences personality judgements in a self-serving manner and (iii) there is a differential effect of the cognitive influences for attitudes in the moral vs non-moral domain.

Flexibility of choices and contrasting moral and non-moral choices. Study 1.

Study 1 focuses on the flexibility of individuals' choices. To contrast moral and non-moral preferences, I needed to employ two tasks, identical except for the respective decision-making domain they concern - moral or non-moral. Constructing such tasks however, turned out to be very challenging. For example, Tassy and colleagues (2013) compared the discrepancy between judgements and choices in moral and non-moral dilemmas; the moral dilemmas involved a choice between abandoning one sailor to save four sailors while in the non-moral dilemma participants chose between spending a 25% discount coupon that expires today and spending a 30% coupon that expires in a year's time. Although the non-moral dilemmas employed were suitable for the purpose of the experiment, they do not map directly into the moral dilemmas (Tassy et al., 2013). Moral dilemmas involve human lives, which might evoke a much stronger emotional response (Greene & Haidt, 2002) than a financial decision about spending discount coupons, which might in turn influenced behaviour in a way that is difficult to disentangle from the moral vs non-moral comparison.

Indeed, a prevailing theoretical approach towards moral preferences is to view them as moral intuitions that are shaped by implicit affective psychological processes (Ditto et al., 2009; Haidt, 2001; Haidt, 2007). Moral Foundations Theory (Graham et al., 2013; Graham et al., 2018; Haidt, 2001; Haidt, 2007) defined five core moral foundations stemming from moral intuitions: care/harm, fairness/cheating, loyalty/betrayal, authority/subversion, and sanctity/degradation. The current Moral Foundations Theory conceptualisation also involves liberty/oppression as a sixth foundation (Iyer et al., 2012). The Moral Foundations Theory premises have recently been validated in a study drawing on a cross-cultural dataset of 30 diverse societies (Doğruyol et al., 2019). A different line of research, conceptualising morality by its function to promote cooperation, suggests a seven-factor model of morality based on family, group, reciprocity, heroism, deference, fairness, and property (Morality-as-Cooperation, Curry et al., 2019). I have chosen to explore “fairness” in Study 1 as a factor, representing essential moral values according to both theoretical accounts (Haidt, 2001; Curry et al., 2019).

Developing a valid experimental design employing two tasks such that the only difference between them is that one of them concerns the fairness moral value, however, had to also take into account that the notion of fairness evokes corresponding perceptions of appropriate social norms (Andreoni & Bernheim, 2009). For instance, when studying social preferences, researchers have commonly found that responses in allocation tasks vary depending on the contextual social norms and tend to cluster around one of two modes: sharing nothing (selfish behaviour) or cooperating by sharing half of the initial endowment (Camerer, 2003). Hence, any task involving fairness concerns might evoke subjective perceptions of social norms, which in turn will influence behaviour (Krupka & Weber, 2013). Therefore, to contrast decisions concerning fairness with a non-moral decision, in a valid experimental design ensuring sufficient individual variation, I needed to employ a task, devoid from contextual social norms and a single interpretation of what constitutes a “fair” behaviour.

To address the above methodological challenge in studying the malleability of moral preferences, my supervisors and I developed a novel task (the “wheel of fortune task”), with a range of potential choices that could not easily be labelled as selfish or prosocial (Appendix 3B). The “wheel of fortune” task meets all the above discussed requirements for ensuring a valid experimental design, however, the task could be classified as having moral implications rather than being a typical moral choice (details are discussed below, please see also Appendix 3B). In brief, participants are provided with an initial endowment (10p), and they can choose how many times a computer will spin a wheel of fortune on their behalf. The wheel of fortune has ninety-nine “good” spaces and one “bad” space. If a spin lands on a “good” space, participants win an extra 2p for themselves. If a spin lands on a “bad” space, the game is over, and all the money accumulated (including the 10p starting endowment) is lost. In the condition with moral implications, participants play on behalf of a charity and the amount won (if any) is to be transferred to a charity of the participant’s choice.

Due to the probabilistic nature of the decision, there is no straightforward interpretation of selfish or prosocial behaviour (and corresponding number of spins) when the task is played on behalf of a charity. Hence, individuals’ responses in the wheel of fortune task should not exhibit any clusters corresponding to perceived social norms. A pilot study (see Appendix 3A for details) confirmed that there is no clustering of responses in the wheel of fortune task, which makes it suitable to contrast choices with and without moral implications. Thus, Study 1 employed the wheel of fortune task to explore the malleability of individuals’ decisions and specifically to investigate whether choices with moral implications are more flexible than choices without moral implications in response to social anchoring.

Flexibility of personality judgements. Contrasting moral and non-moral personality judgements. Studies 2 to 5

The results of Study 1 suggested that the “wheel of fortune” task might not have managed to capture the essence of moral choices and thus provide an informative base for comparison between moral and non-moral choices (please see Chapter 3 for discussion). Hence, I continued searching for a suitable task that could contrast moral and non-moral preferences in a way that is both experimentally valid while the moral task represents a typical moral choice or judgment. After discussion with my supervisors, I decided to continue my exploration by measuring self-perceptions. The domain of the self has the advantage that the distinction between moral and non-moral personality traits could be very clear if the personality traits are carefully chosen (e.g., “kind” vs “intelligent”) while self-rankings on both moral and non-moral personality traits could be indicated on the same measurement scale (0-100).

A potential caveat of employing a measurement scale for indicating personality judgements could be that participants are not aware what the minimum and the maximum of the measurement scale expresses, which might influence their personality judgements. For example, if one is providing a self-rating for honesty, it is not clear what the maximum of the scale stands for - locally impressive or historically impressive examples of honesty. Indeed, research has long shown that the standard for comparison influences personality judgements; for example, participants’ evaluations of how competent or athletic they are, depend on the salient other that they compare with (Dunning & Hayes, 1996; Gilbert et al., 1995; Morse & Gergen, 1970). To provide clarity on the benchmarks for comparison, our instructions in Studies 2 to 4 asked participants to report their personality judgements relative to 100 other (anonymous) participants. Participants were debriefed at the end of each experiment that although there were more than 100 participants in the experiment, they were told to compare to 100 others so as to facilitate their understanding of the comparison in question. To avoid even this mild type of deception, in Study 5, I asked respondents to indicate their self-rankings

relative to “a large number” of other participants and the measurement scale ranged from “less than others” to “more than others”.

To investigate whether moral attitudes are more flexible than preferences in the non-moral domain, I needed to expose both types of choices (with and without moral implications in Study 1) and personality judgements (Studies 2 to 5), to the same psychological phenomenon and measure their respective susceptibility. I employed the anchoring phenomenon in Studies 1 to 4 due to its robustness and replicability (Röseler & Schütz, 2022). In Study 5, I utilised the choice blindness paradigm as it offers a unique framework for exploring the psychological motives shaping behaviour (Johansson et al., 2005). In the choice blindness paradigm, participants’ choices or judgements are swapped inconspicuously. If respondents accept the manipulation and justify their alleged attitudes, their behaviour reveals flexible attitudes that might be guided by a wish to be consistent with perceived previous behaviour (Hall et al., 2013; Johansson et al., 2005; Strandberg et al., 2020). Extending both the choice blindness paradigm and anchoring to the domain of the self is novel and has the potential to provide additional insight not only about the way personality judgements are constructed and adjusted, but also about the psychological mechanisms triggered by the anchoring and choice blindness frameworks.

Psychological paradigm employed to test the malleability of choices and personality judgements: anchoring (Studies 1 to 4)

The anchoring effect refers to the impact of a previously considered comparative value (anchor) on a subsequent absolute judgment and reveals that individuals’ estimates of various quantities are flexible and can be influenced by a salient random value (Tversky & Kahneman, 1974). The typical anchoring framework involves two stages and was introduced in the seminal work of Tversky and Kahneman (1974). These researchers asked participants whether their estimate of a given quantity (the percentage of African countries in the UN) was lower or higher than a randomly

generated number (the outcome of a wheel of fortune). At the second stage of the experiment, respondents had to provide their own estimate of the same quantity.

Study 1 utilised a modified version of the typical anchoring framework: there was no comparative question, instead, respondents indicated their choices on a slider line and the slider cursor was set at the anchor value, i.e., participants had to manually adjust the slider cursor away from the anchor value to indicate their self-rankings. Studies 2 to 4 employed the standard anchoring paradigm, i.e., the subjective judgment indicated on the slider line was preceded by answering a comparative question.

In Study 1, I used 92 as an anchor value while in Studies 2 to 4, the anchors employed were 95 for the high anchor and 5 for the low anchor (I utilised a measurement scale from 0 to 100 for all studies). Previous research has shown that average judgements for desirable moral traits are approximately 84 out of 100 while average judgements for undesirable moral traits are around 28 out of 100 (Tappin & McKay, 2017). I chose 92 in Study 1, and 95 and 5 for the high and low anchors in Studies 2 to 4, to ensure that the anchor values would be perceived as high and low values respectively, which was essential for the experimental design. In Study 1 the anchor value was presented as the average performance (spins chosen) in a previous round of the experiment, integrating social information to the anchor value (there was no previous round of the experiment, however participants were fully debriefed about the deception at the end of the experiment and had the opportunity to withdraw from the study).

Psychological paradigm employed to test the malleability of personality judgements: choice blindness (Study 5)

The choice blindness paradigm is also a robust cognitive phenomenon, replicated in various decision-making domains, such as eye-witness testimony (Sagana et al., 2016), financial decisions (McLaughlin & Somerville, 2013), and political and moral views (Hall et al., 2013; Hall et al., 2013; Strandberg et al., 2020). In Study 5, I employed a modified

version of the choice blindness framework to explore the malleability of personality judgements. In the choice blindness framework, choices and judgements are swapped unbeknownst to the participants, who are then asked to justify their alleged behaviours. For instance, in the seminal study, participants had to indicate the more attractive of two facial images (Johansson et al., 2005). The chosen image was afterwards ostensibly handed to the participant; however, the experimenter swapped the two images while handing them over. Subsequently, participants were asked why they selected this image (their alleged choice) and 74% of the participants went on to eloquently justify a choice they had not actually made (Johansson et al., 2005).

Importantly, the choice blindness paradigm commonly relies on a face-to-face interaction between the experimenter and the participants, which rendered conducting the study challenging in the pandemic environment. Initially, I also planned to collect data in person, however, I came to realise that this would not be feasible and started searching for a way to modify the framework so as to use it in an online environment. As one of the creators of the choice blindness paradigm, Petter Johansson, is part of my supervisory team, I benefited from his insights and our team developed a modified version of the choice blindness paradigm that could be applied in an online setting.

Researchers have also recently applied the choice blindness paradigm in an online experiment to measure and manipulate political attitudes (Strandberg et al., 2020). Participants provided their opinions on different personality traits (e.g., trustworthy) that the two 2016 candidates for President of the USA, Hilary Clinton or Donald Trump, might possess; the choices were indicated by drawing a cross on a scale for each personality trait, having an image of Hilary Clinton on the one side and Donald Trump on the opposite end. The researchers defined (0% - 35%) and (65% -100%) as extreme zones and the manipulation involved selecting the five most extreme responses from the extreme zones and moving them randomly to a more open-minded position (in the middle 30% of the scale). Subsequently, participants were told that research has shown

that the order in which questions are presented might influence behaviour and were asked to review and potentially revise 5 of their 12 responses (Strandberg et al., 2020).

Although Strandberg et al. (2020) were able to successfully replicate the choice blindness paradigm in an online setting, the revision rates were much higher than in the face-to-face choice blindness studies. A potential drawback of Strandberg et al. (2020)'s online choice blindness version could have been the way the revision phase was introduced to participants, which might have led them to believe that they were expected to revise their answers. Hence, our goal was to design an online study in a way that replicated as close as possible the face-to-face settings. To this end, in Study 5, I first asked participants to indicate their self-rankings on a set of personality traits (moral and non-moral). Next, respondents were shown their self-ranking (manipulated or not manipulated) for 20 seconds and asked to reflect on it. Last, a page with a few follow-up questions was shown, concluding with a request for participants to indicate again their self-ranking while their original personality judgement (manipulated or non-manipulated) was set as the starting position of the slider cursor with a paler colour.

Another challenge I had to face when adapting the choice blindness paradigm to the domain of the self was that personality judgements are typically high for desirable personality traits and low for undesirable traits (e.g., Ziano et al., 2021). At the same time, Study 2 had underscored the potential importance of maintaining a positive self-image on self-rankings, hence I wanted to contrast enhancing and diminishing choice blindness manipulations. However, if most self-rankings fall into the upper 30% of the scale for desirable traits and the lower 30% for undesirable traits, the choice blindness manipulations would mostly be in a diminishing direction, which might prevent a valid comparison between diminishing and enhancing choice blindness manipulations.

To ensure the credibility of the manipulations and as many as possible self-rankings that could be manipulated both in an enhancing and diminishing direction, we modified the choice blindness paradigm as follows. Personality judgements were moved either up or down by 20 units, based on the following rule: self-rankings that were lower or equal

to 25 were manipulated up only, self-rankings that were higher or equal to 75 were manipulated down only, while self-rankings that were higher than 25 and lower than 75 were manipulated either up or down (randomly). The threshold values (25 and 75) were based on the distributions of responses to different personality traits in Study 2, the aim being to maximise the number of personality judgements in the middle range (between 25 and 75) that could be manipulated both up and down. Again, we had to carefully choose the personality traits to ensure a clear distinction between moral and non-moral traits (e.g., “kind” or “competent”); based on previous research (Tappin & McKay, 2017; Ziano et al., 2021), we also selected moral and non-moral traits that had comparable average desirability rankings. This procedure guaranteed a valid comparison between moral and non-moral attitudes as both types of personality judgements were indicated and manipulated in an analogous way.

Experimental design and sampling

All studies in this thesis employ an experimental design as it allows inferring causality between the manipulated variable and the dependent variable (Kirk, 2012). Initially some of the studies reported here were planned as lab experiments, however, the outbreak of the COVID 2019 pandemic rendered the option for collecting data face-to-face virtually impossible. In addition, as pointed out above, the general technique used throughout the studies was to expose moral and non-moral behaviours to the same psychological phenomenon (anchoring or the choice blindness paradigm), which naturally called for a between-subject design in all the studies. A between-subject design, however, requires a relatively large number of participants to detect a small effect size. For example, for Study 1, 416 participants were needed to achieve 80% power to detect a small effect of $d = 0.2$ at $\alpha = .05$ (calculated using G Power).

Considering the relatively large sample sizes needed for the studies and the outbreak of COVID-19, collecting data online seemed the only feasible option.

Accordingly, participants for all the studies in this thesis were recruited online. I relied on *Prolific* (www.prolific.com) for all studies, except for the pilot of Study 1, which was conducted using Amazon's Mechanical Turk. The strengths and weakness of online research in relation to their potential effects on the quality of the data reported in this thesis are briefly discussed below.

Conducting research online has become widespread in recent years due to the easy access it offers to relatively inexpensive data (Anwyl-Irvine et al., 2021; Manago et al., 2021). As mentioned above, I generally needed large sample sizes for the experiments and relying on the online platforms allowed me to collect data for each study within hours. Indeed, even after applying three pre-screening criteria (UK nationality, monolingual English speakers and approval rate higher than 90%), there were more than 33,000 matching participants on *Prolific* available to take part in our studies. Such rapid data collection helped me perform all the studies planned in this thesis despite the challenges of the pandemic. In addition, as online platforms provide more diverse samples than the university pools (Buhrmester et al., 2011; Casler et al., 2013), recruiting respondents online contributed to a better generalisability of the obtained results.

Recently, however, concerns have been raised that although not suffering all the limitations of the convenient university sample, sampling from online platforms might systematically affect the collected data (Burnham et al., 2018; Chandler et al., 2017; Stewart et al., 2017). For instance, online samples might be systematically different in certain important characteristics such as religiosity with a relatively high number of Amazon's Mechanical Turk's workers identifying as atheists or agnostics (Burnham et al., 2018). In addition, Chandler et al. (2014) reported evidence for non-naivety of the respondents: data from 132 studies conducted on Amazon's Mechanical Turk was pooled and the analysis revealed that the most active 1% of the workers provided 11% of the total responses collected while 10% of the most active workers provided 41% of the total data. This is one of the reasons I relied on *Prolific* instead, which is a relatively

recently developed platform with lesser risk of encountering workers who are acquainted to some degree with the tasks (Palan & Schitter, 2018).

Data collected via *Prolific* has been found to be of a higher quality, based on naivety of respondents, attention, comprehension, and reliability than data gathered via Amazon's Mechanical Turk, CloudResearch and panels, Qualtrics and Dynata (Peer et al., 2021). Indeed, *Prolific* was specifically designed for conducting academic research and has valuable inbuilt functions such as a provision for pre-screening participants by approval rate, which tackles inattention, one of the biggest potential weaknesses of online data collection (Peer et al., 2017). Research has shown that Amazon's Mechanical Turk workers self-report being together with other people or engaging in parallel activities such as watching TV or listening to music while participating in experimental studies, which might result in low quality or even misleading data (Chandler et al., 2017). Recent research also found that removing data from inattentive participants from the sample might lead to substantially different results (Sulik et al., 2023).

Although using *Prolific* did not guarantee I would have participants' full attention, the platform has a few features that facilitate filtering inattentive participants. For instance, a time-out period is automatically set, and a submission is not accepted if it has taken an unreasonably long time. Moreover, participants might decide to return a submission, which is an easy way to withdraw data, if they change their mind during the study. This is also beneficial from an ethical point of view as researchers would like participants to be able to withdraw their data at any point of the experiment, but also from a financial point of view as returned or timed-out submissions do not need to be reimbursed.

To ensure a high quality of the collected data, I applied pre-screening, attention check questions, manipulation checks and minimum completion time requirements throughout the studies. For Study 1, I did not use pre-screening, but relied on attention and manipulation checks as well as minimum completion time. I lost a substantial amount of data in Study 1: 44% of the recruited participants failed at least one of the

checks. In retrospect, I considered potential inattentiveness or lack of understanding of the task as potential reasons for the high percentage of respondents who failed the checks. Therefore, for the rest of the studies I applied three pre-screening criteria: approval rate higher than 90%, UK nationality and monolingual English speakers (to ensure task comprehension). With all these safeguards in place, I expected to have ensured a high passing rate and high quality of data. Frustratingly, I still lost a lot of data in all but the last study: 30% in Study 2; 25% in Study 3; and 35% in Study 4¹. The data from the remaining samples should be of a high quality as these were pre-screened respondents, who also passed both the attention and the manipulation checks.

Study 5 was by far the most successful with respect to passing rate as I had to exclude only 2% of the participants due to failing to answer correctly at least one of the attention check questions. The percentage of excluded data is relatively low also because inattentive participants were filtered out during the study itself - Prolific provides the option to discontinue participation of respondents who fail both attention checks, which was the case for 51 participants (9% of the total submissions). Although I had to accept and reimburse participants who failed only one of the attention check questions during the study, I excluded them from the analysis to ensure high quality of the data (this exclusion criterion was specified in our pre-registration document).

Another important potential caveat of conducting online research is the comparability of research evidence between online and lab experiment. However, previous research has shown that online participants exhibit similar behaviour patterns to respondents in lab experiments (Manago et al., 2021; Casler et al., 2013; Gosling et al., 2004). The comparability between the results of online and lab experiments has been demonstrated across various judgment and decision-making tasks such as framing effects (Berinsky et al., 2012), reaction time in lexical decision tasks (Hilbig, 2016) and speech recognition (Byun et al., 2015) among others. Most relevant for the tasks involved in this thesis, research has shown online reproducibility of lab-based results on

¹ The pilot for Study 3 and Study 4 showed lower rates of lost data (16%).

anchoring (Röseler & Schütz, 2022), personality measures (Clifford et al., 2015) and behaviour in allocation choices (Amir et al, 2012; Hergueux & Jacquemet, 2015).

Another factor that affects behaviour in allocation choices is whether the choice has monetary consequences (Forsythe et al., 1994). For example, when playing a Dictator Game, participants are more generous in hypothetical games than when playing with real stakes (Amir et al, 2012). Once real stakes are introduced however, behaviour is affected very little, if at all, by the size of the stakes (Keuschnigg et al., 2016; Larney et al., 2019). To ensure our task captures real life behaviour, I introduced additional financial incentives in the experiments. Besides participation fees (which depended on the pre-set duration of the experiment), participants were paid their respective reward from the game played. For example, in Study 1, the wheel of fortune task was simulated and participants (or the respective charity they chose) received the amounts won. In Study 4, respondents also played the Dictator Game with real stakes and dictators kept the portion of the amount they indicated. The financial incentives implemented in some of the studies (Study 1 and Study 4) should have contributed to higher levels of participants' engagement and higher quality of the collected data.

Measuring choices and personality judgements

To ensure a robust experimental design in investigating choices with and without moral implications (Study 1), I measured revealed behaviour instead of intention statements or self-assessed hypothetical behaviour. Theoreticians have long explored the discrepancies between attitudes, stated preferences and actual behaviour, the so-called "attitude-behaviour" or "intention-behaviour" gaps (Ajzen, 1991). A meta-analysis showed that a medium to large change in intentions results in only small to medium change in actual behaviour (Rhodes & Dickau, 2012). For instance, researchers registered a gap between stated intentions of purchasing environmentally friendly

products and actual purchases (Grimmer & Miles, 2017). Moreover, stated intentions to vote did not lead to higher voting turnover (Nickerson & Rogers, 2010).

Due to the intrinsic difficulties of creating a valid experimental design that compares moral and non-moral choices highlighted above however, I conducted Studies 2 to 5 in the domain of the self. Collecting data on personality judgements had the advantage of providing a clear distinction between moral and non-moral traits as well as comparable measurement, however I had to rely on self-assessment rather than revealed behaviour. Nevertheless, as Studies 2 to 5 explored the susceptibility to a certain psychological framework (anchoring or the choice blindness paradigm) and test whether the cognitive influence differentially affects moral and non-moral self-rankings, a potential exaggeration of the respective self-rankings should have affected all conditions. As discussed above, self-rankings were reported relative to 100 other participants (Studies 2 to 4) or in comparison to “a large number” of participants (Study 5), which ensured a similar benchmark for the extremities of the scale across respondents. Relying on self-assessment also allowed us to explore fascinating biases like self-serving biases.

Furthermore, research in personality predominantly relies on self-reports (Paulhus & Vazire, 2007). Self-rankings predict behaviour and life outcomes, and academic and job performance (Beck & Jackson, 2022; Ozer & Benet-Martinez, 2006; Paunonen & Ashton, 2001; Roberts et al., 2007; Zell & Lesick, 2022) to a similar degree as well-established predictors such as cognitive abilities and socioeconomic status (Heckman & Kautz, 2012). As individuals are prone to maintain a stable and positive self-image despite conflicting evidence however (e.g., Stanley et al., 2019), it is important to discuss the accuracy of personality judgements.

To test the accuracy of personality judgements researchers have compared them to judgements of knowledgeable others (Back & Vazire, 2012) and showed that self- and other-perceptions have similar success in predicting behaviour and life outcomes (Kolar et al., 1996; Vazire, 2010; Vazire & Mehl, 2008). A recent meta-analysis (Oltmanns et al., 2020) provided further evidence of strong self–other agreement on longitudinal

personality change in older adults. Nevertheless, the predictive validity depends on the type of personality trait: other-judgements are more accurate predictors of behaviour stemming from evaluative traits (e.g., intelligence) while self-rankings are more accurate predictors of behaviour related to internal traits such as self-esteem (Connelly & Ones, 2010; Vazire, 2010).

Yet, the gold standard in demonstrating accuracy of personality judgements is by predicting overt behaviour (Back & Vazire, 2012). Hence, Study 3 and Study 4 were designed to test whether anchoring self-rankings on two moral traits (honest and considerate) would impact subsequent personality judgements across a range of personality traits (Study 3) and prosocial choices (Study 4). We chose these specific personality traits as honesty-humility is theorised as an additional dimension to the Big Five (HEXACO model of personality, Ashton & Lee 2020, Thalmayer & Saucier 2014), encompassing prosociality (Ashton & Lee 2014; Zettler et al., 2020). Research has provided evidence of the positive association between prosociality and honesty (Isler & Gächter, 2022; Soraperra et al., 2019). High self-assessments on the honesty-humility dimension of HEXACO are also associated with prosocial behaviour in the Dictator Game (Hilbig et al., 2013; Thielmann et al., 2020; Zettler et al., 2020).

Pre-registration

All studies in this thesis were pre-registered on AsPredicted (the links to the pre-registration documents are provided in each respective study). Pre-registering a study involves specifying the hypotheses, research methods, sample size and analysis strategy before the data is collected. Pre-registration, along with ensuring that the data and the analysis code are publicly available, are important Open Science practices, contributing to more credible and reproducible psychological research and addressing the recent replication crisis in psychology (Chambers et al., 2014; Lindsay, 2017; Nosek & Lackens, 2014; Nosek et al., 2018; Open Science Collaboration, 2015).

Pre-registration benefits the research process and has gradually become the norm in social and behavioural sciences (Logg & Dorison, 2021). The main advantages of

pre-registration are three-fold (Nosek et al., 2019; Wagenmakers et al., 2012): (i) distinguishing between confirmatory and exploratory research by specifying which analyses were planned a priori, and thus maintaining the generally accepted 5% false positive error rate in null hypothesis significance testing; (ii) preventing researchers from Hypothesising After the Results are Known (HARKing, Kerr, 1998) and (iii) mitigating the effect of publication bias by providing an accessible and searchable database of planned studies, regardless of whether these studies were published.

Pre-registering the studies in this thesis required thorough and detailed planning of each experiment. I also piloted most of the studies to test the suitability of the respective tasks and gather feedback. This preliminary work allowed us to anticipate potential issues and tailor the design accordingly to avoid them. It was also more cost-efficient – as all the studies required relatively large sample sizes, it was prudent to carefully think about all potential pitfalls before spending resources on the studies.

The only drawback of pre-registration, especially for the first studies, was that it was quite challenging to anticipate all the steps involved in analysing the experiments. For instance, when preparing the pre-registration document for Study 2, I specified ANOVA as the planned statistical analysis. However, I had both attention and comprehension checks in place and the comprehension checks only applied to participants in the anchoring conditions. Because of this, I lost more data in the anchoring condition than in the control condition, which led to an unbalanced final sample. Fitting a linear mixed model (LMM) was eventually more suitable to analyse the data. Both the ANOVA and the LMM analysis conveyed the same message and for transparency reasons, I reported both the LMM and the ANOVA analysis, yet it shows some of the challenges involved in pre-registering a study.

Methodology summary

To summarise, all studies presented in this thesis rely on a unified concept: the stability of preferences was investigated by exposing choices or personality judgements

to a certain psychological phenomenon (anchoring or choice blindness). In addition, I tested for self-serving effects of the cognitive influences as well as for differential effects depending on attitude domain - moral or non-moral. In line with Open Science practices, all studies were pre-registered. The studies required relatively large sample sizes and were conducted via online research platforms, which contributed to more diverse and representative samples. I also put in place safeguards against all known potential drawbacks of online data collection, such as attention and comprehension checks and pre-screening participants. Moreover, financial incentives were applied when suitable to ensure participants' full engagement with the tasks as well as more reliable and generalisable data.

Chapter 3. Study 1. The Effect of Socially Imbued Anchors on Choices with and without Moral Implications

Study 1. The Effect of Socially Imbued Anchors on Choices with and without Moral
Implications

Elitza Ambrus

Bjoern Hartig

Ryan McKay

Manuscript in preparation

Word count: 4390 excluding references and appendices

Abstract

Individuals typically express their moral views with strong conviction as if guided by an inner “moral compass”. Research has shown however that our moral attitudes are malleable both across contexts and over time, that is our “moral compass” has no fixed magnetic North. Here, we test whether moral choices are *more malleable* than choices in other domains. To this aim, we designed a novel task (“wheel of fortune” task) and investigated the susceptibility of choices with and without moral implications to anchors, imbued with social meaning. In an online, incentivised experiment, participants (N = 432) indicated their desired number of spins in the wheel of fortune task on a slider bar (0-140, no numeric values were displayed), playing either for themselves or on behalf of a chosen charity. The data showed that individuals’ choices both with and without moral implications are flexible and susceptible to social anchoring. Contrary to our expectations however, the social anchoring effect was not more pronounced for choices with moral implications. Limitations of the study, such as the intrinsic challenges of contrasting classic moral choices to choices in a non-moral domain in an experimentally valid way are discussed.

Keywords: anchoring, social influence, choices with moral implications

The Effect of Socially Imbued Anchors on Choices with and without Moral Implications

Individuals typically express their moral views with strong conviction (Haidt & Graham, 2007) as if guided by an inner “moral compass”. Although moral preferences have commonly been modelled as stable (e.g., Fehr & Schmidt, 1999), an alternative line of research suggests that moral attitudes are malleable both across contexts and over time (e.g., Cialdini et al., 1999). For instance, individuals tend to behave in line with what they believe the majority would recognise as normative behaviour in the respective task context (Krupka & Weber, 2013). Such findings imply that our “moral compass” has no fixed magnetic North. Rather, psychological factors such as observations of others’ behaviour and of our own previous behaviour, exert their own attractions on the needle. This study investigates whether moral preferences are *more malleable* than preferences in other domains. To this end, we integrate two well-established psychological influences: the anchoring effect (Tversky and Kahneman, 1974) and social impact (e.g., Krupka & Weber, 2013), and explore their joint effect on choices with and without moral implications.

Theoretical framework

Researchers in judgment and decision-making have traditionally modelled agents’ preferences as stable and consistent (Von Neumann & Morgenstern, 1947). In the moral domain, preferences have commonly been assumed as stable with an ongoing debate on whether most of us possess selfish or prosocial moral preferences (Knoch & Fehr, 2007; Rand et al., 2014). However, there is substantial empirical evidence of flexible, context-dependent preferences across various decision-making domains (Hoffman et al., 1996; Payne et al., 1992; Slovic, 1995). For instance, perceived social distance between participants and the experimenter was found to influence generosity in an income distribution task (Hoffman et al., 1996).

In a similar vein, in-group moral norms seem to guide agents' behaviour (Ellemers et al., 2013). Individuals tend to adjust their moral preferences to accommodate concerns about self-image and reputation (Akerlof & Kranton, 2000; Bénabou & Tirole, 2006), aiming at *presenting* their behaviour as prosocial (Andreoni & Bernheim, 2009; Dana et al., 2006). Therefore, moral preferences might be stable only to the degree they consistently comply with varying social norms (Krupka & Weber, 2013) and to the extent that they yield social recognition and self-esteem (Heintz et al., 2016).

Furthermore, the literature on moral credentials/cleansing suggests that moral preferences are malleable not only across task contexts but also over time: both previous moral behaviour as well as intentions to engage in future moral behaviour systematically influence current moral choices (Blanken et al., 2015; Cascio & Plant, 2015; Mullen & Monin, 2016; cf. Blanken et al., 2014). Agents engage in dynamic "moral licensing" with their own previous moral behaviour, and that of close others, being negatively correlated with subsequent moral behaviours (Brañas-Garza et al., 2013). For instance, close others' environmentally friendly behaviour seems to "license" individuals to subsequently behave in a less environmentally friendly manner (Meijers et al., 2019). Therefore, rather than evincing stability, moral preferences adapt dynamically over time and social contexts and might be more sensitive to social influence than preferences in other domains.

From a methodological point of view, however, contrasting moral choices with choices in other domains is intrinsically challenging as moral decisions are characterised by an integral understanding of fairness and appropriate social norms (Andreoni & Bernheim, 2009). Researchers in the domain of social preferences have found that contextual social norms influence choices. For instance, there are commonly two modes in income-distribution tasks: either sharing nothing, representing selfish behaviour, or donating half of one's endowment, reflecting prosocial behaviour (Camerer, 2003). Contextual factors, such as anonymity of the choice, seem to direct behaviour to one of these two modes, swaying behaviour in income distribution tasks from prosocial to

selfish and vice versa (List, 2007). As subjective perception of social norms is pertaining to moral tasks, it influences subsequent behaviour (Krupka & Weber, 2013). To investigate moral and non-moral choices in an experimentally valid design therefore, a task that would not elicit any contextual social norms or a straightforward interpretation of fair behaviour is needed.

We have developed a novel task that is devoid from contextual social norms (the “wheel of fortune task”) to explore the malleability of moral preferences. The wheel of fortune task involves risky choices that are hard to be classified as “selfish” or “prosocial”. In particular, respondents receive an initial endowment of 10p; they should then indicate the number of times a computer should spin a wheel of fortune. The wheel of fortune has ninety-nine “good” spaces, which bring gains (an extra 2p) and one “bad” space, which terminates the game and all the money accumulated so far, including the initial endowment is lost. When distributed in the condition with moral implications, the task is equivalent, however instead of on their behalf, respondents play on behalf of a charity. Therefore, if they win any money, the amount is transferred to a charity of participants’ choice, if they lose all the money, the charity receives nothing.

The settings of the “wheel of fortune task” are probabilistic, which prevents single interpretation of what constates “selfish” or “altruistic” behaviour in the given context. Subsequently, when a participant is playing on behalf of a charity, it is not clear, how many spins one should choose to behave prosaically (or selfishly). This should prevent any clustering of responses, reflecting perceived social norms. We have piloted the task (Appendix 3A) and the results showed that distribution of responses in the wheel of fortune task does not exhibit clustering around a certain value (number of spins). We have thus employed the wheel of fortune task to contrast individuals’ decisions with and without moral implications by exposing them both to the anchoring effect and social influence.

The anchoring effect is typically demonstrated in a two-stage framework, introduced in the seminal work of Tversky and Kahneman (1974). The researchers asked

whether participants' estimates of a given quantity (the percentage of African countries in the UN) was lower or higher than a randomly generated number (the outcome of a wheel of fortune). At the second stage of the experiment, respondents had to provide their own estimate on the same quantity. The results supported a strong effect of the value used in the comparative question (the anchor) on the consecutive own estimates with participants providing an average estimate of 25 and 45 percentage for anchor values of 10 and 65 respectively (Tversky & Kahneman, 1974). The anchoring effect has been replicated across a wide range of decision domains (Furnham & Boo, 2010; Röseler & Schütz, 2022; Strack et al., 2016; Yoon et al., 2019), including value estimates (Ariely et al., 2003) and negotiations (Galinsky & Mussweiler, 2001). The anchoring effect persists over time (Mussweiler, 2001), when participants are experts in the respective field (Englich et al., 2006) or when they are forewarned about the potential influence of anchoring (Wilson et al., 1996).

We have incorporated social impact into the anchoring paradigm as individuals' behaviour is embedded in and interacts dynamically with the social environment (Mischel & Shoda, 1995; Stets & Burke 2000). Social psychological research has provided plenty of evidence of social influence on individuals' behaviour (e.g., Cialdini et al., 1990). Regardless of whether social influence is explained as stemming from conformity (Asch, 1956), pressure (Latane, 1981) or comparison with others (Festinger 1954), agents' behaviour seems to strive to align with others' behaviour. Conversely, research in the moral domain points towards moral preferences that consistently adapt to beliefs about what the majority would consider socially appropriate behaviour (Krupka & Weber, 2013) with consistent interindividual susceptibility to follow social norms (Kimbrough & Vostroknutov, 2016). For instance, a field experiment by Schultz et al. (2015) showed that providing consumers with real-time feedback about the electricity consumption of similar households influenced individuals' energy consumption. To the best of our knowledge, previous research has not investigated whether anchors imbued with social meaning differentially affect choices with and without moral implications.

This study employs the wheel of fortune task to investigate the impact of two key factors on choices with and without social implications: (i) nature of the beneficiary (self or others) and (ii) social anchoring; as well as their interaction. Instead of the comparative question, typically used in the anchoring paradigm, we have introduced a much more natural way of engagement with the anchor value: the slider cursor on the response scale is pre-set at the anchor with the exact numeric value displayed above the slider cursor (please see Appendix 3B). The participants have to click on the anchor value and move the slider cursor in order to indicate their responses, which effectively leads them to adjust the value upwards or downwards. The anchor value is presented as the average number of spins chosen by participants in the first round of the experiment, imbuing the anchor with social connotations.

To distinguish whether social information serves as some kind of heuristic for acceptable social behaviour, we also record and analyse participants' response time. Based on the bounded rationality concept (Simon, 1956) that individuals' resources such as time, knowledge and cognitive abilities are limited, individuals often simplify complex cognitive tasks by relying on heuristics (Todd & Gigerenzer, 2001). People tend to be cognitive misers (Taylor, 1981) and following others' behaviour might be perceived as a useful heuristic especially in an uncertain context (Gigerenzer, 2010). Flexible moral preferences, based on the simple heuristic of taking into account what the majority does, seems to ensure both social inclusion as well as choices that could be justified from bounded rationality point of view (Borah & Kops, 2019). Therefore, if the social connotation of the anchor value we provide serves as heuristics, choices in the anchoring condition should be quicker on average than choices in the control condition.

Our experimental design involves measuring actual behaviour as opposed to self-assessment and stated intentions as research have found substantial differences between actual behaviour and stated preferences, the so-called "attitude-behaviour" or "intention-behaviour" gap (Ajzen, 1991). For example, there is a discrepancy between actual purchases of environmentally friendly products and stated intentions (Grimmer & Miles, 2017). Moreover, respondents who believed they are superior to others in terms

of moral qualities, do not reveal superior levels of fairness, trust and sharing with others in their overt behaviour (Tappin & McKay, 2019).

Furthermore, observed behavioural patterns of social preferences differ systematically between hypothetical tasks and tasks with monetary consequences (Forsythe et al., 1994). For example, participants were significantly more generous when sharing part of their hypothetical initial endowment than when real stakes were introduced (Amir & Rand, 2012). Once monetary incentives are introduced, however, the particular stake size has very small or no effect on choices (Keuschnigg et al., 2016; Larney et al., 2019). Our study involves monetary consequences either for the participants themselves or for their chosen charity, which ensures participants' engagement with the task.

We have conducted both the pilot and the study online to ensure we could collect a large sample of participants coming from more diverse cultural and socio-economic backgrounds than a lab-based experiment would allow. This approach facilitates better generalisability of the experimental results. In addition, research indicates there is no significant difference between social preferences revealed on online platforms and in lab settings (e.g., Amir & Rand, 2012). Furthermore, to address a potential drawback of conducting online research, namely potential distraction due to the lack of control over the experimental environment, we have included two comprehension check questions. Only data from participants who have correctly answered both comprehension questions are included in the analysis.

Based on the robustness of the anchoring effect and the guiding role of moral norms for behaviour (Ellemers et al., 2013), our main hypotheses are that there will be an anchoring effect both on choices with and without moral implications (H1a) however the anchoring effect will have a stronger influence on choices with moral implications (H1b). As secondary hypotheses, we expect the response speed to be faster in the anchoring than in the control condition for both tasks as well as that the response speed will be faster for choices with moral implications than for choices without moral implications (H2a and H2b); and that the variance of choices will be smaller in the

anchoring than in the control condition for both tasks with smaller variance for choices with moral implications (H3a and H3b).

Method

Participants

Participants were recruited via the online platform *Prolific* (www.prolific.com). We aimed to recruit 416 participants (the sample size was calculated using G Power to achieve 80% power to detect a small effect of $d = 0.2$ at $\alpha = .05$). We had specified in our preregistration document that we would not analyse data until we had at least 416 respondents who passed the exclusion criteria - either failing to correctly answer any of the two comprehension questions or completing the study in less than 60 seconds. Both comprehension questions relied on different scenarios of options chosen in the wheel of fortune task, thus checking participants' understanding of the rules of the task. Initially, 440 participants were recruited to allow for exclusions. However, only 55% managed to pass the comprehension check, so we continued recruiting participants (in batches of 50 participants) until we reached the required number. We recruited 772 respondents (432 female, 332 male, 4 other, 4 chose not to specify, $M = 33.94$, $SD = 11.86$), of whom only 432 participants passed the comprehension check (204 female, 223 male, 2 other, 3 chose not to specify, $M = 33.03$, $SD = 11.86$) and were included in the analysis. All participants took more than 60 seconds to fill in our study, so there was no data excluded based on this criterion.

Participants were paid a flat participation fee (63p, the equivalent of £7.50 per hour) as well as a bonus depending on the outcome from spinning the wheel of fortune their chosen number of spins. The money won by the participants who were assigned to the "charity" condition (i.e., played on behalf of a charity) was transferred to the respective charities. The study was self-certified in accordance with the Royal Holloway, University of London Ethics Committee procedure.

Materials and Procedure

The study was designed in Qualtrics. Participants were assigned to one of the four experimental conditions, resulting from the intersection of the anchor and beneficiary factors: personal risk without anchoring, risk-for-charity without anchoring, personal risk with anchoring and risk-for-charity with anchoring. In each condition, respondents had to indicate their desired number of spins in the wheel of fortune task described above (please see also Appendix 3B), framed either as personal risk or risk-for-charity. Respondents in the risk-for-charity conditions were first provided with a list of charities and had the option to select one of them after they had indicated their chosen number of spins. In all conditions, participants indicated their choices on a slider bar. In the anchoring conditions, the slider cursor was pre-set to 92 (the anchor value) and participants were told that this value represented the average number of spins in the first round of the experiment (this was a deception, however participants were fully debriefed at the end of the experiment and given the opportunity to withdraw their data). In the control conditions, the slider cursor was invisible, and participants were instructed to click on the slider bar in order for the slider cursor to appear. The range of the slider was between 0 to 140 spins, inclusive, with no numeric values displayed to participants.

Design and Analysis

The experiment employed a 2 (personal risk task vs risk-for-charity task) x 2 (socially meaningful anchor vs no anchor) between-groups design. The DVs were the number of chosen spins in the wheel of fortune game (H1 and H3) and response time (H2). To test our hypotheses H1 and H2, we ran two-way ANOVAs on the full sample and post-hoc tests for binary comparisons. To test H3, we performed the 'Asymptotic test for the equality of coefficients of variation from k populations' and the 'Modified signed-likelihood ratio test (SLRT) for equality of CVs', included in the R package `cvequality` (Version 0.1.3; Marwick & Krishnamoorthy, 2019). All analyses were conducted using R Studio 3.5.1. Our hypotheses, data collection, and analysis protocol, including the criteria for data exclusions were pre-registered at (https://aspredicted.org/9GN_F5C).

De-identified data and analysis scripts are available on the Open Science Framework:
https://osf.io/6z75m/?view_only=c3f6578c29d140b7bed2e5befa2f9c89

Results

First, the data was explored visually, Figure 1 depicts the histograms for the distributions of the number of spins chosen for both the personal risk and the risk-for-charity conditions either in the no anchor (left panel) or in the anchor (right panel) condition. The histograms for the personal risk and the risk-for-charity in the anchoring condition demonstrate certain differences in the distributions of the number of spins chosen (right panel of Figure 1). Investigating the violin plots at Figure 2 however, suggests that the differences between the two types of risks in each of the anchoring conditions might not be significant.

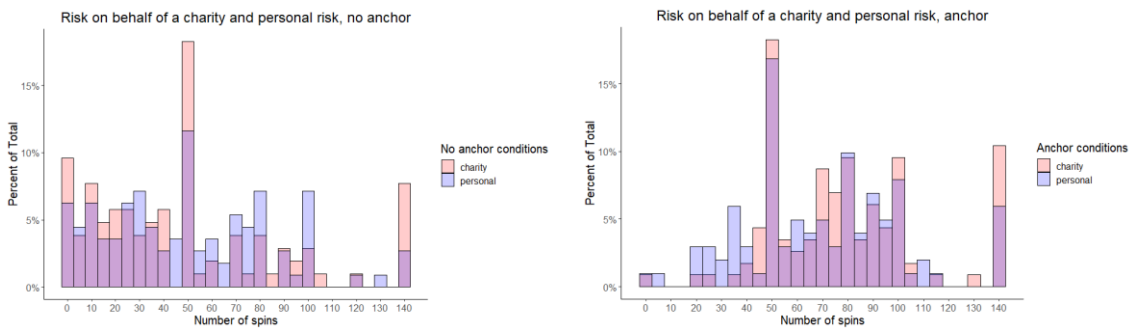


Figure 1. Histograms of the number of spins chosen for each type of task in the no anchor (left panel) and anchor (right panel) by condition.

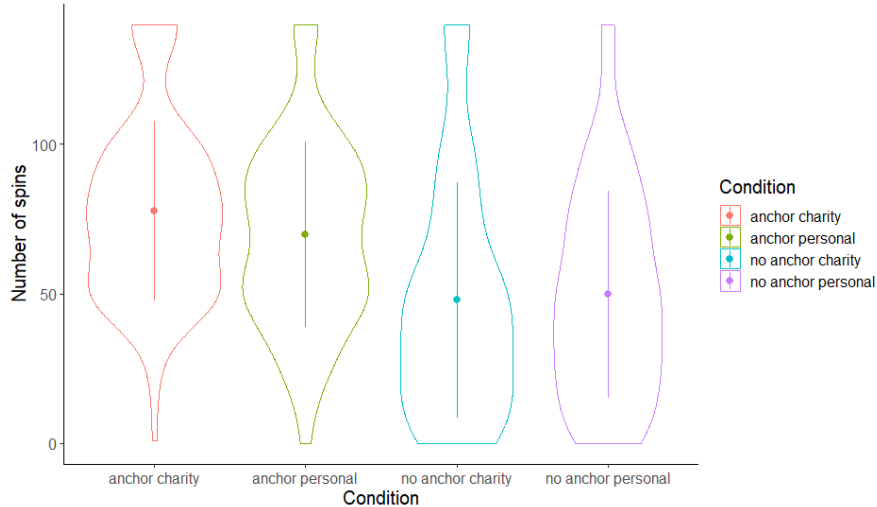


Figure 2. Violin plots for the number of spins chosen by condition (anchor value = 92 spins). The point range represents 2SD around the mean.

To formally explore the effects of anchor, beneficiary, and their interaction on the chosen number of spins (H1a and H1b), we conducted a two-way ANOVA. There was a highly significant effect of anchoring on choices, $F(1, 428) = 59.95, p < .001$. However, neither the main effect of beneficiary, nor the interaction between anchor and beneficiary were significant, $F(1, 428) = 0.88, p = .35$ and $F(1, 428) = 2.33, p = .13$, respectively.

Post-hoc pairwise comparisons (using the Holm method) showed that participants chose to spin the wheel of fortune fewer times in the personal risk task without anchoring ($M = 49.75, SD = 34.56$) than in the personal risk with anchoring task ($M = 69.78, SD = 30.89$), $t(428) = 4.31, p < .001$. Similarly, fewer spins were indicated in the risk-for-charity task without anchoring ($M = 47.85, SD = 39.43$) than in the risk-for-charity with anchoring condition ($M = 77.83, SD = 30.05$), $t(428) = 6.54, p < .001$. Nevertheless, the difference between personal risk with anchoring ($M = 69.78, SD = 30.89$) and risk-for-charity with anchoring ($M = 77.83, SD = 30.05$) was not statistically significant, $t(428) = 1.74, p = .08$. The analysis provides strong support for the anchoring effect on choices (H1a), however there was no evidence in support of a differential effect of anchoring depending on type of task (H1b), $p = .08$. Although the interaction

plot (Figure 3) also suggests some evidence in support of H1b, the coefficient for the interaction did not reach statistical significance.

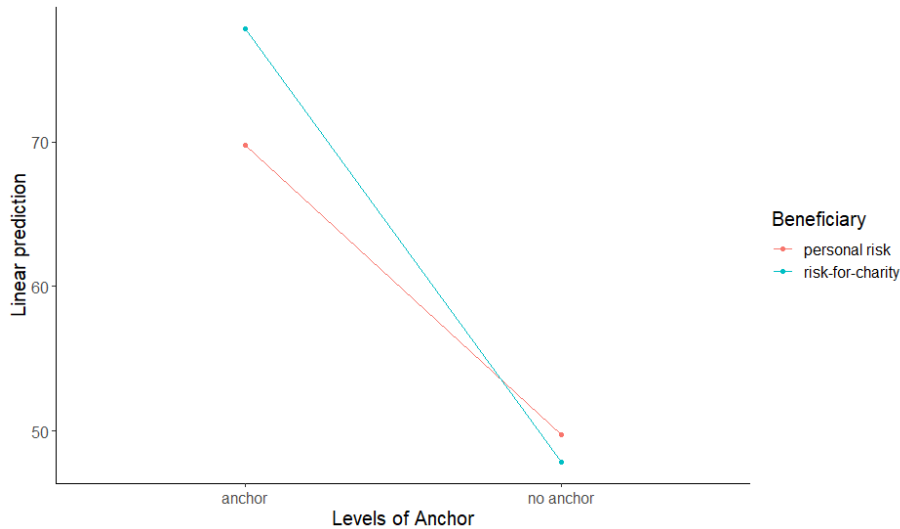


Figure 3. Effect of anchoring on number of spins in the personal risk and risk-for-charity task.

The data did not meet all of the assumptions for performing a two-way ANOVA as the Shapiro-Wilk tests showed that the data did not follow a normal distribution, $W=0.97, p < .001$. Consequently, we performed Kruskal-Wallis test, which revealed a significant difference in the number of spins chosen among the experimental conditions, $\chi^2(3) = 62.28, p < .001$. Follow-up pairwise comparisons with the Dunn test showed that there was a highly significant effect of anchoring on choices both in the personal risk task ($n = 213, p < .001$) and the risk-for-charity task ($n = 219, p < .001$). However, there was no significant difference in the effect of anchoring between personal risk and risk-for-charity ($n = 216, p = .08$).

To test whether the response speed would be quicker in the anchoring than in the no anchor condition for both tasks (H2a) as well as quicker in the risk-for-charity than in the personal risk task (H2b), we performed a Kruskal-Wallis test to explore the effects of anchor and beneficiary on response time. Contrary to our expectations

however, the analysis of the response time showed that there was no difference in response time among groups ($W=4.04$, $p = .26$).

Finally, to test whether the variance of choices would be smaller in the anchoring than in the control condition for both tasks (H3a), we conducted an analysis of variance, namely the “Asymptotic test for the equality of coefficients of variation from k populations” and the “Modified signed-likelihood ratio test (SLRT) for equality of CVs” (cvequality, Version 0.1.3; Marwick & Krishnamoorthy, 2019). Both tests showed that there was a significant difference in the variance between the anchor condition and the control condition (Asymptotic test $AT = 43.71$, $p < .001$ and M-SLRT= 46.14, $p < .001$), thus providing evidence in support of H3a. Next, we applied the same tests for H3b, analysing the difference in variance between the two tasks, personal risk, and risk-for-charity, in the anchor and the non-anchor condition (H3b). Both asymptotic tests for the equality of coefficients of variation showed that there was no significant difference between the variance of the two tasks either in the no anchor or anchor conditions ($p = .22$ and $p = .23$ respectively). Hence, the data did not show support for a smaller variance of risk-for-charity than personal risk in the anchoring condition.

Discussion

Although moral views are commonly expressed with strong conviction, research has shown that moral attitudes are malleable both across contexts and over time. Here, we investigated whether moral choices are *more malleable* than non-moral choices. The data showed that individuals’ choices both with and without moral implications are flexible and susceptible to social anchoring. These findings are in line with research showing the robustness and replicability of the anchoring effect across various decision-making domains (Yoon et al., 2019; Röseler & Schütz, 2022). In addition, the data provided further evidence for the malleability of choices with moral implications, which points towards moral attitudes being shaped by cognitive influences and social

comparisons (Kimbrough & Vostroknutov, 2016). Contrary to our expectations however, the social anchoring effect was not stronger for choices with moral implications.

Participants' response time also did not differ significantly across conditions. Nevertheless, the analysis showed that there was a smaller variance of choices in the anchoring condition compared to the no anchor condition. Social anchoring seems to render choices somewhat cognitively easier, however there was no differential effect depending on the type of choice. Although the data did not provide evidence for a magnified effect of social anchoring on choices with moral implications, it provided strong evidence for the effect of anchoring on both types of choices as well as for the influence of social information on the variability of choices.

A potential explanation for the lack of evidence for a magnified social anchoring effect on choices with moral implications might be the fact that our task involved a choice with moral implications rather than a typical moral choice. We chose "fairness" to tap into moral identity, however participants might have differed in the degree to which they perceived being fair as central to their moral identity (Aquino & Reed, 2002). In addition, as discussed, judgements and choices in the moral domain in general assume an integral understanding of fairness and social norms (Andreoni & Bernheim, 2009). Therefore, we developed and employed a novel task (the "wheel of fortune task"), devoid of contextual social norms and a single interpretation of what constitutes fair, altruistic or selfish behaviour. In constructing a valid experimental design, however, relying on equivalent tasks (with and without moral implications), we might have created a moral task that is devoid of the sensitive context that would potentially exacerbate the effect of social anchoring.

Research has shown that individuals' attitudes in moral dilemmas reveal strong emotional engagement (Greene et al., 2001). Moreover, Rozin (1999) provided evidence for a stronger emotional and behavioural response for moral than non-moral violations. If the elicitation of strong emotions triggers the exacerbated responses in moral judgment and choices, constructing a non-moral equivalent to a classical moral dilemma such as the trolley problem (see for e.g., Greene et al., 2001) might not be feasible.

Therefore, compromising on the essence of the moral choice to achieve comparability across domains might defeat the goal of exploring potential differential effects between moral and non-moral behaviour.

In summary, the data provided strong evidence in support of malleable attitudes, susceptible to anchoring. Nevertheless, there was no support for a magnified effect of social anchoring on choices with moral implications. The response time also did not differ between tasks with and without moral implications; however, the analysis of variance showed that the variance in the anchoring condition was smaller than the one in the no anchor condition, revealing that socially imbued anchors render choices cognitively easier. Although the pilot study pointed towards the suitability of our newly developed task to compare moral and non-moral attitudes, the manipulation we used to introduce moral considerations (risk on behalf of a charity as opposed to personal risk) might not have been decisive enough to contrast moral and non-moral choices. The intrinsic strong emotional engagement in moral choices might render constructing a non-moral equivalent task of a classic moral choice not feasible, thus preventing the experimental exploration of a differential anchoring effect in the moral domain.

References

- Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, 50(2), 179-211. [https://doi.org/10.1016/0749-5978\(91\)90020-T](https://doi.org/10.1016/0749-5978(91)90020-T)
- Akerlof, G. A., & Kranton, R. E. (2000). Economics and identity. *The Quarterly Journal of Economics*, 115(3), 715-753. <https://doi.org/10.1162/003355300554881>
- Amir, O., & Rand, D. G. (2012). Economic games on the internet: The effect of \$1 stakes. *PLoS one*, 7(2). <https://doi.org/10.1371/journal.pone.0031461>
- Andreoni, J., & Bernheim, B. D. (2009). Social image and the 50–50 norm: A theoretical and experimental analysis of audience effects. *Econometrica*, 77(5), 1607-1636. <https://doi.org/10.3982/ECTA7384>
- Ariely, D., Loewenstein, G., & Prelec, D. (2003). “Coherent arbitrariness”: Stable demand curves without stable preferences. *The Quarterly Journal of Economics*, 118(1), 73-106. <https://doi.org/10.1162/00335530360535153>
- Asch, S. E. (1956). Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psychological monographs: General and applied*, 70(9), 1. <https://doi.org/10.1037/h0093718>
- Aquino, K., & Reed, A. II. (2002). The self-importance of moral identity. *Journal of Personality and Social Psychology*, 83(6), 1423–1440. <https://doi.org/10.1037/0022-3514.83.6.1423>
- Bénabou, R., & Tirole, J. (2006). Incentives and prosocial behavior. *American Economic Review*, 96(5), 1652-1678. <https://doi.org/10.1257/aer.96.5.1652>
- Blanken, I., van de Ven, N., & Zeelenberg, M. (2015). A meta-analytic review of moral licensing. *Personality and Social Psychology Bulletin*, 41(4), 540–558. <https://doi.org/10.1177/0146167215572134>

- Blanken, I., van de Ven, N., Zeelenberg, M., & Meijers, M. H. C. (2014). Three attempts to replicate the moral licensing effect. *Social Psychology, 45*(3), 232–238. <https://doi.org/10.1027/1864-9335/a000189>
- Borah, A., & Kops, C. (2019). Rational choices: an ecological approach. *Theory and Decision, 86*(3-4), 401-420. <https://doi.org/10.1007/S11238-019-09689-5>
- Brañas-Garza, P., Bucheli, M., Espinosa, M. P., & García-Muñoz, T. (2013). Moral cleansing and moral licenses: experimental evidence. *Economics & Philosophy, 29*(2), 199-212. <https://doi.org/10.1017/S0266267113000199>
- Camerer, C. F. (2003). Behavioural studies of strategic thinking in games. *Trends in cognitive sciences, 7*(5), 225-231. [https://doi.org/10.1016/S1364-6613\(03\)00094-9](https://doi.org/10.1016/S1364-6613(03)00094-9)
- Cascio, J., & Plant, E. A. (2015). Prospective moral licensing: Does anticipating doing good later allow you to be bad now? *Journal of Experimental Social Psychology, 56*, 110-116. <https://doi.org/10.1016/j.jesp.2014.09.009>
- Cialdini, R. B., Reno, R. R., & Kallgren, C. A. (1990). A focus theory of normative conduct: recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology, 58*(6), 1015. <https://doi.org/10.1037/0022-3514.58.6.1015>
- Dana, J., Cain, D. M., & Dawes, R. M. (2006). What you don't know won't hurt me: Costly (but quiet) exit in dictator games. *Organizational Behavior and Human Decision Processes, 100*(2), 193-201. <https://doi.org/10.1016/j.obhdp.2005.10.001>
- Ellemers, N., Pagliaro, S., & Barreto, M. (2013). Morality and behavioural regulation in groups: A social identity approach. *European Review of Social Psychology, 24*(1), 160-193. <https://doi.org/10.1080/10463283.2013.841490>
- Englich, B., Mussweiler, T., & Strack, F. (2006). Playing dice with criminal sentences: The influence of irrelevant anchors on experts' judicial decision making. *Personality*

and *Social Psychology Bulletin*, 32(2), 188-200.

<https://doi.org/10.1177/0146167205282152>

Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, 114(3), 817-868.

<https://doi.org/10.3758/s13428-021-01694-3>

Festinger, L. (1954). A theory of social comparison processes. *Human Relations*, 7(2), 117-140. <https://doi.org/10.1177/001872675400700202>

Forsythe, R., Horowitz, J. L., Savin, N. E., & Sefton, M. (1994). Fairness in simple bargaining experiments. *Games and Economic Behavior*, 6(3), 347-369.

<https://doi.org/10.1006/game.1994.1021>

Frederick, S. W., & Mochon, D. (2012). A scale distortion theory of anchoring. *Journal of Experimental Psychology: General*, 141(1), 124.

<https://doi.org/10.1037/a0024006>

Galinsky, A. D., & Mussweiler, T. (2001). First offers as anchors: the role of perspective-taking and negotiator focus. *Journal of Personality and Social Psychology*, 81(4), 657.

<https://doi.org/10.1037/0022-3514.81.4.657>

Gigerenzer, G. (2010). Moral satisficing: Rethinking moral behavior as bounded rationality. *Topics in Cognitive Science*, 2(3), 528-554.

<https://doi.org/10.1111/j.1756-8765.2010.01094.x>

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, 293(5537), 2105-2108.

<https://doi.org/10.1126/science.1062872>

Grimmer, M., & Miles, M. P. (2017). With the best of intentions: a large sample test of the intention-behaviour gap in pro-environmental consumer behaviour. *International Journal of Consumer Studies*, 41(1), 2-10.

<https://doi.org/10.1111/ijcs.12290>

- Harris, A. J., Blower, F. B., Rodgers, S. A., Lagator, S., Page, E., Burton, A., ... & Speekenbrink, M. (2019). Failures to replicate a key result of the selective accessibility theory of anchoring. *Journal of Experimental Psychology: General*, 148(9), e30. <https://doi.org/10.1037/xge0000644>
- Haidt, J., & Graham, J. (2007). When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research*, 20(1), 98-116. <https://doi.org/10.1007/S11211-007-0034-Z>
- Heintz, C., Karabegovic, M., & Molnar, A. (2016). The co-evolution of honesty and strategic vigilance. *Frontiers in Psychology*, 7, 1503. <https://doi.org/10.3389/fpsyg.2016.01503>
- Keuschnigg, M., Bader, F., & Bracher, J. (2016). Using crowdsourced online experiments to study context-dependency of behavior. *Social Science Research*, 59, 68-82. <https://doi.org/10.1016/j.ssresearch.2016.04.014>
- Kimbrough, E. O., & Vostroknutov, A. (2016). Norms make preferences social. *Journal of the European Economic Association*, 14(3), 608-638. <https://doi.org/10.1111/jeea.12152>
- Knoch, D., & Fehr, E. (2007). Resisting the power of temptations: the right prefrontal cortex and self-control. *Annals of the New York Academy of Sciences*, 1104(1), 123-134. <https://doi.org/10.1196/annals.1390.004>
- Krupka, E. L., & Weber, R. A. (2013). Identifying social norms using coordination games: Why does dictator game sharing vary? *Journal of the European Economic Association*, 11(3), 495-524. <https://doi.org/10.1111/jeea.12006>
- Larney, A., Rotella, A., & Barclay, P. (2019). Stake size effects in ultimatum game and dictator game offers: A meta-analysis. *Organizational Behavior and Human Decision Processes*, 151, 61-72. <https://doi.org/10.1016/j.obhdp.2019.01.002>
- List, J. A. (2007). On the interpretation of giving in dictator games. *Journal of Political economy*, 115(3), 482-493. <https://www.jstor.org/stable/10.1086/519249>

- Marwick, B. and K. Krishnamoorthy 2019 cvequality: Tests for the Equality of Coefficients of Variation from Multiple Groups. R software package version 0.1.3. <https://github.com/benmarwick/cvequality>
- Meijers, M. H., Noordewier, M. K., Verlegh, P. W., Zebregs, S., & Smit, E. G. (2019). Taking Close Others' Environmental Behavior Into Account When Striking the Moral Balance? Evidence for Vicarious Licensing, Not for Vicarious Cleansing. *Environment and Behavior*, 51(9-10), 1027-1054. <https://doi.org/10.1177/0013916518773148>
- Mischel, W., & Shoda, Y. (1995). A cognitive-affective system theory of personality: reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review*, 102(2), 246. <https://doi.org/10.1037/0033-295X.102.2.246>
- Mullen, E., & Monin, B. (2016). Consistency versus licensing effects of past moral behavior. *Annual Review of Psychology*, 67, 363-385. <https://doi.org/10.1146/annurev-psych-010213-115120>
- Mussweiler, T. (2001). The durability of anchoring effects. *European Journal of Social Psychology*, 31, 431–442. <https://doi.org/10.1002/ejsp.52>
- Payne, J. W., Bettman, J. R., & Johnson, E. J. (1992). Behavioral decision research: A constructive processing perspective. *Annual Review of Psychology*, 43(1), 87-131. <https://doi.org/10.1146/annurev.ps.43.020192.000511>
- Rand, D. G., Peysakhovich, A., Kraft-Todd, G. T., Newman, G. E., Wurzbacher, O., Nowak, M. A., & Greene, J. D. (2014). Social heuristics shape intuitive cooperation. *Nature Communications*, 5(1), 1-12. <http://dx.doi.org/10.2139/ssrn.2222683>
- Röseler, L., & Schütz, A. (2022, March 9). Hanging the Anchor Off a New Ship: A Meta-Analysis of Anchoring Effects. PsyArXiv. <https://doi.org/10.31234/osf.io/wf2tn>

- Schultz, P. W., Estrada, M., Schmitt, J., Sokoloski, R., & Silva-Send, N. (2015). Using in-home displays to provide smart meter feedback about household electricity consumption: A randomized control trial comparing kilowatts, cost, and social norms. *Energy*, *90*, 351-358. <https://doi.org/10.1371/journal.pone.0218702>
- Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review*, *63*, 129–138. <https://doi.org/10.1037/h0042769>
- Slovic, P. (1995). The construction of preference. *American Psychologist*, *50*(5), 364. <https://doi.org/10.1037/0003-066X.50.5.364>
- Stets, J. E., & Burke, P. J. (2000). Identity theory and social identity theory. *Social Psychology Quarterly*, 224-237. <https://doi.org/10.2307/2695870>
- Strack, F., Bahník, Š., & Mussweiler, T. (2016). Anchoring: accessibility as a cause of judgmental assimilation. *Current Opinion in Psychology*, *12*, 67-70. <https://doi.org/10.1016/j.copsyc.2016.06.005>
- Tappin, B. M., & McKay, R. T. (2019). Investigating the relationship between self-perceived moral superiority and moral behavior using economic games. *Social Psychological and Personality Science*, *10*(2), 135-143. <https://doi.org/10.1177/1948550617750736>
- Todd, P. M., & Gigerenzer, G. (2001). Putting naturalistic decision making into the adaptive toolbox. *Journal of Behavioral Decision Making*, *14*(5), 381-383. <https://doi.org/10.1002/bdm.396>
- Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124-1131. <https://doi.org/10.1126/science.185.4157.1124>
- Von Neumann, J. & Morgenstern, O. (1947). *Theory of Games and Economic Behavior*. Princeton University Press. Retrieved from <http://press.princeton.edu/chapters/i7802.pdf>

Wilson, T.D. Houston, C., Etling, K.M., & Brekke, N. (1996). A new look at anchoring effects: Basic anchoring and its antecedents. *Journal of Experimental Psychology: General*, 4, 387-402. <https://doi.org/10.1037/0096-3445.125.4.387>

Yoon, S., Fong, N. M., & Dimoka, A. (2019). The robustness of anchoring effects on preferential judgements. *Judgment & Decision Making*, 14(4), 470-487. <https://EconPapers.repec.org/RePEc:jdm:journl:v:14:y:2019:i:4:p:470-487>

Appendix 3A. Pilot study

A pre-registered pilot study was conducted to test the suitability of the “wheel of fortune” task to contrast decisions with and without moral implications and explore their susceptibility to anchoring and social influence.

Method

Participants

We recruited 50 participants (17 female, $M = 39.64$, $SD = 8.34$) via Amazon’s Mechanical Turk (www.mturk.com). Participants were paid a flat participation fee as well as a potential reward depending on the outcome from spinning the wheel of fortune their chosen number of spins. The amount won (if any) by participants in the “charity” condition (i.e., who played on behalf of a charity) was transferred to the respective chosen charity. The pilot study was self-certified in accordance with the Royal Holloway, University of London Ethics Committee procedure.

Materials and Procedure

The study was designed in Qualtrics. After providing online consent, participants were presented with the wheel of fortune task (Appendix 3B) phrased either as a personal risk or as a risk-for-charity. Respondents in the risk-for-charity condition were provided with a list of charities and had the option to select one of them after they had indicated their responses. Participants indicated their choices on a slider line; they were instructed to click on the slider bar in order for the slider cursor to appear and move the slider cursor to indicate their choices. The slider scale ranged between 0 to 140 spins, inclusive, with no minimum or maximum values displayed.

Design and Analysis

The pilot had a between-subject design with two conditions, the two levels of the factor beneficiary: personal risk and risk-for-charity. The two conditions of the pilot constitute the control condition of Study 1. The data was analysed with RStudio 3.5.1. The distribution of responses was explored with histograms and boxplots. I performed an independent samples t-test to compare the mean responses in the two conditions. No data was excluded from the analysis.

Results & Discussion

The primary goal of the pilot was to check the distribution of responses. As expected, there was no clustering of responses around certain values (Figure 1). Therefore, we deemed the task suitable to explore anchoring of choices through the prism of morality.

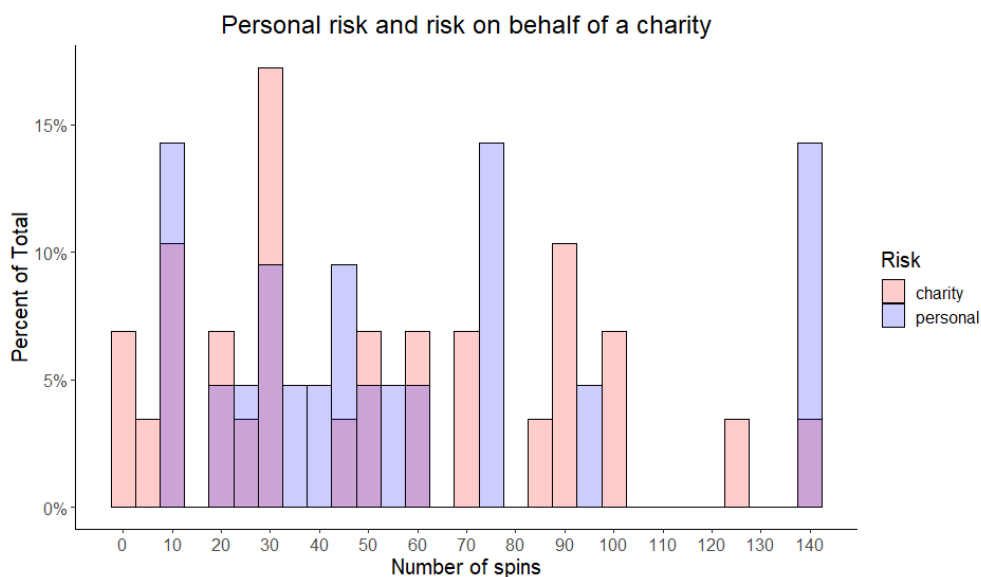


Figure 1. Histograms for the personal risk and the risk-for-charity conditions.

There was no significant difference between the average number of spins chosen in the personal risk ($M = 57.69, SD = 41.48$) and the risk-for-charity task ($M = 50.69, SD = 38.69$), $t(48) = -0.61, p > .050$. As the pilot study constitutes the control condition of Study 1, the non-significant difference between the responses in the two tasks allows potential justification of significant differences in the anchoring condition of Study 1 as stemming from the implemented anchor.

Appendix 3B. Wheel of fortune task

Instructions for one of the four experimental conditions in Study 1 (risk-for-charity with anchoring)

In this study you will play a game with real money (i.e., your decision is not just hypothetical). Please read carefully through the rules of the game and make sure you understand how your decision affects how much money can be won and what the odds of winning are. After you have made your choice in the game, we will ask you two comprehension questions to double-check your understanding of the rules. It is therefore important that you understand everything correctly.

Wheel of fortune game:

In this game, you can earn money for a charity (again, your decision is not just hypothetical – real money is at stake).

You start with a budget of 10p (£0.10). Next, you must choose how many times the computer should spin a wheel of fortune on behalf of the charity.

The wheel of fortune has one hundred spaces: **ninety-nine “good” spaces** and **one “bad” space**. On each spin, the spinner will land on one of these spaces – all spaces are equally likely to be landed on (please see image below).

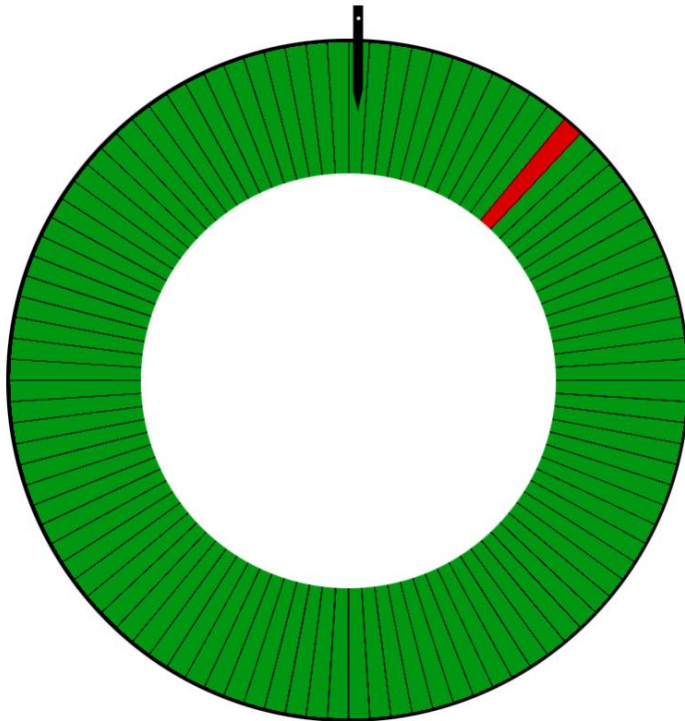


Figure 1. Wheel of fortune. *The ninety-nine “good” spaces are shown in green and the one “bad” space is shown in red.*

On **EACH** spin, the computer checks the outcome:

- If a spin lands on a “good” space, you win an extra 2p for the charity.
- If a spin lands on a “bad” space, the game is over and all the money accumulated so far (including the 10p you started with) is lost, so the charity receives no money from this game.

*The above rules apply for each spin, so the more spins you choose, the more money is earned in total if you never land on a “bad” space **but** there are also more chances to land on a “bad” space and lose everything on behalf of the charity.*

What you need to decide is how many times to spin the wheel. All spins will be

processed instantaneously by the computer, so choosing more spins will not take any more time than choosing fewer.

After you have made your decision, you can select one of the following charities to receive the money (if any):

World Wide Fund For Nature (Wildlife preservation)

Doctors Without Borders (Humanitarian medical aid)

Shelter (Housing and homelessness)

Cancer Research UK (Medical research)

Transparency International (Anti-corruption)

Oxfam (Global poverty)

Battersea Dogs & Cats Home (Animal shelter)

Please proceed to the next screen to decide how many spins you would like the computer to make.

Risk-for-charity task

anchoring condition

Please move the slider to indicate the number of times you would like the computer to spin the wheel of fortune on behalf of your chosen charity.

Please remember that on each spin, there is a ninety nine percent chance of winning an extra two pence and one percent chance of ending the game with the loss of all the money accumulated for the charity.

The slider bar is set initially at 92, the number of spins chosen on average by participants in the first round of this experiment.

Number of times the computer should spin the wheel:



no anchor condition

Please click carefully on the line below to make the slider appear, then move the slider to indicate the number of times you would like the computer to spin the wheel of fortune on behalf of your chosen charity.

Please remember that on each spin, there is a ninety nine percent chance of winning an extra two pence and one percent chance of ending the game with the loss of all the money accumulated for the charity.

Number of times the computer should spin the wheel:



Chapter 4. Study 2. Self-serving Anchoring of Personality Judgements

Study 2. Self-serving Anchoring of Personality Judgements

Elitza Ambrus

Bjoern Hartig

Ryan McKay

Submitted for publication at the British Journal of Social Psychology

Word count: 4,401 excluding references and appendices

Abstract

Human judgements are notoriously susceptible to *anchoring*, such that people's estimates of various quantities can be influenced by a salient arbitrary value. Here, we asked whether anchors could influence judgements about that with which we are most intimately acquainted – our own personal qualities. Moreover, we investigated whether we are particularly susceptible to such influences when they flatter us by enhancing our self-evaluations. Participants (N = 248) first indicated whether they ranked themselves higher or lower than 95 (high anchor) or 5 (low anchor) out of 100 other participants on eight personality characteristics. Subsequently, participants provided their specific self-rankings for each trait in comparison to others (participants in a no-anchor condition provided these specific self-rankings at the outset). The data showed that personality judgements were susceptible to anchoring in a self-serving manner: while enhancing anchors strongly impacted self-rankings, diminishing anchors had little or no influence on personality judgements. We discuss the implications of the self-serving anchoring of personality judgements for both the anchoring mechanism and the way personality traits are constructed and adjusted.

1. Introduction

Human judgements are notoriously susceptible to *anchoring*. What percentage of the United Nations are African nations? How old was Gandhi when he died? Our estimated answers to questions like these can be influenced by an arbitrary reference point (e.g., a roulette wheel set to stop at a particular number; Tversky & Kahneman, 1974; see also Strack & Mussweiler, 1997). Few of us, however, are deeply invested in our answers to such trivia questions. But what about more fundamental judgements? Might our judgements about our own personal qualities be vulnerable to anchoring? The present study extends the anchoring paradigm to the domain of the personality. We investigate whether personality judgements can be anchored by arbitrary reference points and whether participants are especially vulnerable to *self-serving* anchors, i.e., anchors that heighten their qualities. As moral traits are essential for the self-concept (Strohminger & Nichols, 2014), we also test whether any self-serving anchoring effect is especially pronounced for *moral* qualities.

The anchoring effect is a robust and replicable psychological phenomenon that has been observed across various decision-making domains (Furnham & Boo, 2010; Li et al., 2021; Strack et al., 2016; Röseler et al., 2022.; Röseler & Schütz, 2022; Yoon et al., 2019). The classical anchoring paradigm has two stages: in the first stage, respondents indicate whether their estimate would be higher or lower than a given number (which functions as the anchor) and in the second stage they provide their specific estimate (Tversky & Kahneman, 1974). For example, legal experts were asked if the sentence in a particular case should be higher or lower than a reference number (which the experts knew was randomly determined) and this random number anchored the lengths of the subsequent sentences they deemed appropriate (Englich et al., 2006).

Several theories have been developed to explain the psychological mechanisms underpinning anchoring. For example, Tversky and Kahneman (1974) proposed that individuals adjust away from the anchor value (a salient starting point) until they reach what they consider a plausible estimate. According to Tversky and Kahneman (1974),

though, the adjustment is typically insufficient and thus yields a biased estimate. Though this “insufficient adjustment” account was initially the dominant theoretical explanation, later empirical results seemed inconsistent with it (e.g., Jacowitz & Kahneman, 1995). Subsequently, Mussweiler and Strack (1999a, 1999b, 2000a, 2000b, 2001b; Bahník & Strack, 2016; Mussweiler, 2003; Mussweiler et al., 2000; Strack & Mussweiler, 1997; cf. Bahník, 2021; Harris et al., 2019) modelled anchoring as a hypothesis-confirmatory search, induced by the comparison with the anchor value and rendering anchor-consistent information “selectively accessible”. But if the anchoring effect is mediated by the accessibility of relevant information, then one might expect anchoring to be more likely when anchors are consistent with information that is selectively accessible for other reasons.

Several lines of evidence suggest that people find it easier to recall positive information about the self than negative information. For example, Ritchie et al. (2017) asked participants to generate instances of positive and negative behaviours they had previously performed. A month later they recalled more of the positive than the negative behaviours (see also Sedikides & Green, 2009; Sedikides et al., 2016).

Individuals also demonstrated much better memory for positive than negative self-referenced words (Symons & Johnson, 1997; Zhang et al., 2018). Carlson et al. (2020) found that individuals tend to recall being more generous in the past than they actually were, even when incentivised for accuracy of memory. Regarding personality judgements, Santioso et al. (1990) showed that the desirability of personality traits influences individuals’ self-rankings on those traits. These researchers manipulated the perceived desirability of personality traits (e.g., making introversion seem desirable in one condition, while making extroversion desirable in another) and provided evidence of selective autobiographical memory recall, biased towards behaviours demonstrating the desirable trait in question (Santioso et al., 1990). If positive self-relevant information is selectively accessible, then anchoring of judgements about the self might be more effective when anchors are self-serving.

To our knowledge, no previous study has investigated whether judgements of our own personal qualities are vulnerable to anchoring. There are reasons to doubt this possibility: after all, we are intimately acquainted with our selves and although researchers have argued for some degree of flexibility of personality traits throughout the lifespan, self-judgements are commonly considered stable during adulthood (Bleidorn et al., 2022; Costa & McCrae, 1992; McCrae et al., 2000; Roberts et al., 2006; Roberts & Yoon, 2022). However, previous research has provided evidence that self-relevant information can be anchored (Cheek et al. 2015; Greenberg et al., 2017; Joel et al., 2017; Mussweiler & Strack, 2000; Plous, 1989). For instance, Cheek et al. (2015) found they could anchor people's judgements of their own recent behaviours such as estimates of the number of math problems they had just solved, or the number of stairs they had just climbed. Subsequently, Joel et al. (2017) anchored individuals' probability judgements about their own futures and showed that while anchors pointing towards high probabilities of desirable events were effective, anchors that suggested high probabilities of undesirable events were not effective in influencing judgements.

Other studies also suggest the self is "selectively stable". Stanley et al. (2019) showed that even if past immoral acts are recalled, individuals report a perception of self-change and a dissociation from their past behaviour; while if previous moral deeds are recollected, agents report a perception of self-continuity and association with their past deeds (Stanley et al., 2019). Indeed, negative changes to moral traits are the most detrimental to individuals' sense of self-continuity (Molouki & Bartels, 2017) as moral traits are perceived central to individuals' self-concept (Heiphetz et al., 2016; Strohinger & Nichols, 2014). Positive changes in moral traits, however, are easily accommodated as a natural development of the fundamentally good values individuals believe they possess (Molouki & Bartels, 2017; Newmann et al., 2014; Newmann et al., 2015). People also manage to preserve a positive self-view by failing to update their self-image in light of their unethical behaviour and attributing such behaviour to contextual factors (Malle et al., 2006).

Such findings are reminiscent of self-serving biases in belief formation, that is people's predisposition to overweight desirable and underweight undesirable information when forming and updating self-relevant beliefs (Lefebvre et al., 2017; Sedikides & Skowronski, 2020; Sharot & Garrett, 2016, cf. Burton et al., 2022; Shah et al., 2016, but see also Garrett & Sharot, 2017). For example, agents readily updated their beliefs about their IQ in response to positive feedback, while being reluctant to incorporate negative feedback even when incentives for reporting accurate self-representations were in place (Eil & Rao, 2011; Möbius et al., 2022). Korn et al. (2012) also found evidence for asymmetric updating of self-beliefs in response to social feedback - positive social feedback was overweighted while negative social feedback was largely dismissed. Recent research also indicates that although individuals are generally conservative in updating their self-beliefs, the update is asymmetric, overweighting positive feedback (Möbius et al., 2022). These findings suggest that the self-image is constructed and updated in a self-serving manner, reinforcing the selective accessibility of positive self-relevant information.

In the present study, we sought to extend the anchoring paradigm to even more intimate and important self-judgements than those investigated by previous authors. Our main hypotheses were that individuals' rankings of their own personal qualities would be vulnerable to anchoring (H1), especially when anchoring is *self-serving* (H2). In particular, given that morality is fundamental to our self-concept (Heiphetz et al., 2016; Strohinger & Nichols, 2014; Molouki & Bartels, 2017) and that previous research indicates our moral preferences are sensitive to certain psychological influences (Alicke & Govorun, 2005; Brown, 1986; Meyers et al., 2019; Tappin & McKay, 2017; Zell et al., 2020), we hypothesised that judgements would be most vulnerable to self-serving anchoring of moral (vs non-moral) traits (H3)².

² For ease of exposition and interpretation, we have split the hypothesis labelled H2 in our pre-registration document into two separate secondary hypotheses, H2 and H3 (Although H2 underpins H3 as a rationale, this was left implicit in our pre-registration).

2. Method

2.1. Overview

All participants ranked themselves in comparison to 100 other anonymous participants on eight personality traits. Participants were randomly assigned to one of three between-subject experimental conditions (*anchor*: high, low or no anchor). The high and low anchor conditions involved two stages: in the first stage, for each trait, participants indicated whether they would rank themselves above or below at least 95 (high anchor condition) or at least 5 (low anchor condition) out of the 100 other participants. In the second stage, respondents provided their specific self-ranking for the relevant trait, again in comparison to the 100 other participants. Participants in the no anchor condition were only asked to indicate their specific self-ranking for each personality characteristic. The eight personality traits were balanced along the levels of two within-subject variables: *morality* (moral, non-moral) and *desirability* (desirable, undesirable). After indicating their self-rankings on all personality traits, respondents answered an attention check question (asking them to place the slider bar at the midpoint of the slider scale) and provided basic demographic data (age and gender).

2.1.1. Transparency and openness

Our hypotheses, data collection, and analysis protocol were pre-registered (https://aspredicted.org/QJH_H3Z), including the criteria for data exclusions. We report all manipulations, and all measures in the study, and we follow JARS (Kazak, 2018). De-identified data, analysis scripts and study materials are available on the Open Science Framework: https://osf.io/x8rqz/?view_only=a919b539fa224bfca48d2cd3428dc82a. Data were analysed using R Studio 4.1.0 (R Development Core Team, 2021) and the lme4 package (Bates et al., 2015) for the linear mixed-effects models. We used the afex package (Singmann et al., 2018) as well as the emmeans (Lenth, 2018), multcomp (Hothorn et al., 2011) and MOTE (Buchanan et al., 2019) packages for post-hoc binary comparisons, *p*-value adjustments and effect sizes. The figures were created in Python

(Van Rossum & Drake, 2009) using the Pandas (McKinney et al., 2010), Numpy (Harris et al., 2020) and Seaborn (Waskom et al., 2017) libraries.

2.2. Participants

We recruited 360 participants via the online platform *Prolific* (www.prolific.com). All participants were monolingual English speakers with UK nationality and had a *Prolific* approval rate higher than 90%. However, after data collection ceased, we found that 2 submissions were incomplete, leaving us with data from 358 participants to analyse. As specified in our pre-registration document we excluded respondents who: (i) failed the attention check question ($N=50$); (ii) provided inconsistent answers in the anchoring paradigm (e.g., participants who stated that their self-ranking for a given trait was below 95 in the first stage of the anchoring paradigm, but who then indicated a self-ranking higher than 95 in the second stage, suggesting they were confused or inattentive; $N=60$) or (iii) completed the study in less than 30 seconds ($N=0$).

The final sample comprised of 248 participants (156 females, 88 males, 3 other and 1 who preferred not to specify their gender; $M = 35.96$, $SD = 13.89$). As the second exclusion criterion was not applicable for the no anchor condition, we had less data exclusion there than in the low and high anchor conditions. The final distribution of participants across anchor conditions was 71 in the high anchor condition, 70 in the low anchor condition and 107 in the no anchor condition. Participants were paid a flat participation fee (70p, the equivalent of £7.50 per hour). The study was self-certified in accordance with the Anonymous University, Ethics Committee procedure.

2.3. Materials and Procedure

The study was Qualtrics-based and conducted online. *Desirability* (desirable and undesirable) and *morality* (moral or non-moral) were counterbalanced across the personality traits, and we selected moral and non-moral traits with comparable average desirability ratings (Tappin & McKay, 2017). The characteristics thus comprised two desirable moral traits (*honest, fair*), two undesirable moral traits (*manipulative,*

deceptive), two desirable non-moral traits (*competent, knowledgeable*) and two undesirable non-moral traits (*lazy, illogical*). The personality traits were presented in a separate random order for each participant. Self-rankings were indicated on a slider, ranging from 0 to 100 (the numeric value was displayed above the slider bar and changed accordingly when participants moved the slider cursor).

Following the classical anchoring paradigm (Tversky & Kahneman, 1974), for each personality characteristic participants were first asked whether they would rank themselves higher than at least 95 (high anchor condition) or 5 (low anchor condition) other participants (out of 100). Subsequently, respondents provided their specific self-ranking for each personality trait, in comparison to 100 other participants. The slider bar in the high and the low anchoring conditions was pre-set at 95 or 5, respectively. In the no anchor condition, participants only provided their self-rankings without any preceding comparative question; in order not to bias their response, the slider cursor was initially invisible in this condition, and participants were instructed to click on the slider bar for the slider cursor to appear and to then move it to indicate their self-rankings.

2.4. Design and Analysis

The experiment employed a 3 (*anchor*: high, low, no anchor; between-subjects) x 2 (*desirability*: desirable, undesirable; within-subjects) x 2 (*morality*: moral, non-moral; within-subjects) mixed design. The DV was self-ranking on the eight personality traits (measured on a 0 to 100 scale) and grouped along the levels of *morality* and *desirability*. We deviated from our pre-registration in two minor respects: first, although we pre-registered an ANOVA analysis, we deemed it more appropriate to fit linear mixed-effects models to the self-rankings measure as our final sample was unbalanced and the two methods are conceptually equivalent³. Second, for the analysis of H2 and H3 we

³ For the sake of completeness, we report the conceptually equivalent pre-registered ANOVA analysis, which yielded qualitatively identical results, in Appendix A.

recoded the data for clarity – reversing the self-ranking values for undesirable traits and constructing dummies for *enhancing* and *diminishing* manipulations.

3. Results

Anchoring influenced judgements of participants’ own personal qualities

To test our main hypothesis (H1) that self-judgements would be subject to anchoring effects we fitted a simple linear mixed-effects model (Model 1) on the self-ranking measure with *anchor* modelled as a fixed effect and *participants* as a random effect. The results of Model 1 demonstrated an overall anchoring effect on self-judgements. Compared to participants in the control condition ($M = 45.83, SD = 25.38$), subjects in the high anchor condition ranked themselves higher ($M = 50.45, SD = 31.09$) whereas those in the low anchor condition ranked themselves lower ($M = 39.82, SD = 29.71$). Both differences were statistically significant (Table 1, Model 1; see also Figure 1).

Table 2. Estimated fixed effects for Model 1 (depicting the effect of anchors on self-rankings).

	Model 1
Intercept	45.83*** (1.13)
High Anchor	4.62* (1.78)
Low Anchor	-6.01*** (1.79)

Baseline level: no anchor; Number of observations: 1984; grouped by participants, N=248 Significance codes: '***' 0.001; '**' 0.01; '*' 0.05;

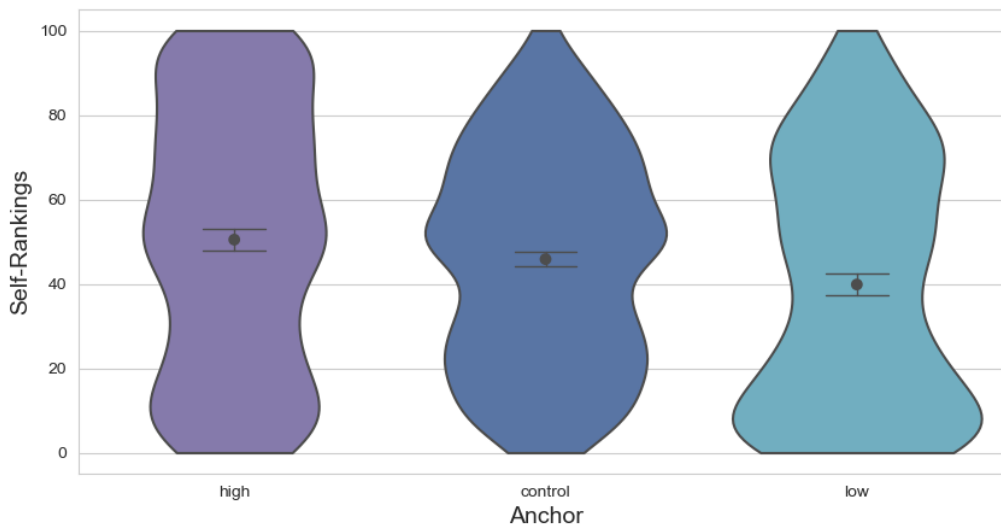


Figure 1. Violin plots of self-rankings by *anchor*. Error bars represent 95% confidence intervals (CI).

Anchoring influenced self-judgements in a self-serving manner, though this was not more pronounced for moral traits

To examine H2, we first reversed the self-ranking values for undesirable traits (i.e., we subtracted them from 100, such that 0 became 100, 25 became 75, etc.) so that higher values for all traits represent a more positive self-view. Second, we replaced the high and low anchor dummy variables with dummies for *enhancing* and *diminishing* anchors. The *enhancing* dummy applies to the combination of high anchor with desirable traits and low anchor with undesirable traits, i.e., when the anchor pulls the self-ranking towards a more positive self-perception. Conversely, the *diminishing* dummy applies when the anchor pulls the self-ranking towards a more negative self-view, i.e., high anchor with undesirable traits and low anchor with desirable traits.

Table 2, Model 2 reports the results of a linear mixed-effects regression predicting the recoded self-rankings with *enhancing* and *diminishing* anchor dummy variables as fixed effects and participants as a random effect (please see also Figure 2 and Figure 3). Model 3 adds *morality* as a fixed effect and Model 4 adds the interactions of *morality*

with the *anchor* dummies. The results show that the effect of the *enhancing* anchors is positive and highly significant, but we cannot reject the null hypothesis for the *diminishing* anchors. Therefore, self-rankings were especially vulnerable to anchors promoting an enhanced self-view, which confirms H2. Such “self-serving anchoring” was not, however, more pronounced for moral (vs non-moral) traits (H3), as the dummy for moral traits did not interact with the dummies for *enhancing* and *diminishing* anchors. There was though a significant main effect of moral traits, indicating that participants ranked themselves more positively for morally relevant than morally irrelevant traits.

Table 2. Estimated fixed effects for Models 2, 3 and 4 (depicting the effect of *enhancing* and *diminishing* anchors, *morality* and their interaction on self-rankings).

	Model 2	Model 3	Model 4
Intercept	64.35*** (1.13)	60.22*** (1.23)	60.95*** (1.35)
Enhancing Anchor	8.75*** (1.63)	8.75*** (1.63)	7.75*** (2.00)
Diminishing Anchor	-1.81 (1.63)	-1.81 (1.63)	-3.36 (2.00)
Morality (moral traits)		8.24*** (0.96)	6.79*** (1.47)
Enhancing * Morality (moral traits)			2.00 (2.33)
Diminishing * Morality (moral traits)			3.09 (2.33)

Baseline level: no anchor, non-moral; Number of observations: 1984; grouped by participants, N=248.
Significance codes: '***' 0.001; '**' 0.01; '*' 0.05;

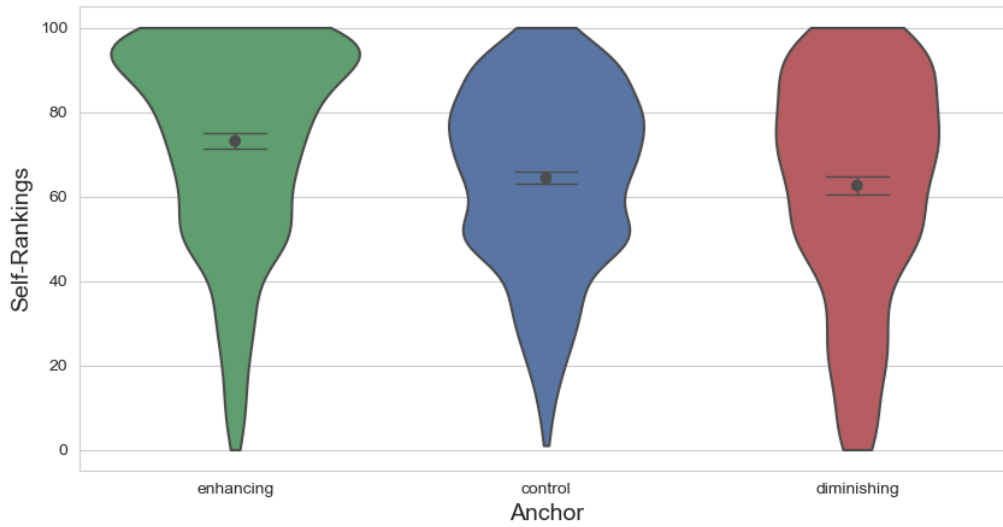


Figure 2. Violin plots of self-rankings by *anchor*. Error bars represent 95% confidence intervals (CI).

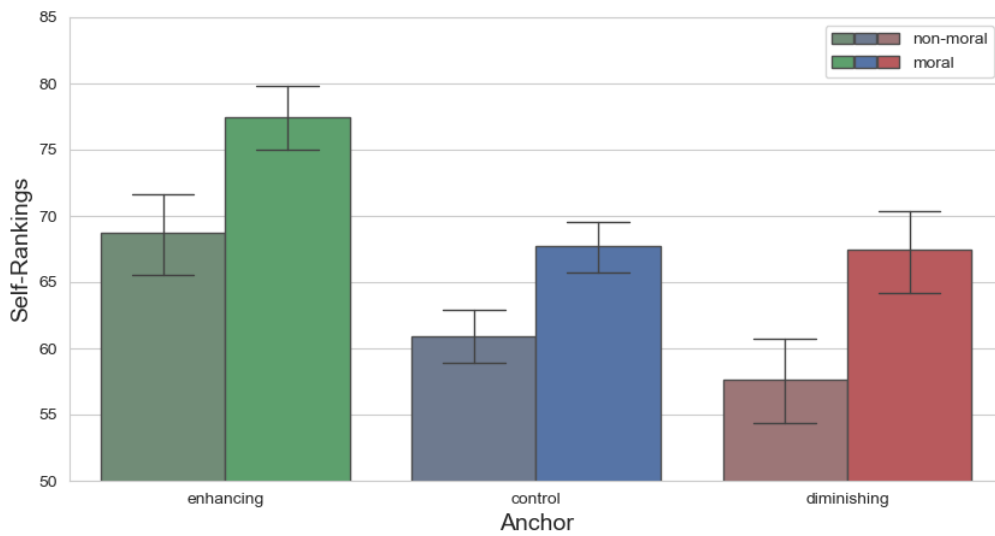


Figure 3. Self-rankings by *anchor* and *morality*. Error bars represent 95% confidence intervals (CI).

4. Discussion

Anchoring effects have been documented across a range of domains, including sentencing decisions (Englich et al., 2006), negotiations (Galinsky & Mussweiler, 2001), judgements of recent behaviour (Cheek et al., 2015) and judgements of future prospects

(Joel et al., 2017). Here we sought evidence of whether anchors could distort judgements about that with which we are most intimately acquainted: our own personal qualities. Our data revealed that such self-judgements are indeed susceptible to anchoring. Moreover, accounting for the effect on the self-image revealed a self-serving element: while enhancing anchors had a strong impact on self-rankings, diminishing anchors had little or no influence on self-judgements. This “self-serving anchoring” was not more pronounced for judgements about moral traits, although participants did rank themselves more positively on moral than non-moral traits, consistent with previous research on the importance of morality to the self-concept (Molouki & Bartels, 2017; Strohminger & Nichols, 2014) and the magnified effect of cognitive biases in the moral domain (Alicke et al., 2001; Brown, 2012; Tappin & McKay, 2017).

Our findings provide further evidence for the robustness of the anchoring effect across judgement and decision-making contexts (e.g., Yoon et al., 2019), and extend the anchoring paradigm to a novel domain. The self-serving anchoring effect we document has several interesting implications. First, the fact that judgements of one’s own personal qualities can be anchored at all has implications for the stability of self-judgements, as personality traits are commonly considered to be fixed in adulthood (e.g., Costa & McCrae, 1992) and can serve as a valid predictor of important life outcomes (Roberts et al., 2007; Beck & Jackson, 2022; Soto, 2021). Our results are consistent with recent evidence that personality traits retain a certain degree of flexibility in adulthood (Bleidorn et al., 2022; Roberts et al., 2007; Soto, 2021) and that they change in response to relatively lasting nonclinical psychological interventions (Bleidorn et al., 2022; Stieger et al., 2020). Our findings, however, demonstrate a more striking flexibility: self-judgements can instantly adjust in response to a salient *enhancing* value, which points towards somewhat different psychological mechanisms shaping the construction of self-judgements than those currently theorised (Roberts & Yoon, 2022).

The differential flexibility of self-judgements in the face of anchors of different types (being readily adjusted in response to enhancing anchors while remaining relatively stable in response to diminishing anchors) echoes existing research on asymmetric updating of self-relevant beliefs: the self-image seems malleable insofar as it quickly updates in response to positive social feedback, but exhibits stability in response to negative social feedback, which is mostly neglected (Korn et al., 2012). Self-representations are also positively distorted in memory, which helps maintain a stable positive self-view (Carlson et al., 2020; Zhang et al., 2018). In a study on anchoring of participants' own recent previous behaviour (e.g., the number of math problems they had just solved), Cheek et al. (2015) found that only a high anchor impacted these judgements. Cheek et al. suggested that a floor effect may have limited the effectiveness of the low anchors, but another possibility is that high anchors may have been more self-serving. Consistent with this interpretation, Joel et al. (2017) subsequently showed that when anchoring individuals' probability judgements about their own futures, anchors suggesting a high probability of undesirable prospects were ineffective.

Alongside the implications for the stability of self-judgements, our results have more general implications for theories of anchoring. In conjunction with the results of Joel et al. (2017), the fact that diminishing anchors had little to no effect on subsequent self-judgements in our study suggests that the undesirability of an anchor limits its effectiveness, implying a strong self-serving bias in anchoring. A recent dynamic meta-analysis showed that factors such as monetary incentives for participation or accuracy, type of experiment (online or lab), and demographic factors such as age or gender do not influence the magnitude of the anchoring effect (Röseler & Schütz, 2022). Moreover, previous research suggests that anchoring persists even when anchors are irrelevant, extreme or incompatible with the estimate (Glöckner & Englich, 2015; Mussweiler, 2001b; Strack & Mussweiler, 1997; Röseler & Schütz, 2022) as well as when anchor-inconsistent information is considered prior to the estimate (Mussweiler et al., 2000). Our findings suggest however that considerations of maintaining a positive self-

image may attenuate or eliminate the effect of anchoring when it comes to judgements about one's own personal qualities.

Although our study was not designed with the aim of differentiating between different theoretical accounts of anchoring, evidence of self-serving anchoring might shed additional light on the psychological mechanisms that underpin the anchoring process. If anchoring results from adjustment of the anchor value (Tversky & Kahneman, 1974) or from numeric priming (Wong & Kwong, 2000), individuals' self-rankings in our study should be influenced to a similar degree by the numeric value of the anchor, regardless of the anchor's desirability (enhancing or diminishing). However, if the anchor values elicit a search for anchor-consistent information (Mussweiler & Strack, 1999a), the enhanced accessibility of positive self-relevant information would produce such a seemingly self-serving effect (Ritchie et al., 2017; Sedikides & Green, 2009; Santoso et al., 1990; Sedikides et al., 2016; Zhang et al., 2018).

Furthermore, all current theories of anchoring assume some sort of process of assimilation towards the anchor value (Strack et al., 2016). This assimilation, however, may also be biased in a seemingly self-serving way. Pinter et al. (2011) theorised that positive self-relevant information is assimilated into self-knowledge (rendering it selectively accessible) while negative self-relevant information is separated/contrasted away from stored self-knowledge. Therefore, enhancing and diminishing anchors might be processed via different psychological mechanisms stemming from the way self-relevant information is integrated and stored: enhancing anchors would tap into highly accessible positive self-relevant information, while diminishing anchors would be ineffective as negative self-relevant information is contrasted away from self-knowledge and thus not easily accessible.

Our results are also consistent with similarity vs dissimilarity testing and subsequent selective accessibility (Hanko et al., 2010; Mussweiler, 2003; Mussweiler, 2001a; Mussweiler & Strack, 2000). According to this proposition, individuals in our study go through an initial stage of assessing how similar they are to the proposed

anchor. The outcome of this assessment determines whether knowledge confirming similarity or dissimilarity is activated and this knowledge is the basis for the subsequent evaluation (Mussweiler, 2003). Given that individuals tend to see themselves as above average (Alicke & Govorun, 2005), subjects faced with a diminishing anchor are more likely to search for dissimilarities with the anchor and activate knowledge inconsistent with the anchor, rendering it largely ineffective. In contrast, an enhancing anchor is likely to induce a similarity search, resulting in the activation of self-positive knowledge, and thereby enhancing the subsequent self-judgement.

In conclusion, our results show that judgements of one's own personal qualities are susceptible to anchoring; in line with previous research on motivated processing of self-relevant information, the data revealed that anchoring is strong when anchors enhance the self-image while diminishing anchors have little to no effect on self-rankings. Our results show that personality judgements exhibit both flexibility and stability: flexibility in response to enhancing adjustments and stability/resistance towards diminishing adjustments. The self-serving anchoring we observed suggests that the selective accessibility of positive self-relevant information might be an important factor that simultaneously limits the effectiveness of (diminishing) anchoring and allows personality traits to adapt instantly to salient *enhancing* values.

Statement of Contribution

What is already known on this subject?

Previous research has shown that personality judgements are stable in middle adulthood and serve as a valid predictor of life outcomes (Beck & Jackson, 2022; Roberts et al., 2007; Soto, 2021). At the same time, the anchoring effect is a robust and replicable psychological phenomenon that has been observed across various decision-making domains (e.g., Röseler & Schütz, 2022). The anchoring effect refers to the impact of a previously considered comparative value (anchor) on a subsequent absolute judgment and reveals that individuals' estimates of various quantities are flexible and can be influenced by a salient random value (Tversky & Kahneman, 1974).

What does this study add?

- Showed that personality judgements can be influenced by salient arbitrary reference points (“anchors”).
- Showed that we are especially influenced by self-serving anchors, i.e., anchors pointing in a flattering direction.

References

- Alicke, M. D., & Govorun, O. (2005). The better-than-average effect. *The Self in Social Judgment*, 1, 85-106. https://doi.org/10.1007/978-3-319-24612-3_300293
- Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. *arXiv preprint arXiv:1506.04967 [stat.ME]*.
<https://doi.org/10.48550/arXiv.1506.04967>
- Bahník, Š. (2021). Anchoring does not activate examples associated with the anchor value. Advance online publication. *PsyArXiv*.
<https://doi.org/10.31234/osf.io/4j5wb>
- Bahník, Š., & Strack, F. (2016). Overlap of accessible information undermines the anchoring effect. *Judgment and Decision Making*, 11(1), 92-98.
<https://doi.org/10.1017/S1930297500007610>
- Beck, E. D., & Jackson, J. J. (2022). A mega-analysis of personality prediction: Robustness and boundary conditions. *Journal of Personality and Social Psychology*, 122(3), 523– 553. <https://doi.org/10.1037/pspp0000386>
- Bleidorn, W., Hopwood, C. J., Back, M. D., Denissen, J. J., Hennecke, M., Hill, P. L., ... & Zimmermann, J. (2021). Personality trait stability and change. *Personality Science*, 2, 1-20. <https://doi.org/10.5964/ps.6009>
- Bleidorn, W., Schwaba, T., Zheng, A., Hopwood, C. J., Sosa, S. S., Roberts, B. W., & Briley, D. A. (2022). Personality stability and change: A meta-analysis of longitudinal studies. *Psychological bulletin*, 148(7-8), 588
<https://doi.org/10.1037/bul0000365>
- Brown, J. D. (1986). Evaluations of self and others: Self-enhancement biases in social judgements. *Social cognition*, 4(4), 353-376.
<https://doi.org/10.1521/SOCO.1986.4.4.353>

- Brown, J. D. (2012). Understanding the better than average effect: Motives (still) matter. *Personality and Social Psychology Bulletin*, 38(2), 209-219.
<https://doi.org/10.1177/0146167211432763>
- Buchanan, E., Gillenwaters, A., Scofield, J., Valentine, K. (2019). *MOTE: Measure of the Effect: Package to assist in effect size calculations and their confidence intervals*. R package version 1.0.2. <http://github.com/doomlab/MOTE>.
- Burton, J. W., Harris, A. J., Shah, P., & Hahn, U. (2022). Optimism where there is none: asymmetric belief updating observed with valence-neutral life events. *Cognition*, 218, 104939.
<https://doi.org/10.1016/j.cognition.2021.104939>
- Carlson, R. W., Maréchal, M. A., Oud, B., Fehr, E., & Crockett, M. J. (2020). Motivated misremembering of selfish decisions. *Nature Communications*, 11(1), 1-11.
<https://doi.org/10.1038/s41467-020-15602-4>
- Cheek, N. N., Coe-Odess, S., & Schwartz, B. (2015). What have I just done? Anchoring, self-knowledge, and judgements of recent behavior. *Judgment and Decision Making*, 10, 76–85.
<https://EconPapers.repec.org/RePEc:jdm:journl:v:10:y:2015:i:1:p:76-85>
- Costa Jr, P. T., & McCrae, R. R. (1992). Four ways five factors are basic. *Personality and Individual Differences*, 13(6), 653-665. [https://doi.org/10.1016/0191-8869\(92\)90236-I](https://doi.org/10.1016/0191-8869(92)90236-I)
- Eil, D., & Rao, J. M. (2011). The good news-bad news effect: asymmetric processing of objective information about yourself. *American Economic Journal: Microeconomics*, 3(2), 114-38. <https://doi.org/10.1257/mic.3.2.114>
- Englich, B. (2008). When knowledge matters—differential effects of available knowledge in standard and basic anchoring tasks. *European Journal of Social Psychology*, 38(5), 896-904. <https://doi.org/10.1002/ejsp.479>

- Englich, B., & Mussweiler, T. (2001). Sentencing under uncertainty: Anchoring effects in the courtroom. *Journal of Applied Social Psychology, 31*, 1535–1551.
<https://doi.org/10.1111/j.1559-1816.2001.tb02687.x>
- Englich, B., Mussweiler, T., & Strack, F. (2006). Playing dice with criminal sentences: The influence of irrelevant anchors on experts' judicial decision making. *Personality and Social Psychology Bulletin, 32*, 188–200.
<https://doi.org/10.1177/0146167205282152>
- Englich, B., & Soder, K. (2009). Moody experts---How mood and expertise influence judgmental anchoring. *Judgment and Decision making, 4*(1), 41-50.
<https://EconPapers.repec.org/RePEc:jdm:journl:v:4:y:2009:i:1:p:41-50>
- Freira, L., Sartorio, M., Boruchowicz, C., Boo, F. L., & Navajas, J. (2020). The irrational interplay between partisanship, beliefs about the severity of the COVID-19 pandemic, and support for policy interventions.
<https://doi.org/10.31234/osf.io/4cgfw>
- Frederick, S. W., & Mochon, D. (2012). A scale distortion theory of anchoring. *Journal of Experimental Psychology: General, 141*(1), 124-133.
<https://doi.org/10.1037/a0024006>
- Furnham, A., & Boo, H. C. (2011). A literature review of the anchoring effect. *The Journal of Socio-Economics, 40*(1), 35-42. <https://doi.org/10.1016/j.socec.2010.10.008>
- Galinsky, A. D., & Mussweiler, T. (2001). First offers as anchors: the role of perspective taking and negotiator focus. *Journal of Personality and Social Psychology, 81*(4), 657-669. <https://doi.org/10.1037/0022-3514.81.4.657>
- Garrett, N., & Sharot, T. (2017). Optimistic update bias holds firm: Three tests of robustness following Shah et al. *Consciousness and Cognition, 50*, 12–22.
<https://doi.org/10.1016/j.concog.2016.10.013>
- Glöckner, A., & Englich, B. (2015). When relevance matters. *Social Psychology, 46*(1), 4-12. <https://doi.org/10.1027/1864-9335/a000214>

- Greenberg, D. L., Bishara, A. J., & Mugayar-Baldocchi, M. A. (2017). Anchoring effects on early autobiographical memories. *Memory*, 25(9), 1303-1308.
<https://doi.org/10.1080/09658211.2017.1297833>
- Harris, A. J., Blower, F. B., Rodgers, S. A., Lagator, S., Page, E., Burton, A., Urlichich, D. & Speekenbrink, M. (2019). Failures to replicate a key result of the selective accessibility theory of anchoring. *Journal of Experimental Psychology: General*, 148(9), e30-e50. <https://doi.org/10.1037/xge0000644>
- Hanko, K., Crusius, J., & Mussweiler, T. (2010). When I and me are different: assimilation and contrast in temporal self-comparisons. *European Journal of Social Psychology*, 40(1), 160-168. <https://doi.org/10.1002/ejsp.625>
- Harris, C. R. et al. (2020). Array programming with NumPy. *Nature*, 585, 357–362.
<https://doi.org/10.1038/s41586-020-2649-2>.
- Heiphetz, L., Strohminger, N., & Young, L. L. (2017). The role of moral beliefs, memories, and preferences in representations of identity. *Cognitive science*, 41(3), 744-767.
<https://doi.org/10.1111/cogs.12354>
- Hothorn, T., Bretz, F., Westfall, P., Heiberger, R. M., & Schuetzenmeister, A. (2011). Multcomp: simultaneous inference in general parametric models. R package version 1.2. Available at: <https://cran.r-project.org/web/packages/multcomp/index.html>
- Jacowitz, K. E., & Kahneman, D. (1995). Measures of anchoring in estimation tasks. *Personality and Social Psychology Bulletin*, 21(11), 1161-1166.
<https://doi.org/10.1177/01461672952111004>
- Joel, S., Spielmann, S. S., & MacDonald, G. (2017). Motivated use of numerical anchors for judgements relevant to the self. *Personality and Social Psychology Bulletin*, 43(7), 972-985. <https://doi.org/10.1177/0146167217702613>
- Kazak, A. E. (2018). Editorial: Journal article reporting standards. *American Psychologist*, 73(1), 1–2. <https://doi.org/10.1037/amp0000263>

- Korn, C. W., Prehn, K., Park, S. Q., Walter, H., & Heekeren, H. R. (2012). Positively biased processing of self-relevant social feedback. *The Journal of Neuroscience*, 32(47), 16832–16844. <https://doi.org/10.1523/JNEUROSCI.3016-12.2012>
- Lefebvre, G., Lebreton, M., Meyniel, F., Bourgeois-Gironde, S., & Palminteri, S. (2017). Behavioural and neural characterization of optimistic reinforcement learning. *Nature Human Behaviour*, 1(4), 1-9. <https://doi.org/10.1038/s41562-017-0067>
- Lenth, R. (2018). emmean: Estimated Marginal Means, aka Least-Squares Means. R Package Version 1.2.3. Available at: <https://CRAN.R-project.org/package=emmeans>.
- Li, L., Maniadis, Z., & Sedikides, C. (2021). Anchoring in economics: a meta-analysis of studies on willingness-to-pay and willingness-to-accept. *Journal of Behavioral and Experimental Economics*, 90, 101629. <https://doi.org/10.1016/j.socec.2020.101629>.
- Lieder, F., Griffiths, T. L., Huys, Q. J., & Goodman, N. D. (2018). Empirical evidence for resource-rational anchoring and adjustment. *Psychonomic Bulletin & Review*, 25(2), 775-784. <https://doi.org/10.3758/s13423-017-1288-6>
- McCrae, R. R., Costa, P. T. J., Jr., Ostendorf, F., Angleitner, A., Hrebícková, M., Avia, M. D., Sanz, J., Sánchez-Bernardos, M. L., Kusdil, M. E., Woodfield, R., Saunders, P. R., & Smith, P. B. (2000). Nature over nurture: Temperament, personality, and life span development. *Journal of Personality and Social Psychology*, 78(1), 173–186. <https://doi.org/10.1037/0022-3514.78.1.173>
- McKinney, W. et al. (2010). Data structures for statistical computing in Python. *Proceedings of the 9th Python in Science Conference*, 445, 51–56). <https://doi.org/10.25080/Majora-92bf1922-00a>

- Malle, B. F. (2006). The actor-observer asymmetry in attribution: A (surprising) meta-analysis. *Psychological Bulletin*, 132(6), 895-919. <https://doi.org/10.1037/0033-2909.132.6.895>
- Meyers, E. A., Białek, M., Fugelsang, J. A., Koehler, D. J., & Friedman, O. (2019). Wronging past rights: The sunk cost bias distorts moral judgment. *Judgment and Decision Making*, 14(6), 721-727. <https://EconPapers.repec.org/RePEc:jdm:journl:v:15:y:2020:i:6:p:909-925>
- Molouki, S., & Bartels, D. M. (2017). Personal change and the continuity of the self. *Cognitive Psychology*, 93, 1-17. <https://doi.org/10.1016/j.cogpsych.2016.11.006>
- Möbius, M. M., Niederle, M., Niehaus, P., & Rosenblat, T. S. (2022). Managing self-confidence: Theory and experimental evidence. *Management Science* 68(11). <https://doi.org/10.1287/mnsc.2021.4294>
- Mussweiler, T. (2003). Comparison processes in social judgment: mechanisms and consequences. *Psychological review*, 110(3), 472. <https://doi.org/10.1037/0033-295X.110.3.472>
- Mussweiler, T. (2001a). 'Seek and ye shall find': Antecedents of assimilation and contrast in social comparison. *European journal of social psychology*, 31(5), 499-509. <https://doi.org/10.1002/ejsp.75>
- Mussweiler, T. (2001b). The durability of anchoring effects. *European Journal of Social Psychology*, 31, 431-442. <https://doi.org/10.1002/ejsp.52>
- Mussweiler, T., & Strack, F. (2000). The "relative self": Informational and judgmental consequences of comparative self-evaluation. *Journal of personality and social psychology*, 79(1), 23. <https://doi.org/10.1037/0022-3514.79.1.23>
- Mussweiler, T., & Strack, F. (1999a). Hypothesis-consistent testing and semantic priming in the anchoring paradigm: A selective accessibility model. *Journal of Experimental Social Psychology*, 35(2), 136-164. <https://doi.org/10.1006/jesp.1998.1364>

- Mussweiler, T., & Strack, F. (1999b). Comparing is believing: A selective accessibility model of judgmental anchoring. *European review of social psychology*, 10(1), 135-167. <https://doi.org/10.1080/14792779943000044>
- Mussweiler, T., & Strack, F. (2000a). The use of category and exemplar knowledge in the solution of anchoring tasks. *Journal of Personality and Social Psychology*, 78, 1038- 1052. <https://doi.org/10.1037/0022-3514.78.6.1038>
- Mussweiler, T., & Strack, F. (2000b). Numeric judgements under uncertainty: The role of knowledge in anchoring. *Journal of Experimental Social Psychology*, 36(5), 495-518. <https://doi.org/10.1006/jesp.1999.1414>
- Mussweiler, T., & Strack, F. (2001). The semantics of anchoring. *Organizational Behavior and Human Decision Processes*, 86(2), 234-255. <https://doi.org/10.1006/obhd.2001.2954>
- Mussweiler, T. (2003). Comparison processes in social judgment: mechanisms and consequences. *Psychological review*, 110(3), 472. <https://doi.org/10.1037/0033-295X.110.3.472>
- Mussweiler, T., Strack, F., & Pfeiffer, T. (2000). Overcoming the inevitable anchoring effect: Considering the opposite compensates for selective accessibility. *Personality and Social Psychology Bulletin*, 26(9), 1142-1150. <https://doi.org/10.1177/01461672002611010>
- Newman, G. E., Bloom, P., & Knobe, J. (2014). Value judgements and the true self. *Personality and Social Psychology Bulletin*, 40(2), 203–216. <https://doi.org/10.1177/0146167213508791>
- Newman, G. E., De Freitas, J., & Knobe, J. (2015). Beliefs about the true self explain asymmetries based on moral judgment. *Cognitive science*, 39(1), 96-125. <https://doi.org/10.1111/cogs.12134>
- Northcraft, G. B., & Neale, M. A. (1987). Experts, amateurs, and real estate: An anchoring-and-adjustment perspective on property pricing decisions.

Organizational Behavior and Human Decision Processes, 39, 84–97.

[https://doi.org/10.1016/0749-5978\(87\)90046-X](https://doi.org/10.1016/0749-5978(87)90046-X)

Olaru, G., Stieger, M., Rügger, D., Kowatsch, T., Flückiger, C., Roberts, B. W., & Allemand, M. (2022). Personality change through a digital-coaching intervention: Using measurement invariance testing to distinguish between trait domain, facet, and nuance change. *European Journal of Personality*.

<https://doi.org/10.1177/089020702211450>

Pinter, B., Green, J. D., Sedikides, C., & Gregg, A. P. (2011). Self-protective memory: Separation/integration as a mechanism for mnemonic neglect. *Social Cognition*, 29(5), 612-624. <https://doi.org/10.1521/soco.2011.29.5.612>

Plous, S. (1989). Thinking the unthinkable: The effects of anchoring on likelihood estimates of nuclear war. *Journal of Applied Social Psychology*, 19(1), 67–91. <https://doi.org/10.1111/j.1559-1816.1989.tb01221.x>

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>

Ritchie, T. D., Sedikides, C., & Skowronski, J. J. (2017). Does a person selectively recall the good or the bad from their personal past? It depends on the recall target and the person's favourability of self-views. *Memory*, 25(8), 934-944.

<https://doi.org/10.1080/09658211.2016.1233984>

Roberts, B. W., Walton, K. E., & Viechtbauer, W. (2006). Patterns of meanlevel change in personality traits across the life course: A meta-analysis of longitudinal studies.

Psychological Bulletin, 132(1), 1–25. <https://doi.org/10.1037/0033-2909.132.1.1>

Roberts, B. W., Kuncel, N. R., Shiner, R., Caspi, A., & Goldberg, L. R. (2007). The power of personality: The comparative validity of personality traits, socioeconomic status,

and cognitive ability for predicting important life outcomes. *Perspectives on Psychological Science*, 2(4), 313-345. <https://www.jstor.org/stable/40212212>

Roberts, B. W., & Yoon, H. J. (2022). Personality psychology. *Annual Review of Psychology*, 73(1), 489–516. <https://doi.org/10.1146/annurevpsych-020821-114927>

Röseler, L., Böglér, H. L., Koßmann, L., Krueger, S., Bickenbach, S., Bühler, R., ... Sing, J. (2022, April 13). Replicating Epley and Gilovich: Need for Cognition, Cognitive Load, and Forewarning do not Moderate Anchoring Effects. *PsyArXiv*. <https://doi.org/10.31234/osf.io/bgp3m>

Röseler, L., & Schütz, A. (2022, March 9). Hanging the Anchor Off a New Ship: A Meta-Analysis of Anchoring Effects. *PsyArXiv*. <https://doi.org/10.31234/osf.io/wf2tn>

Röseler, L., Weber, L., Helgerth, K. A. C., Stich, E., Günther, M., Tegethoff, P., Wagner, F. S., Ambrus, E., Antunovic, M., Barrera, F., Halali, E., Ioannidis, K., McKay, R., Milstein, N., Molden, D. C., Papenmeier, F., Rinn, R., Schreiter, M. L., Zimdahl, M., Allen, E., Bahník, S., Baumeister, R. F., Bermeitinger, C., Bickenbach, S. L. C., Blank, P. A., Blower, F. B. N., Böglér, H. L., Boo, F. L., Boruchowicz, C., Bühler, R. L., Burgmer, P., Cheek, N., N., Dohle, S., Dorsch, L., Dück, M. S., Fels, S.-A., Fischer, A. L., Frech, M.-L., Freira, L., Friedinger, K., Genschow, O., Harris, A., Hartig, B., Häusser, J. A., Hedgebeth, M., Henkel, M., Horvath, D., Hügel, J. C., Igna, E. L. E., Imhoff, R., Intelmann, P., Karg, A. H., Klamar, A., Klein, C., Klusmann, B., Knappe, E., Köppel, L.-M., Koßmann, L., Kraft, P., Kroworsch, M. K., Krueger, S. M., Kühling, S., Lagator, S., Lammers, J., Loschelder, D. D., Navajas, J., Norem, J., K., Novak, J. Onuki, Y., Page, E., Panse, F., Pavlovic, Z., Pearton, J., Rebholz, T. R., Rodgers, S., Röseler, J. J., Rostekova, A., Roßmaier, K. V., Sartorio, M., Scheelje, L., Schindler, S., Schreiner, N. B., Seida, C., Shanks, D. R., Siems, M.-C., Stitz, M., Starkulla, M., Stäglich, M., Thies, K., Thum, E., Undorf, M., Unger, B. D., Urlichich, D., Vadillo, M. A., Wackershauser-Sablotny, V., Wessel, I., Wolf, H., Zhou, A., & Schütz, A. (2022). OpAQ: Open Anchoring Quest, Version 1.1.43.96. <https://dx.doi.org/10.17605/OSF.IO/YGNVB>

- Sanitioso, R., Kunda, Z., & Fong, G. T. (1990). Motivated recruitment of autobiographical memories. *Journal of Personality and Social Psychology*, *59*(2), 229-241.
<https://doi.org/10.1037/0022-3514.59.2.229>
- Sedikides, C., & Green, J. D. (2009). Memory as a self-protective mechanism. *Social and Personality Psychology Compass*, *3*(6), 1055–1068.
<https://doi.org/10.1111/j.1751-9004.2009.00220.x>
- Sedikides, C., Green, J. D., Saunders, J., Skowronski, J. J., & Zengel, B. (2016). Mnemic neglect: Selective amnesia of one's faults. *European Review of Social Psychology*, *27*, 1–62. <https://doi.org/10.1080/10463283.2016.1183913>
- Sedikides, C., & Skowronski, J. J. (2020). In human memory, good can be stronger than bad. *Current Directions in Psychological Science*, *29*(1), 86-91.
<https://doi.org/10.1177/0963721419896363>
- Sharot, T., & Garrett, N. (2016). Forming beliefs: Why valence matters. *Trends in Cognitive Sciences*, *20*(1), 25–33. <https://doi.org/10.1016/j.tics.2015.11.002>
- Shah, P., Harris, A. J. L., Bird, G., Catmur, C., & Hahn, U. (2016). A pessimistic view of optimistic belief updating. *Cognitive Psychology*, *90*, 71–127.
<https://doi.org/10.1016/j.cogpsych.2016.05.004>
- Singmann, H., Bolker, B., Westfall, J., and Aust, F. (2018). afex: Analysis of Factorial Experiments. R Package Version 0.21-2. Available at: <https://CRAN.R-project.org/package=afex>.
- Smith, A. R., & Windschitl, P. D. (2015). Resisting anchoring effects: The roles of metric and mapping knowledge. *Memory & Cognition*, *43*(7), 1071-1084.
<https://doi.org/10.3758/s13421-015-0524-4>
- Smith, A. R., Windschitl, P. D., & Bruchmann, K. (2013). Knowledge matters: Anchoring effects are moderated by knowledge level. *European Journal of Social Psychology*, *43*(1), 97-108. <https://doi.org/10.1002/ejsp.1921>

- Soto, C. J. (2021). Do links between personality and life outcomes generalize? Testing the robustness of trait–outcome associations across gender, age, ethnicity, and analytic approaches. *Social Psychological and Personality Science*, *12*(1), 118-130. <https://doi.org/10.1177/1948550619900572>
- Stanley, M. L., Henne, P., & De Brigard, F. (2019). Remembering moral and immoral actions in constructing the self. *Memory & Cognition*, *47*(3), 441-454. <https://doi.org/10.3758/s13421-018-0880-y>
- Stavrova, O., Köneke, V., & Schlösser, T. (2016). Overfulfilling the Norm: The better-than-average effect in judgements of attitudes. *Social Psychology*, *47*, 288-293. <https://doi.org/10.1027/1864-9335/a000280>
- Strack, F., Bahník, Š., & Mussweiler, T. (2016). Anchoring: accessibility as a cause of judgmental assimilation. *Current Opinion in Psychology*, *12*, 67-70. <https://doi.org/10.1016/j.copsyc.2016.06.005>
- Strack, F., & Mussweiler, T. (1997). Explaining the enigmatic anchoring effect: Mechanisms of selective accessibility. *Journal of Personality and Social Psychology*, *73*, 437–446. <https://doi.org/10.1037/0022-3514.73.3.437>
- Stieger, M., Wepfer, S., Rügger, D., Kowatsch, T., Roberts, B. W., & Allemand, M. (2020). Becoming more conscientious or more open to experience? Effects of a two-week smartphone-based intervention for personality change. *European Journal of Personality*, *34*(3), 345–366. <https://doi.org/10.1002/per.2267>
- Strohinger, N., & Nichols, S. (2014). The essential moral self. *Cognition*, *131*(1), 159–171. <https://doi.org/10.1016/j.cognition.2013.12.005>
- Symons, C. S., & Johnson, B. T. (1997). The self-reference effect in memory: A meta-analysis. *Psychological Bulletin*, *121*(3), 371–394. <https://doi.org/10.1037/0033-2909.121.3.371>

- Tappin, B. M., & McKay, R. T. (2017). The illusion of moral superiority. *Social Psychological and Personality Science*, 8(6), 623-631.
<https://doi.org/10.1177/1948550616673878>
- Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157), 1124-1131.
<https://doi.org/10.1126/science.185.4157.1124>
- Van Rossum, G. & Drake, F. L. (2009). *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace.
- Waskom, M. et al. (2017). *mwaskom/seaborn: v0.8.1 (September 2017)*, Zenodo.
Available at: <https://doi.org/10.5281/zenodo.883859>.
- Wong, K. F. E., & Kwong, J. Y. Y. (2000). Is 7300 m equal to 7.3 km? Same semantics but different anchoring effects. *Organizational Behavior and Human Decision Processes*, 82(2), 314-333. <https://doi.org/10.1006/obhd.2000.2900>
- Yoon, S., Fong, N. M., & Dimoka, A. (2019). The robustness of anchoring effects on preferential judgements. *Judgment & Decision Making*, 14(4), 470-487.
<https://EconPapers.repec.org/RePEc:jdm:journl:v:14:y:2019:i:4:p:470-487>
- Zhang, Y., Pan, Z., Li, K., & Guo, Y. (2018). Self-serving bias in memories. *Experimental Psychology*, 65(4), 236–244. <https://doi.org/10.1027/1618-3169/a000409>
- Zell, E., Strickhouser, J. E., Sedikides, C., & Alicke, M. D. (2020). The better-than-average effect in comparative self-evaluation: A comprehensive review and meta-analysis. *Psychological Bulletin*, 146(2), 118–149. <https://doi.org/10.1037/bul0000218>

Appendix A. Pre-registered ANOVA analysis

We ran a 3 (high, low, no anchor) x 2 (desirable, undesirable) x 2 (moral, non-moral) mixed ANOVA to analyse the full sample and followed up with post hoc tests for binary comparisons. The ANOVA analysis revealed highly significant main effects of *anchor*, $F(2,245) = 14.75$, $p < .001$, $\eta_{ges}^2 = .041$ and *desirability* $F(1, 245) = 483.63$, $p < .001$, $\eta_{ges}^2 = .416$. The interaction between *desirability* and *morality* was also highly significant $F(1, 245) = 66.87$, $p < .001$, $\eta_{ges}^2 = .044$ while the interaction between *anchor* and *desirability* was significant at the 10% level only, $F(2, 245) = 2.70$, $p = .069$, $\eta_{ges}^2 = .008$. No other effects were significant (Table A1).

Table A1. ANOVA analysis for the self-rankings measure

	F-value	df	MSE	ges	p-value
Anchor	14.75***	245	543.56	0.041	< .001
Desirability	483.63***	245	547.59	0.416	< .001
Morality	0.08	245	169.42	< .001	.773
Interaction Anchor * Desirability	2.70	245	547.59	0.008	0.069
Interaction Anchor * Morality	0.96	245	169.42	< .001	.385
Interaction Desirability * Morality	66.87***	245	256.64	0.044	< .001
Interaction Anchor * Desirability * Morality				0.002	
Morality	1.78	245	256.64		.170

N = 248; Significance codes: '***' 0.001; '**' 0.01; '*' 0.05;

The main effect of *anchor* pertained to our primary hypothesis of anchoring of self-judgements (H1), and we explored it further by running post-hoc binary comparison tests. The follow-up analysis showed that self-rankings were significantly different among all levels of the *anchor* factor. Consistent with H1, participants provided significantly higher self-rankings in the high anchor condition ($M= 50.45$, $SD = 31.09$) than in both the no anchor condition ($M= 45.83$, $SD = 25.38$), $p = .010$, $d_{avg}^2 = 0.16$ and the low anchor condition ($M= 39.82$, $SD = 29.71$), $p < .001$, $d_{avg}^2 = 0.35$.

Self-rankings were also significantly higher in the no anchor than in the low anchor condition, $p = .002$, $d_{avg}^2 = -0.22$. These results confirm an anchoring effect for self-judgements.

To investigate whether anchoring was more pronounced when anchors were *self-serving* (H2), we examined the interaction between *anchor* and *desirability* (significant at 10%). Comparing the anchoring conditions at each level of the *desirability* factor suggested that anchoring was most effective when people were anchored in a self-serving direction: for desirable traits self-rankings in the high anchor condition ($M=68.47$, $SD = 22.99$) were significantly higher than self-rankings in both the no anchor condition ($M=60.17$, $SD = 19.22$), $p = .004$, $d_{avg}^2 = .39$ and in the low anchor condition ($M=57.42$, $SD = 24.24$), $p < .001$, $d_{avg}^2 = .47$. However, the difference between the no anchor and the low anchor condition was not significant, $p = .481$. In contrast, for undesirable traits, the low anchor influenced self-rankings while the high anchor did not have a significant impact: self-rankings in the low anchor condition ($M=22.21$, $SD = 23.65$) were significantly lower than self-rankings in both the no anchor condition ($M=31.48$, $SD = 22.54$), $p = .001$, $d_{avg}^2 = -.40$ and the high anchor condition ($M=32.43$, $SD = 27.50$), $p = .001$, $d_{avg}^2 = -.40$, but there was no significant difference between the no anchor and the high anchor conditions, $p = .710$, see Figure A1.

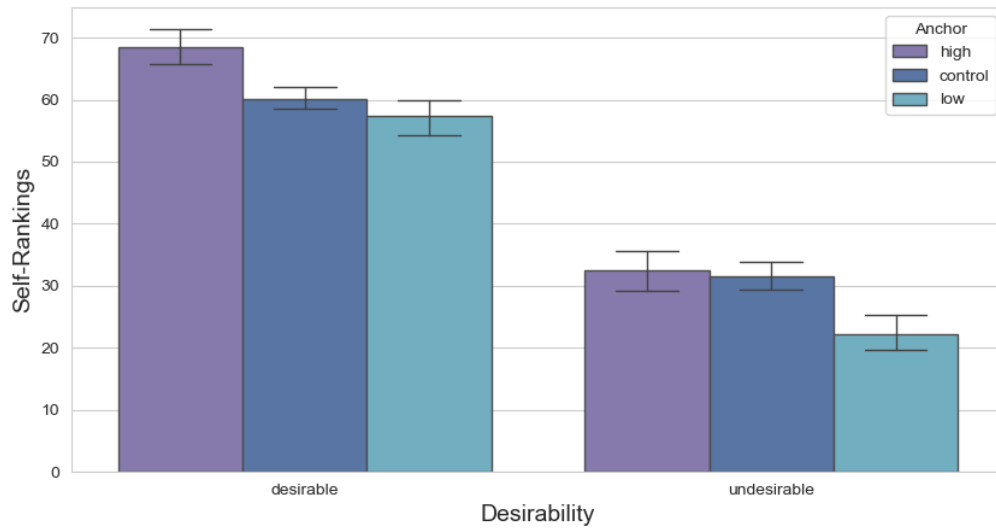


Figure A1. Self-rankings by *anchor* condition and *desirability*. Error bars represent 95% confidence intervals (CI).

Chapter 5. Study 3. Anchoring of Personality Judgements and Its impact on
Subsequent Personality Judgements

Study 3. Anchoring of Personality Judgements and Its impact on Subsequent
Personality Judgements

Elitza Ambrus

Bjoern Hartig

Ryan McKay

Manuscript in preparation

Word count: 4020 excluding references and appendices

Abstract

Personality traits are important predictors of life outcomes, with recent wellbeing research focussed on shaping them via nonclinical psychological interventions. Our previous work shows that judgements of one's personal qualities are prone to anchoring when the anchor value elevates the self-image. Here, we explore whether the self-serving anchoring effect on personality judgements can influence subsequent self-rankings. Participants (N = 228) first responded whether they would rank themselves higher than 95 (high anchor) or 5 (low anchor) out of 100 other participants on two morally relevant personality traits. Participants then indicated their self-rankings in comparison to others on eight different personality traits. The data replicated the self-serving anchoring effect on personality judgements, however there was no support for a self-serving *aftereffect* on subsequent personality judgements. In addition, an exploratory analysis showed an interesting phrasing effect: *enhancing* anchors had a stronger impact for traits phrased in a negative rather than positive way. Limitations of the study are discussed as well as implications for future efforts to design interventions influencing personality characteristics.

Keywords: personality judgements; self-serving anchoring; anchoring aftereffect

Anchoring of Personality Judgements and Its impact on Subsequent Personality Judgements

Personality traits predict a wide range of life outcomes, such as work performance, health, and well-being (e.g., Beck & Jackson, 2022) and are commonly considered stable in adulthood (Costa & McCrae, 1992; McCrae et al., 2000). In the last two decades, however, research has shown that personality traits continue to change in adulthood, albeit slowly (Roberts et al., 2006; Bleidorn et al., 2022; Soto, 2021). This has occasioned a surge of research on the effectiveness of nonclinical psychological interventions to change personality traits (Allemand & Flückiger, 2022; Bleidorn et al., 2021; Olaru et al., 2022; Stieger et al., 2020; Roberts et al., 2017). Our previous work (Ambrus et al., under review) provides evidence for a different type of flexibility of personality judgements: self-rankings are influenced by anchoring (Tversky & Kahneman, 1974), when the influence (the anchor value) elevates the self-image. The current study builds on the instantaneous flexibility of personality judgements and explores whether the self-serving anchoring effect on personality judgements carries over to subsequent self-rankings.

The anchoring effect is a robust phenomenon affecting judgements and choices across a wide range of decision-making domains (Furnham & Boo, 2010; Strack et al., 2016; Röseler & Schütz, 2022; Yoon et al., 2019). In the classic anchoring framework, participants are first asked to compare their estimate of a particular quantity with a salient value (the anchor), and then provide their specific estimate (Tversky & Kahneman, 1974). For example, real estate agents were provided with information, such as the listed price of a house (which they knew was the seller's best guess rather than a professional appraisal of the property value), yet their subsequent estimates of the property value were anchored by the irrelevant data provided (Northcraft & Neale, 1987). The anchoring effect endures despite monetary incentives for accuracy, occurs even when extreme or implausible anchor values are used and is observed whether the study is conducted online or as a lab experiment (Strack & Mussweiler, 1997; Röseler & Schütz, 2022). The effect also persists – though in mitigated form – when anchors are irrelevant (Glöckner & Englich, 2015; Sugden et al., 2013) or when anchor-inconsistent

information is considered prior to the subject providing their estimate (Mussweiler et al., 2000). Even explicit instructions to correct for the potential influence of the anchoring effect, do not outweigh the effect (Wilson et al., 1996).

Meanwhile, the literature on personality development has demonstrated that personality traits are stable in young and middle adulthood (Lucas & Donnellan, 2011; Soto et al., 2021) and can serve as a valid predictor of life outcomes (Beck & Jackson, 2022; Roberts et al., 2007; Soto, 2021). Although stability underlies the very definition of personality characteristics as consistent behavioural patterns over contexts (Roberts, 2009), recent research shows that while some personality traits exhibit stability, other personality traits remain flexible in middle adulthood too (Bleidorn et al., 2022). The observed malleability of personal characteristics is important as it provides theoretical premises for successful nonclinical psychological interventions (Allemand & Flückiger, 2022). Yet, none of the existing personality theories can accommodate quick adjustments of personality judgements in response to cognitive influences.

Ambrus et al., (in prep) found that self-judgements of personality characteristics are swiftly adjusted in response to anchoring, in a self-serving manner: *enhancing* anchors have a strong effect while *diminishing* anchors have little or no effect. These findings are in line with research showing that self-relevant information can be anchored (Cheek et al. 2015; Greenberg et al., 2017; Joel et al., 2017). For instance, Joel et al. (2017) found that individuals' probability estimates of future life outcomes are susceptible to anchoring, however anchors that pointed in undesirable direction, e.g., high probability of undesirable future events, were not effective. Therefore, considerations of maintaining a positive self-image and favourable future prospects might interact with the anchoring effect when self-relevant information is concerned.

Indeed, constructing and maintaining a stable and positive self-image is of paramount importance to us (Alicke et al., 2013; Leary, 2007). Ziano et al., (2021) showed that individuals consider desirable traits as more descriptive of themselves than of others; most participants also believe they perform better than the average person (Alicke & Govorun, 2005; Brown, 1986). In addition, morality has a central role in our

self-concept with moral traits being perceived as the essence of the human soul (Heiphetz et al., 2016; Strohminger, 2018; Strohminger & Nichols, 2014). Cognitive biases, such as the better-than-average effect and the sunk-cost effect are more pronounced for moral traits (Brown, 2012; Meyers et al., 2019; Tappin & McKay, 2017). Due to the defining role of moral traits for the self-concept (Strohminger & Nichols, 2014), manipulating judgements of morally relevant personality traits might be effective in achieving a change that impacts the overall self-image and thus subsequent personality judgements too.

In the present study, our main hypotheses are that anchoring of morally relevant personality traits will be effective, particularly if the anchor elevates the self-image. In addition, anchoring of morally relevant personality traits will have an *aftereffect* on subsequent personality judgements, especially if the anchor elevates the self-image.

Method

Overview

The study was designed in Qualtrics and conducted online via the platform *Prolific* (www.prolific.com). We piloted the design of this study and Study 4 (Chapter 6) also on *Prolific* (please see Appendix 5A). The study had two stages: in the first stage of the experiment, participants reported their self-rankings on two morally relevant personality traits (both desirable or both undesirable). In the anchoring conditions, respondents first indicated whether their self-ranking on each respective trait was higher or lower than the anchor value, and then provided their specific self-ranking. In the second stage of the experiment, all respondents provided their self-rankings on a set of eight different personality traits (a combination of desirable and undesirable, moral and non-moral traits). All participants also responded to an attention check question as well as provided data on their age and gender. Our hypotheses, data collection, and analysis protocol were pre-registered (https://aspredicted.org/PXR_W4K). De-identified data and analysis scripts are available on the Open Science Framework:

https://osf.io/r7fvw/?view_only=875e761272844a12a1d2df3d8efc55b4

Participants

We recruited 300 participants in a pre-registered online experiment. All participants were monolingual English speakers with UK nationality and had a *Prolific* approval rate higher than 90%. As specified in our pre-registration document, we excluded participants who: (i) failed the attention check question, asking participants to place the slider bar at 50, the midpoint of the scale (N=37); (ii) provided inconsistent answers in the anchoring paradigm, e.g., stating they would rank themselves higher than 95 at the first stage of the study, however providing a self-ranking lower than 95 at the second stage, suggesting they misunderstood the task or did not pay attention (N=35), or (iii) completed the study in less than 30 seconds (N=0). The final sample comprised 228 participants (164 females, 62 males, 2 other; $M = 32.83$, $SD = 10.88$). The distribution of participants across conditions was as follows: 33 in the high anchor, desirable trait condition; 42 in the high anchor, undesirable trait condition; 36 in the low anchor, desirable trait condition; 31 in the low anchor, undesirable trait condition; 45 in the no anchor, desirable trait condition and 41 participants in the no anchor, undesirable trait condition. Respondents were paid a flat participation fee (70p, the equivalent of £7.50 per hour). The study was self-certified in accordance with the Royal Holloway, University of London Ethics Committee procedure.

Materials and Procedure

Participants were randomly assigned to one of the six experimental conditions, resulting from the cross-section of the two between-subject variables anchor and desirability. In the first stage of the experiment, participants ranked themselves either on two desirable traits (*honest*, *considerate*) or on two undesirable traits (*dishonest*, *inconsiderate*). The self-rankings on these two morally relevant traits followed the classic anchoring paradigm (Tversky & Kahneman, 1974): participants first indicated whether they would rank themselves higher than at least 95 other participants (out of 100 other anonymous participants) in the high anchor condition or higher than at least 5

other participants in the low anchor group (the slider cursor was pre-set at 95 and 5 respectively), and then provided their specific self-rankings on the respective traits.

Self-rankings were indicated on a slider bar, ranging from 0 to 100 (no numeric values were displayed on the slider bar, however the corresponding numeric value of the slider cursor was shown just above it and changed dynamically when the slider cursor changed position). The comparative questions for both traits were shown on the same screen, followed by a new screen asking for the specific self-rankings on both traits (Appendix 5B). In the no anchor condition, participants indicated their personality judgements without answering a comparative question first (the slider cursor was invisible until participants clicked on the slider line to indicate their self-rankings).

In the second stage of the experiment, participants provided their self-rankings in comparison to the 100 other participants on a set of eight personality traits. The traits were balanced along the levels of desirability and morality so that we had a pair of traits for each of the four combinations of desirable/undesirable and moral/nonmoral (*trustworthy, principled, manipulative, prejudiced, creative, easy going, illogical, uptight*). The same eight traits were presented in a separate random order for each participant (Appendix 5C). The eight traits were selected to ensure comparable average desirability levels between the moral and non-moral personality characteristics (Tappin & McKay, 2017).

Design and Analysis

The first stage of the experiment had a 3 (anchor: high, low, no anchor) x 2 (desirability: desirable, undesirable) between-subject design. The second stage of the experiment employed a 3 (anchor: high, low, no anchor, between-subject) x 2 (desirability: desirable, undesirable, between-subject) x 2 (trait: moral, non-moral, within-subject) mixed design. The DVs were self-rankings on the first and second stage of the experiment (measured on a 0 to 100 scale). The self-rankings on undesirable traits were reverse coded (i.e., we subtracted the respective value from 100). The analyses were conducted in R version 4.1.0 (R Development Core Team, 2021). We

fitted linear mixed models for the self-rankings of personality traits measure, using the lme4 package (Bates et al., 2015).

Results

Anchoring at the first stage of the experiment

To analyse the data, we first constructed two dummy variables: *enhancing* anchor and *diminishing* anchor. *Enhancing* anchor suggests an improved self-view; a dummy that takes a value of 1 for a high anchor on a desirable trait or a low anchor on an undesirable trait, and 0 otherwise; *Diminishing* anchor reflects deterioration of the self-view; a dummy that takes a value of 1 for a high anchor on an undesirable trait or a low anchor on a desirable trait, and 0 otherwise. There were 64 participants in the *enhancing* anchor condition, 78 in the *diminishing* anchor condition and 86 participants in the no anchor condition (*control*). We fitted a linear mixed model to analyse the self-rankings measure with *enhancing* anchor and *diminishing* anchor as fixed effects and participants as a random effect.

The results provide support for our first main hypothesis: as expected, the coefficient for the *enhancing* anchor was positive and significant ($p = .013$), showing that participants who were exposed to *enhancing* anchors provided on average higher self-rankings than participants in the no anchor condition (Table 1). We also expected that *diminishing* anchor will lead to significantly lower self-rankings than *enhancing* anchors, possibly also lower than self-judgements of participants who were not anchored. The coefficient for the *diminishing* anchor was negative and bordering on significance at 5% ($p = .050$), which supports that *diminishing* anchors might lead to lower self-rankings than the ones indicated in the no anchor condition. To compare the effect of *enhancing* and *diminishing* anchors we used the Wald test for equality of coefficients, which showed that the two coefficients are significantly different $W(1) = 18.29, p < .001$. Therefore, the data provided support for our first and second main hypotheses and replicated the self-serving anchoring effect (Study 2, Chapter 4).

Table 1. Estimated fixed effects for personality judgements at the first stage of the study

	Estimate	SE	t-value	p-value
Intercept	74.13	2.06	36.06	< .001
<i>Enhancing</i> anchor	7.88	3.15	2.50	.013
<i>Diminishing</i> anchor	-5.87	2.98	-1.97	.050

Baseline levels: no anchor; Number of observations: 456; grouped by participants, N=228

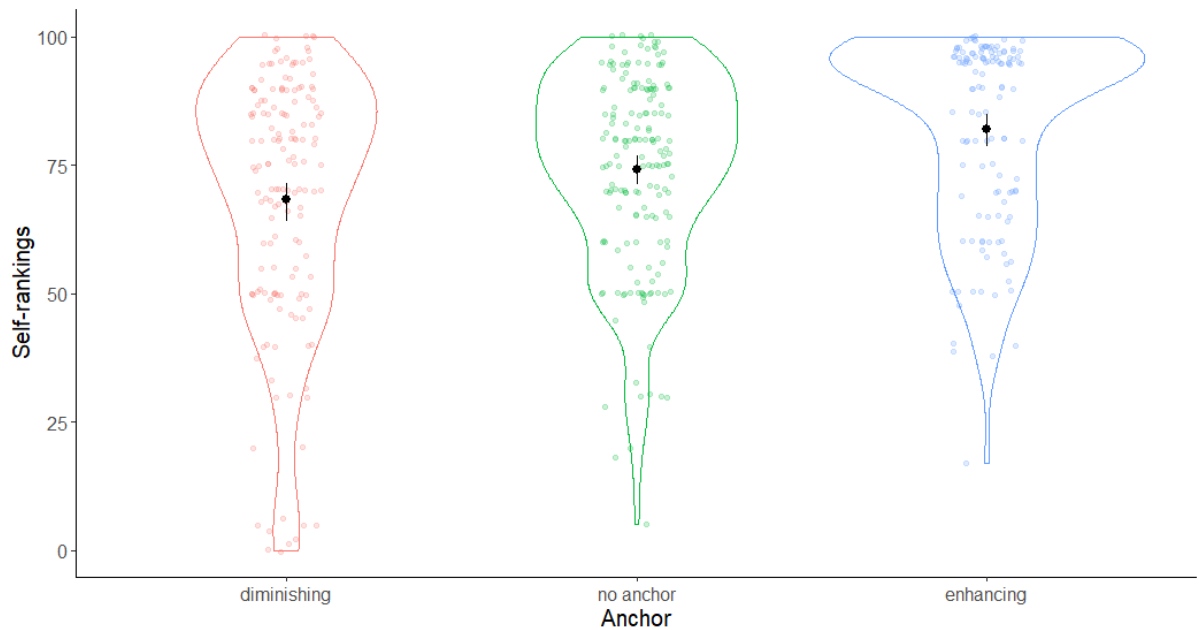


Figure 1. Violin plots of the self-rankings data by anchor conditions.

Anchoring *aftereffect* at the second stage of the experiment

To test whether anchoring at the first stage of the study will influence self-rankings at the second stage too, we fitted a linear mixed models explaining the second stage self-rankings with the anchor types. In Model 1, *enhancing* anchor and *diminishing anchor* were modelled as fixed effects while participants were modelled as a random effect. Model 2 relies on the same specification as Model 1, however morality was added as a fixed effect (Table 2).

The analysis did not support our expectations for an *aftereffect* on subsequent personality judgements as neither the coefficient for *enhancing* ($p = .972$) nor *diminishing* anchor ($p = .912$) was significant, Model 1 and 2, Table 2.

Table 2. Estimated fixed effects of *enhancing*, *diminishing*, morality and their interaction on personality judgements at the second stage of the experiment

	Model 1	Model 2
Intercept	64.37*** (1.13)	57.61*** (1.27)
<i>Enhancing</i> anchor aftereffect	-0.06 (1.73)	-0.06 (1.73)
<i>Diminishing</i> anchor aftereffect	-0.18 (1.64)	-0.18 (1.64)
Moral traits		13.51*** (1.15)

Baselines: no anchor, non-moral traits, Number of observations 1824, grouped by participants, N=228

The highly significant coefficient for moral traits (13.51, Table 2) was somewhat unexpected since we selected moral and non-moral traits with comparable average desirability levels (Tappin & McKay, 2017). We explored the average self-rankings for each personality traits (Table 3) and it seems that individuals had the tendency to rank the specific moral traits we chose higher than the non-moral traits. This resulted in the large positive value of the coefficient for morality in Model 2 (Table 2).

Table 3. Means and Standard Deviations for self-rankings at the second stage of the study.

	Mean (SD)	Moral	Average
Trustworthy	70.62 (20.25)	Yes	
Principled	56.49 (22.43)	Yes	71.04
Manipulative*	77.68 (21.58)	Yes	(22.80)
Prejudiced*	79.38 (19.50)	Yes	
Easy-going	53.04 (26.10)	No	57.53
Creative	41.94 (26.03)	No	(27.55)
Illogical*	73.85 (22.48)	No	
Uptight*	61.30 (25.20)	No	

*Scores for undesirable traits are reverse-coded (subtracted from 100)

Since there was no evidence for an *aftereffect* of anchoring on subsequent self-rankings we did not conduct any further investigation of our secondary hypotheses because these were concerned with moderation of the (absent) *aftereffect*. However, the data showed that when reverse-coded, self-rankings on undesirable traits (e.g., “dishonest”) were higher on average than the personality judgements on desirable traits (e.g., “honest”). This finding led us to perform an exploratory analysis of a potential effect on self-rankings of the way the personality traits were phrased at the first stage of the experiment (i.e., “honest” and “considerate” vs. “dishonest” and “inconsiderate”).

Exploratory analysis. The effect of the way the personality traits are phrased (positively or negatively)

At the first stage of the experiment, we elicited self-rankings using pairs of the same personality traits, phrased both either positively or negatively (honest/dishonest, considerate/inconsiderate). Subsequently, we reverse scored the undesirable traits (subtracting the provided self-rankings from 100) for the analyses. Rationally, one would expect that for a given anchoring condition, the average personality judgements for

honesty should be the same as the reversed average personality judgements for dishonesty. To test this assumption, we fitted a linear mixed model on the self-rankings measure provided in the first stage of the experiment, including *enhancing* anchor, *diminishing* anchor, desirability as well as their interaction as fixed factors and participants as a random effect (please see Table 4 and Figure 2).

The effect of desirability was negative and highly significant ($p = .002$), showing that reverse-scored average self-rankings on negatively phrased traits were higher than average self-rankings on positively phrased traits. We followed up on the significant effect of desirability with post-hoc binary comparisons. The data showed that in all three anchoring conditions (enhancing, diminishing, control), participants ranked themselves significantly higher (i.e., as “better”) when the traits were phrased negatively (dishonest, inconsiderate). In all three conditions, going from positive phrasing to negative phrasing increased the average ranking substantially (+17-30%).

Specifically, the post-hoc binary comparisons revealed that in the no anchor condition, respondents provided on average lower personality judgements for desirable (positively phrased) traits ($M = 68.39, SD = 18.92$) than for undesirable (negatively phrased) traits ($M = 80.43, SD = 17.51$), $p = .023$. In a similar vein, personality judgements in the *enhancing* anchor, desirable traits condition ($M = 75.52, SD = 18.04$) were lower than self-rankings in the *enhancing* anchor, undesirable traits condition ($M = 88.92, SD = 16.63$), $p = .032$. Personality judgements in the *diminishing* anchor, desirable traits condition ($M = 58.58, SD = 25.16$) were lower than self-rankings in the *diminishing* anchor, undesirable traits condition ($M = 76.55, SD = 21.90$), $p < .001$.

Table 4. Estimated linear model for the self-rankings indicated at the first stage of the study

	Estimate	SE	t-value	p-value
Intercept	80.43	2.76	29.10	< .001
<i>Enhancing</i> anchor	-3.88	3.89	-0.10	.319
<i>Diminishing</i> anchor	8.49	4.21	2.02	.045
Desirable traits	-12.04	3.82	-3.15	.002
Interaction <i>Enhancing</i>				
anchor*Desirable trait	-1.37	5.85	-0.23	.815
Interaction <i>Diminishing</i>				
anchor*Desirable trait	-5.93	5.55	-1.07	.286

Baselines: no anchor, undesirable traits, N=228

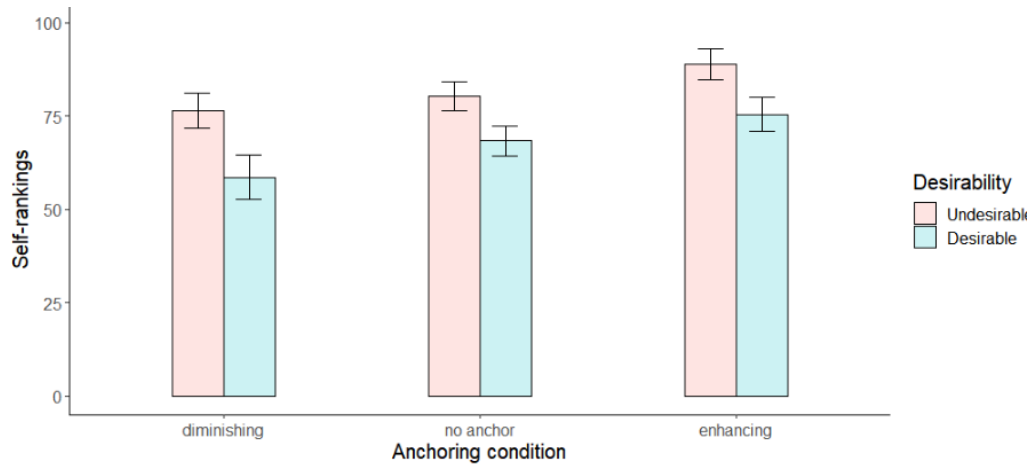


Figure 2. Bar plot of self-rankings by anchoring conditions.

Discussion

The anchoring effect is a robust cognitive phenomenon, affecting judgements and choices in a wide range of decision-making domains (Röseler & Schütz, 2022). Our previous research (Ambrus et al., under review) extended anchoring to the domain of the self and showed that personality judgements are susceptible to anchoring in a self-

serving manner. The current study built on the self-serving anchoring effect on personality judgements by exploring potential *aftereffects* on subsequent personality judgements.

The results showed that anchoring influenced self-rankings in the first stage of the experiment, which is in line with the robustness of the anchoring effect across decision-making domains (e.g., Strack et al., 2016). These findings also replicate the self-serving anchoring effect (Ambrus et al., under review): *enhancing* anchors increased self-rankings, while the effect of *diminishing* anchors was borderline. In the second stage of the experiment, we found no evidence for an *aftereffect* of *enhancing* or *diminishing* anchors on subsequent personality judgements. The lack of *aftereffect* of *diminishing* anchors is consistent with our expectations and is also in line with literature showing individuals' reluctance to internalise negative self-relevant information (Joel et al., 2017; Möbius et al., 2022; Ritchie et al., 2017; Sedikides & Green, 2009). Nevertheless, we expected that *enhancing* anchors will contribute to elevated subsequent self-rankings. The analysis did not support exacerbated *enhancing* anchors *aftereffect* for morally relevant traits, rather morality had a general elevating effect on self-rankings regardless of the type of anchor employed.

In retrospect, the current study has several limitations: in Ambrus et al. (under review) we anchored self-rankings on eight personality traits (desirable and undesirable, moral and non-moral) while due to budget considerations the current study relied on two personality traits only. It is possible that individuals need to be exposed to anchoring of a range of personality traits (encompassing both desirable and undesirable, moral and non-moral traits) so that the effect on subsequent self-rankings is more pronounced. For instance, Aquino and Reed (2002) outlined nine central moral personality characteristics as evoking individuals' moral identity.

In addition, although we selected moral and non-moral traits with comparable average desirability ratings (Tappin & McKay, 2017), the analysis showed that participants indicated substantially higher self-rankings on moral traits, which precluded a valid investigation of a potential differential self-serving effect on moral traits. Yet, the

relatively high self-rankings on moral traits are in line with literature on the importance of morality to the self-image (e.g., Strohminger, 2018).

Furthermore, due to the data exclusions and the employed between-subject design, our final sample comprised a relatively small number of participants per condition (ranging from 64 to 86). Despite all the precautionary measures we took (three pre-screening criteria), a large portion of the recruited participants (24%) failed the attention checks, which means that there might still be too much noise in the data. We excluded inattentive participants in the anchoring conditions based on both attention and comprehension checks while in the control condition we relied on an attention check only. Recent research has provided evidence that excluding data from inattentive participants might result in substantially different findings (Sulik et al., 2023). Failure to filter out all the inattentive participants from the control condition in the current study might have rendered the potential *aftereffect* of self-serving anchors on subsequent personality judgements difficult to capture. Having in mind the above discussed limitations, the lack of self-serving *aftereffect* we find does not necessarily mean that there is no potential transfer of self-serving anchoring to subsequent self-rankings. Future research, relying on a carefully selected set of personality traits, triggering the moral self-concept might further explore this question.

In an exploratory analysis, we investigated another factor that might potentially facilitate the anchoring effect on self-rankings, namely the way the personality traits are phrased at the anchoring stage of the experiment (positively or negatively). The average personality judgements of positively and negatively phrased traits should have been similar as these are personality judgements on essentially the same trait (self-rankings on undesirable traits were reverse scored). However, the analysis showed that the average personality judgements on the negatively phrased traits (when reverse scored) were significantly higher than the average personality judgements on the same traits, phrased in a positive manner.

The fact that the negatively phrased traits trigger even more elevated personality judgements is reminiscent of the loss aversion effect with losses looming

larger than gains (Tversky & Kahneman, 1981). For instance, self-rankings of how “dishonest” one is might be perceived as a potential “loss” in one’s self-image. In contrast, self-rankings of how “honest” one is might be perceived as potential “gains” in the self-image. This would explain the observed relatively higher personality judgements on negatively phrased traits (when reverse scored). The “loss aversion” effect on personality judgements underscores the importance of potential self-image considerations for the adjustments of personality judgements. This is also in line with literature outlining self-protecting behavioural motives as more important than self-enhancing motives (Alicke & Sedikides, 2009). The observed interplay of cognitive biases involved in self-assessment also illuminates the way personality judgements are constructed and adjusted, which could be utilised to facilitate designing successful interventions to shape personality traits and promote wellbeing.

In conclusion, this study replicated the self-serving anchoring effect on personality judgements. We found no evidence for an *enhancing aftereffect* on subsequent personality judgements. Having in mind the limitations of our study, such as a relatively small sample and anchoring on two personality traits, our results show potential for further research. Our findings support a different type of malleability of personality judgements than that previously theorised – self-rankings can instantly adjust in response to a cognitive influence and enhancing adjustments might transfer to the general self-image too. A potential self-serving anchoring *aftereffect* on the overall self-image would have implications for the recent surge in research effort in designing interventions shaping personality traits.

References

- Allemand, M., & Flückiger, C. (2022). Personality change through digital-coaching interventions. *Current Directions in Psychological Science*, 31(1), 41-48. <https://doi.org/10.1177/0963721421106778>
- Alicke, M. D., & Govorun, O. (2005). The better-than-average effect. *The Self in Social Judgment*, 1, 85-106. https://doi.org/10.1007/978-3-319-24612-3_300293
- Alicke, M. D., & Sedikides, C. (2009). Self-enhancement and self-protection: What they are and what they do. *European review of social psychology*, 20(1), 1-48. <https://doi.org/10.1080/10463280802613866>
- Alicke, M. D., Zell, E., & Guenther, C. L. (2013). Social self-analysis: Constructing, protecting, and enhancing the self. *Advances in Experimental Social Psychology*, 48, 173–234. <https://doi.org/10.1016/B978-0-12-407188-9.00004-1>
- Ambrus, E. Z., Hartig, B., McKay, R. T. (2023). Self-serving anchoring of personality judgements. *Under review*
- Aquino, K., & Reed, A. II. (2002). The self-importance of moral identity. *Journal of Personality and Social Psychology*, 83(6), 1423–1440. <https://doi.org/10.1037/0022-3514.83.6.1423>
- Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. *arXiv preprint arXiv:1506.04967*.
- Beck, E. D., & Jackson, J. J. (2022). A mega-analysis of personality prediction: Robustness and boundary conditions. *Journal of Personality and Social Psychology*, 122(3), 523– 553. <https://doi.org/10.1037/pspp0000386>
- Bleidorn, W., Hopwood, C. J., Back, M. D., Denissen, J. J., Hennecke, M., Hill, P. L., ... & Zimmermann, J. (2021). Personality trait stability and change. *Personality Science*, 2, 1-20. <https://doi.org/10.5964/ps.6009>

- Bleidorn, W., Schwaba, T., Zheng, A., Hopwood, C. J., Sosa, S. S., Roberts, B. W., & Briley, D. A. (2022). Personality stability and change: A meta-analysis of longitudinal studies. *Psychological bulletin*, *148*(7-8), 588
<https://doi.org/10.1037/bul0000365>
- Brown, J. D. (1986). Evaluations of self and others: Self-enhancement biases in social judgments. *Social cognition*, *4*(4), 353-376.
<https://doi.org/10.1521/SOCO.1986.4.4.353>
- Brown, J. D. (2012). Understanding the better than average effect: Motives (still) matter. *Personality and Social Psychology Bulletin*, *38*(2), 209-219.
<https://doi.org/10.1177/0146167211432763>
- Costa Jr, P. T., & McCrae, R. R. (1992). Four ways five factors are basic. *Personality and Individual Differences*, *13*(6), 653-665. [https://doi.org/10.1016/0191-8869\(92\)90236-l](https://doi.org/10.1016/0191-8869(92)90236-l)
- Furnham, A., & Boo, H. C. (2011). A literature review of the anchoring effect. *The Journal of Socio-Economics*, *40*(1), 35-42. <https://doi.org/10.1016/j.socec.2010.10.008>
- Glöckner, A., & Englich, B. (2015). When relevance matters. *Social Psychology*, *46*(1), 4-12. <https://doi.org/10.1027/1864-9335/a000214>
- Heiphetz, L., Strohminger, N., & Young, L. L. (2017). The role of moral beliefs, memories, and preferences in representations of identity. *Cognitive science*, *41*(3), 744-767.
<https://doi.org/10.1111/cogs.12354>
- Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*, 263-291. [https://doi.org/0012-9682\(197903\)47:2<263:PTAAOD>2.0.CO;2-3](https://doi.org/0012-9682(197903)47:2<263:PTAAOD>2.0.CO;2-3)
- Leary, M. R. (2007). Motivational and emotional aspects of the self. *Annual Review of Psychology*, *58*, 317-344. <https://doi.org/10.1146/annurev.psych.58.110405.085658>

- Lucas, R. E., & Donnellan, M. B. (2011). Personality development across the life span: longitudinal analyses with a national sample from Germany. *Journal of Personality and Social Psychology*, *01*(4), 847-861.
<https://doi.org/10.1037/a0024298>
- McCrae, R. R., Costa, P. T. J., Jr., Ostendorf, F., Angleitner, A., Hrebícková, M., Avia, M. D., Sanz, J., Sánchez-Bernardos, M. L., Kusdil, M. E., Woodfield, R., Saunders, P. R., & Smith, P. B. (2000). Nature over nurture: Temperament, personality, and life span development. *Journal of Personality and Social Psychology*, *78*(1), 173–186. <https://doi.org/10.1037/0022-3514.78.1.173>
- Meyers, E. A., Białek, M., Fugelsang, J. A., Koehler, D. J., & Friedman, O. (2019). Wronging past rights: The sunk cost bias distorts moral judgment. *Judgment and Decision Making*, *14*(6), 721-727.
<https://EconPapers.repec.org/RePEc:jdm:journl:v:15:y:2020:i:6:p:909-925>
- Molouki, S., & Bartels, D. M. (2017). Personal change and the continuity of the self. *Cognitive Psychology*, *93*, 1-17. <https://doi.org/10.1016/j.cogpsych.2016.11.006>
- Möbius, M. M., Niederle, M., Niehaus, P., & Rosenblat, T. S. (2022). Managing self-confidence: Theory and experimental evidence. *Management Science* *68*(11).
<https://doi.org/10.1287/mnsc.2021.4294>
- Mussweiler, T. (2001). The durability of anchoring effects. *European Journal of Social Psychology*, *31*, 431–442. <https://doi.org/10.1002/ejsp.52>
- Mussweiler, T., Strack, F., & Pfeiffer, T. (2000). Overcoming the inevitable anchoring effect: Considering the opposite compensates for selective accessibility. *Personality and Social Psychology Bulletin*, *26*(9), 1142-1150.
<https://doi.org/10.1177/01461672002611010>
- Northcraft, G. B., & Neale, M. A. (1987). Experts, amateurs, and real estate: An anchoring-and-adjustment perspective on property pricing decisions.

Organizational Behavior and Human Decision Processes, 39, 84–97.

[https://doi.org/10.1016/0749-5978\(87\)90046-X](https://doi.org/10.1016/0749-5978(87)90046-X)

Olaru, G., Stieger, M., Rügger, D., Kowatsch, T., Flückiger, C., Roberts, B. W., & Allemand, M. (2022). Personality change through a digital-coaching intervention: Using measurement invariance testing to distinguish between trait domain, facet, and nuance change. *European Journal of Personality*.

<https://doi.org/10.1177/089020702211450>

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>

Ritchie, T. D., Sedikides, C., & Skowronski, J. J. (2017). Does a person selectively recall the good or the bad from their personal past? It depends on the recall target and the person's favourability of self-views. *Memory*, 25(8), 934-944.

<https://doi.org/10.1080/09658211.2016.1233984>

Roberts, B. W. (2009). Back to the future: Personality and assessment and personality development. *Journal of research in personality*, 43(2), 137-145.

<https://doi.org/10.1016/j.jrp.2008.12.015>

Roberts, B. W., Kuncel, N. R., Shiner, R., Caspi, A., & Goldberg, L. R. (2007). The power of personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. *Perspectives on Psychological Science*, 2(4), 313-345. <https://www.jstor.org/stable/40212212>

Roberts, B. W., Luo, J., Briley, D. A., Chow, P. I., Su, R., & Hill, P. L. (2017). A systematic review of personality trait change through intervention. *Psychological Bulletin*, 143(2), 117-141. <https://doi.org/10.1037/bul0000088>

Roberts, B. W., Walton, K. E., & Viechtbauer, W. (2006). Patterns of meanlevel change in personality traits across the life course: A meta-analysis of longitudinal studies. *Psychological Bulletin*, 132(1), 1–25. <https://doi.org/10.1037/0033-2909.132>

- Roberts, B. W., & Yoon, H. J. (2022). Personality psychology. *Annual Review of Psychology*, 73(1), 489–516. <https://doi.org/10.1146/annurevpsych-020821-114927>
- Röseler, L., & Schütz, A. (2022, March 9). Hanging the Anchor Off a New Ship: A Meta-Analysis of Anchoring Effects. PsyArXiv. <https://doi.org/10.31234/osf.io/wf2tn>
- Sedikides, C., & Green, J. D. (2009). Memory as a self-protective mechanism. *Social and Personality Psychology Compass*, 3(6), 1055–1068. <https://doi.org/10.1111/j.1751-9004.2009.00220.x>
- Sedikides, C., Green, J. D., Saunders, J., Skowronski, J. J., & Zengel, B. (2016). Mnemic neglect: Selective amnesia of one's faults. *European Review of Social Psychology*, 27, 1–62. <https://doi.org/10.1080/10463283.2016.1183913>
- Stanley, M. L., Henne, P., & De Brigard, F. (2019). Remembering moral and immoral actions in constructing the self. *Memory & Cognition*, 47(3), 441-454. <https://doi.org/10.3758/s13421-018-0880-y>
- Stieger, M., Wepfer, S., Rügger, D., Kowatsch, T., Roberts, B. W., & Allemand, M. (2020). Becoming more conscientious or more open to experience? Effects of a two-week smartphone-based intervention for personality change. *European Journal of Personality*, 34(3), 345–366. <https://doi.org/10.1002/per.2267>
- Strack, F., Bahnik, Š., & Mussweiler, T. (2016). Anchoring: accessibility as a cause of judgmental assimilation. *Current Opinion in Psychology*, 12, 67-70. <https://doi.org/10.1016/j.copsyc.2016.06.005>
- Strack, F., & Mussweiler, T. (1997). Explaining the enigmatic anchoring effect: Mechanisms of selective accessibility. *Journal of Personality and Social Psychology*, 73, 437–446. <https://doi.org/10.1037/0022-3514.73.3.437>
- Strohinger, N. (2018). Identity is essentially moral. *Atlas of moral psychology*, 141-148. <https://doi.org/10.1016/j.cognition.2013.12.005>

- Strohminger, N., & Nichols, S. (2014). The essential moral self. *Cognition*, *131*(1), 159–171. <https://doi.org/10.1016/j.cognition.2013.12.005>
- Soto, C. J. (2021). Do links between personality and life outcomes generalize? Testing the robustness of trait–outcome associations across gender, age, ethnicity, and analytic approaches. *Social Psychological and Personality Science*, *12*(1), 118-130. <https://doi.org/10.1177/1948550619900572>
- Sugden, R., Zheng, J., & Zizzo, D. J. (2013). Not all anchors are created equal. *Journal of Economic Psychology*, *39*, 21-31. <https://doi.org/10.1016/j.joep.2013.06.008>
- Sulik, J., Ross, R. M., Balzan, R., & McKay, R. (2023). Delusion-like beliefs and data quality: Are classic cognitive biases artifacts of carelessness? *Journal of Psychopathology and Clinical Science*. Advance online publication. <https://doi.org/10.1037/abn0000844>
- Tappin, B. M., & McKay, R. T. (2017). The illusion of moral superiority. *Social Psychological and Personality Science*, *8*(6), 623-631. <https://doi.org/10.1177/1948550616673878>
- Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124-1131. <https://doi.org/10.1126/science.185.4157.1124>
- Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*(4481), 453-458. <https://www.jstor.org/stable/1685855>
- Wilson, T. D., Houston, C. E., Etling, K. M., & Brekke, N. (1996). A new look at anchoring effects: Basic anchoring and its antecedents. *Journal of Experimental Psychology: General*, *125*(4), 387–402. <https://doi.org/10.1037/0096-3445.125.4.387>
- Yoon, S., Fong, N. M., & Dimoka, A. (2019). The robustness of anchoring effects on preferential judgements. *Judgment & Decision Making*, *14*(4), 470-487. <https://EconPapers.repec.org/RePEc:jdm:journl:v:14:y:2019:i:4:p:470-487>

Ziano, I., Mok, P. Y., & Feldman, G. (2021). Replication and extension of Alicke (1985) better-than-average effect for desirable and controllable traits. *Social Psychological and Personality Science*, 12(6), 1005-1017.
<https://doi.org/10.1177/1948550620948973>

Appendix 5A. Pilot of Study 3 and Study 4

We piloted Study 3 and 4 simultaneously with the aim to collect feedback on both studies and test for potential technical issues. We also needed to test whether the way we phrased the tasks was clear and understandable enough to avoid potential data exclusions due to lack of task comprehension. We also used the pilot to determine the average duration of each study to anticipate the financial costs for running Study 3 and Study 4.

Method

Overview

First, all participants ranked themselves, compared to 100 other participants, on two morally relevant personality characteristics (both desirable or both undesirable). In the anchoring conditions, respondents were asked to first indicate whether they think they would rank higher or lower than a given anchor value, followed by an estimate of their specific relative rank. The slider cursor was pre-set at the anchor values of 95 for the high anchor condition and 5 for the low anchor condition. In the control condition, there was no comparison question and participants only provided their self-rankings. The slider cursor in the control condition was initially invisible and appeared once the participants clicked on the slider line.

Next, half of the participants provided their self-rankings on a set of personality traits (piloting the second stage of Study 3) while the other half of the participants were presented with a Dictator Game (piloting the second stage of Study 4). The set of eight personality characteristics were balanced along the levels of the desirability and morality factors; the traits were presented within-subject and in an individual random order for each participant. The respondents assigned to play the Dictator Game were allocated a 100p windfall endowment and had the option to share a portion with another participant. Each dictator was randomly assigned a recipient from the other half of the participants, who received the respective donation (if any). All respondents answered an attention check question (asking participants to place the slider cursor at the midpoint of the slider

scale) and provided data on their age and gender. In addition, all respondents were asked to provide feedback on the study and report any technical issues.

Participants

We recruited 52 participants via the online platform *Prolific* (www.prolific.com). All participants were monolingual English speakers with UK nationality and had a *Prolific* approval rate higher than 90%. We excluded data from 2 participants who failed the attention check question and 8 participants who provided inconsistent answers in the anchoring paradigm (for example, participants who stated that their self-ranking for a given trait is above 5 at the first stage of the anchoring framework, but afterwards proceeded to indicate a self-ranking lower than 5). The final sample comprised of 42 participants (17 females, 24 males, 1 other; $M = 31.26$, $SD = 12.51$).

Participants were paid a flat participation fee (53p, the equivalent of £7.50 per hour) as well as a reward from the Dictator Game. The experiment was self-certified in accordance with the Royal Holloway, University of London Ethics Committee procedure.

Materials and Procedure

The experiment was designed in Qualtrics and conducted online. Participants were randomly assigned to one of the twelve experimental conditions, resulting from the cross-section of the three between-subject variables *anchor*, *desirability*, and *task*.

At the first stage of the experiment, participants ranked themselves either on two desirable traits (*honest*, *considerate*) or on two undesirable traits (*dishonest*, *inconsiderate*). Self-rankings were indicated on a slider scale, ranging from 0 to 100 (no numeric values were displayed on the slider bar, however the corresponding numeric value of the slider cursor was shown just above it and changed dynamically when the slider cursor changed position).

Respondents assigned to provide their self-rankings on a set of personality characteristics: four moral traits, desirable and undesirable (*trustworthy*, *principled*, *manipulative*, *prejudiced*) and four non-moral traits, desirable and undesirable (*creative*,

easy going, illogical, uptight). The eight traits were selected in a way to ensure comparable average desirability levels between the moral and non-moral personality characteristics (Tappin & McKay, 2017).

Participants presented with the Dictator Game (Forsythe et al., 1994) were allocated a 100p windfall endowment and had the option to share a portion with another participant. The Dictator Game donations were also indicated on a slider line, ranging from 0 to 100 (no numeric values were displayed on the slider bar, however the corresponding numeric value of the slider cursor was shown just above it and changed dynamically when the slider cursor moved along the slider scale).

Design and Analysis

The experimental design employed was a 3 (*anchor*: high, low, control, between-subject) x 2 (*desirability*: desirable, undesirable, between-subject) x 2 (*trait*: moral, non-moral, within-subject) x 2 (type of task: self-judgements, dictator game, between-subject) mixed design.

Our primary goal when piloting Study 3 and Study 4 was to gather feedback on how comprehensible the tasks were as well as to test for potential technical issues. While there were no complaints or any technical problems during the experiment, four participants reported having difficulties selecting the correct response in the attention check question. Although the rest of the feedback was positive, we still had to exclude 16% of the sample for inconsistent answers.

Based on the above results, we further improved the way we phrased the self-ranking task and the comparison question. We also refined the Java-script code in Qualtrics to ensure that Studies 3 and 4 would run without any technical issues. The pilot also provided us with data on the potential average duration of Study 3 and Study 4. As Study 3 seemed to take longer on average, we estimated that the most financially efficient way to run the studies would be one after each other, rather than simultaneously.

Appendix 5B. Anchoring task

Please compare yourself to the other participants in this study with respect to how **HONEST** and how **CONSIDERATE** you are.

Specifically, we want to know whether you think you are **MORE HONEST** than at least **5** of the other participants and whether you think you are **MORE CONSIDERATE** than at least **5** of the other participants.

Please take a moment to reflect on your own past behaviour and on the behaviour of others before answering.

Please compare yourself to the other participants in this study with respect to how **HONEST** you are. First, do you think you are **MORE HONEST** than at least **5** of the other participants?

YES, I am MORE HONEST than at least **5** of the other participants.

NO, I am NOT MORE HONEST than at least **5** of the other participants.

Now please compare yourself to the other participants in this study with respect to how **CONSIDERATE** you are. Do you think you are **MORE CONSIDERATE** than at least **5** of the other participants?

YES, I am MORE CONSIDERATE than at least **5** of the other participants.

NO, I am NOT MORE CONSIDERATE than at least **5** of the other participants.

Please move the slider to indicate how **HONEST** you think you are, compared to the other people in the experiment.

I am MORE HONEST than 5 of the other participants.



Please move the slider to indicate how **CONSIDERATE** you think you are, compared to the other people in the experiment.

I am MORE CONSIDERATE than 5 of the other participants.



Appendix 5C. Subsequent self-rankings

You will now be asked to provide your self-rankings on **eight** more personal characteristics.

Again, we're interested in how you think you compare on each personality trait to the other one hundred participants.

Please proceed to the next screen to indicate your rankings.



Please click carefully on the line below to make the slider appear, then move the slider to indicate how **PRINCIPLED** you think you are, compared to the other people in the experiment.

I am MORE PRINCIPLED than ___ of the other participants.



Please click carefully on the line below to make the slider appear, then move the slider to indicate how **PRINCIPLED** you think you are, compared to the other people in the experiment.

I am MORE PRINCIPLED than 77 of the other participants.



Chapter 6. Study 4. Anchoring of Personality Judgements and Its impact on
Subsequent Prosocial Choices

Study 4. Anchoring of Personality Judgements and Its impact on Subsequent
Prosocial Choices

Elitza Ambrus

Bjoern Hartig

Ryan McKay

Manuscript in preparation

Word count: 4197 excluding references and appendices

Abstract

Interindividual differences influence prosocial attitudes, which in turn shape the way individuals interact with each other. Although personality characteristics have traditionally been considered stable in adulthood, research shows that personality traits continue to change throughout the lifespan. Recent research shows that personality judgements are influenced by anchoring when the anchors elevate the self-view. The current study builds on this self-serving anchoring effect on personality judgements and asks if the self-serving anchoring effect carries over to subsequent prosocial choices. Participants (N = 193) first indicated whether they would rank themselves higher than 95 (high anchor) or 5 (low anchor) out of 100 other participants on two morally relevant personality characteristics, before providing their specific self-rankings for each trait. In the second stage of the study, participants played a Dictator Game with the option to share a portion of a 75p windfall endowment with another participant. Contrary to our expectations, the results from the first stage of the study did not provide evidence for a self-serving anchoring effect on personality judgements. However, the analysis of the second stage of the study showed that participants in the enhancing anchor condition donated nearly 15% more than participants in the diminishing anchor condition. Despite some limitations of the study, such as small sample size, our results are novel, interesting, and worthy of future research.

Keywords: self-serving anchoring, personality judgements; prosocial choices

Anchoring of Personality Judgements and Its impact on Subsequent Prosocial Choices

Individuals' interactions with each other are guided by their prosocial or selfish attitudes, which have commonly been assumed as stable with research focussing on investigating whether most individuals possess selfish or prosocial moral intuitions (Knoch & Fehr, 2007; Rand et al., 2014). An alternative line of research, however, has shown that prosocial behaviours are influenced by contextual factors (e.g., List, 2007), strategic reasoning (Rand et al., 2016) and interindividual differences (Yamagishi et al., 2013). Meanwhile, despite the traditional view of personality judgements as stable in adulthood (McCrae et al., 2000), a growing body of literature demonstrates that they continue to change in adulthood too (Roberts et al., 2007; Bleidorn et al., 2022). Personality judgements are a robust predictor of life outcomes (Beck & Jackson, 2022; Roberts et al., 2007; Soto, 2021) and a surge of recent research has supplied empirical evidence on shaping them via nonclinical psychological interventions (e.g., Stieger et al., 2020). Yet, personality traits are considered as relatively stable behavioural patterns over task contexts (Roberts, 2009), which should render them resilient towards cognitive biases. Our previous work (Ambrus et al., under review) however, shows that personality judgements are flexible in response to anchoring when the anchors elevate the self-concept. The present study builds on the self-serving anchoring effect on personality judgements and explores whether it influences behaviour in subsequent prosocial choices.

The anchoring effect refers to the influence of a random salient value (anchor) on a consecutive absolute judgement (Tversky & Kahneman, 1974). Anchoring has been documented across a variety of decision-making domains (Furnham & Boo, 2010; Strack et al., 2016; Yoon et al., 2019) and persists even when participants are explicitly forewarned for its influence (Wilson et al., 1996). The anchoring effect size is not influenced by factors such as monetary incentives, type of experiment (online or lab) or demographic factors (Röseler & Schütz, 2022). The anchoring effect also persists, though in mitigated form, when extreme (Mussweiler, 2001; Strack & Mussweiler, 1997; Wegener et al., 2001) or irrelevant (Glöckner & Englich, 2015; Smith & Windschitl, 2011)

anchors are presented. For instance, Cheek et al. (2015) showed that even an extreme and nonsensical anchor (a negative number) influenced judgements about the number of math problems participants had just solved.

The anchoring effect on personality judgements, however, reveals a self-serving pattern: enhancing anchors have a strong influence while diminishing anchors have little or no effect on personality judgements (Ambrus et al., under review). These findings are in line with research showing that self-relevant information could be anchored (Cheek et al. 2015; Greenberg et al., 2017) and that motivated reasoning might render anchors ineffective if they point towards unfavourable future prospects (Joel et al., 2017). Here, we build on the impact of enhancing anchors on personality judgements and explore potential *aftereffects* on subsequent prosocial behaviour. We anchor participants' personality judgements on morally relevant personality characteristics as moral traits define individuals' self-concept (Molouki & Bartels, 2017; Strohminger & Nichols, 2014; Strohminger, 2018). To measure behaviour, we employ the Dictator Game (Forsythe et al., 1994; Kahneman et al., 1986) as it has been widely used to capture prosocial behaviour and to measure perceptions of fairness (Fehr & Schmidt, 1999).

Furthermore, we chose the traits "honest" and "considerate" (and their respective antonyms "dishonest" and "inconsiderate") as research has shown that the honesty-humility dimension of the HEXACO model of personality (Ashton & Lee 2020, Thalmayer & Saucier 2014) is associated with prosocial behaviour as measured by the Dictator Game (Hilbig et al., 2013; Thielmann et al., 2020; Zettler et al., 2020). The honesty-humility facet of the HEXACO model of personality has been theorised as encompassing prosociality (Ashton & Lee 2014; Zettler et al., 2020) and empirical evidence has linked prosociality to honesty (Isler & Gächter, 2022; Soraperra et al., 2019).

Given the evidence on self-serving anchoring of personality judgements, we expect that personality judgements will again be susceptible to anchoring when the anchors enhance individuals' self-image. In addition, we anticipate that anchoring of self-rankings at the first stage of the experiment will influence donations in the Dictator

Game at the second stage of the experiment, such that enhancing anchors will lead to higher Dictator Game donations than diminishing anchors.

Method

Overview

The study was Qualtrics-based and conducted via the online platform *Prolific* (www.prolific.com). Participants were randomly assigned to one of the six experimental conditions, resulting from the cross-section of the two between-subject variables, anchor and desirability. In the first stage of the experiment, participants ranked themselves, compared to 100 other participants, on two morally relevant personality characteristics. In the anchoring conditions, respondents were asked to first indicate whether they thought they would rank higher or lower than a given anchor value and then provided their self-rankings on both traits. In the no anchor condition, participants indicated their self-rankings on both traits without any preceding comparisons. In the second stage of the study, all participants were presented with a Dictator Game: a 75p windfall endowment was allocated to each participant and they had the option to share a portion with another participant. Each participant received a participation fee for taking part in the experiment and a (potential) reward from the Dictator Game. The respective Dictator Game donations were randomly distributed to participants who took part in a different experiment (Study 3, Chapter 5). Our hypotheses, data collection, and analysis protocol were pre-registered (https://aspredicted.org/PXR_W4K). Deidentified data and analysis are available on the Open Science Framework (https://osf.io/dt9wb/?view_only=764c265388dc45c59fd6dd8d62bdc435).

Participants

We recruited 300 participants in a pre-registered online experiment via the *Prolific* platform. All participants were monolingual English speakers with UK nationality and had a *Prolific* approval rate higher than 90%. As specified in our pre-registration

document, we excluded participants who: (i) failed the attention check question, asking participants to indicate 50, the midpoint of the slider scale (N=84); (ii) provided inconsistent answers in the anchoring paradigm, e.g., participants who indicated they would rank themselves higher than 5 on the trait in question, however subsequently indicated a self-ranking lower than 5, suggesting a lack of comprehension or attention (N=23) and (iii) completed the study in less than 30 seconds (N=0). The final sample comprised 193 participants (137 females, 53 males, 3 other; $M = 33.29$, $SD = 10.72$). The distribution of participants across conditions was as follows: 25 in the high anchor, desirable trait condition; 31 in the high anchor, undesirable trait condition; 34 in the low anchor, desirable trait condition; 24 in the low anchor, undesirable trait condition; 38 in the control, desirable trait condition and 41 participants in the control, undesirable trait condition. Respondents were paid a flat participation fee (70p, the equivalent of £7.50 per hour) as well as the amount that they chose to keep from their initial Dictator Game endowment. The study was self-certified in accordance with the Royal Holloway, University of London Ethics Committee procedure.

Materials and Procedure

In the first stage of the study, all participants were randomly assigned to one of six between-subject conditions, resulting from the cross section of the factors anchor and desirability. Participants then reported their self-rankings on two morally relevant personality traits, either desirable (*honest, considerate*) or undesirable (*dishonest, inconsiderate*). We employed the classic two-stage anchoring paradigm (Tversky & Kahneman, 1974): first, respondents answered a comparative question about whether they would rank themselves higher than at least 95 other participants (out of 100) in the high anchor condition or higher than at least 5 other participants in the low anchor group (the slider cursors were pre-set at 95 and 5 respectively) on the two morally relevant traits. Next, participants indicated their specific personality judgements on these two traits. The comparison questions for both traits were presented at the same screen, followed by indications of the specific personality judgements. Self-rankings

were indicated on a slider bar, ranging from 0 to 100 (no numeric values were displayed on the slider bar, only the corresponding numeric value of the slider cursor was shown just above it and changed dynamically when the slider cursor moved along the slider line). In the no anchor condition, the slider cursor and its corresponding value appeared only when participants clicked on the slider line to indicate their self-rankings. The first stage of the present study is identical to the first stage of Study 3 (Chapter 5), please see Appendix 5B for screenshots of the anchoring task.

In the second stage of the study, participants played a Dictator Game (Kahneman et al., 1986). All participants played as dictators, they received a 75p budget and had the opportunity to share part of it with anonymous partner. The donations in the Dictator Game were indicated on a slider scale, ranging from 0 to 75. No numeric values were displayed on the slider line, however, if participants moved the slider cursor, a text displayed above the slider bar dynamically showed the chosen amount to keep and the corresponding amount that would be given to the other participant (the amounts shown moved in steps of 5p). The phrasing of the Dictator Game choice was counterbalanced, i.e., when moving the slider cursor, half of the participants were presented a text stating first the amount they were to keep to themselves, followed by the amount they were to donate; and vice versa for the other half of the participants (Appendix 6A). The Dictator Game was played with real stakes, to ensure full engagement with the task and to better capture real-life behaviour (Amir et al., 2016).

Design and Analysis

The first stage of the experiment employed a 3 (*anchor*: high, low, no anchor) x 2 (*desirability*: desirable, undesirable) between-subject design and the DV was self-rankings (measured on a 0 to 100 scale). The self-rankings on undesirable traits were reverse coded (i.e., subtracting the provided self-rankings from 100). We then fitted a linear model for the personality judgements measure in the first stage of the experiment to test for a self-serving anchoring effect.

To investigate whether self-serving anchoring had an *aftereffect* in the second stage of the study, we run t-tests comparing the average donations in the Dictator Game (measured on a 0 to 75 scale) across conditions. All analyses were conducted via R Studio 3.5.1, using the lme4 package for the linear model (Bates et al., 2015).

Results

Anchoring at the first stage of the experiment

To analyse the data we first constructed two dummy variables: *enhancing* anchor, denoting anchors elevating the self-view, this variable took a value of 1 for a high anchor on a desirable trait or a low anchor on an undesirable trait and 0 otherwise; and *diminishing* anchor, denoting anchors diminishing the self-image, this variable took a value of 1 for a high anchor on an undesirable trait or a low anchor on a desirable trait and 0 otherwise. There were 49 participants in the *enhancing* anchor condition, 65 in the *diminishing* anchor condition and 79 respondents in the no anchor condition. Next, a linear mixed model was fitted to analyse the self-rankings measure with *enhancing* anchor and *diminishing* anchor modelled as fixed effects and participants as a random effect.

The analysis showed that the coefficient for *enhancing* anchor was non-significant ($p = .630$) while the coefficient for *diminishing* anchor was negative and significant ($p = .024$), see Table 1 (please see also Figure 1 for violin plots). The results did not support a stronger effect of *enhancing* anchors on self-rankings, i.e., a self-serving anchoring effect on self-ranking: the coefficient for *enhancing* anchor is positive, however it did not reach significance within our sample size. The coefficient for *diminishing* anchor is negative, resulting in lower self-rankings than in the no anchor condition, which is in line with our expectations. We run a Wald test for equality of coefficients to formally compare the effect of *enhancing* and *diminishing* anchors, which showed that, as expected, on average self-rankings in the *diminishing* anchor condition were lower than self-rankings in the *enhancing* anchor condition, $W(1) = 6.15$, $p = .013$. Yet, we anticipated a strong elevating effect of *enhancing* anchors on self-rankings.

Table 1. Estimated fixed effects, the effect of *enhancing* and *diminishing* anchors on self-rankings at the first stage of the experiment

	Estimate	SE	t-value	p-value
Intercept	78.40	2.04	38.45	< .001
<i>Enhancing</i> anchor	1.59	3.30	0.48	0.630
<i>Diminishing</i> anchor	-6.91	3.04	-2.28	0.024

Baseline level: no anchor; Number of observations: 386; grouped by participants, N=193

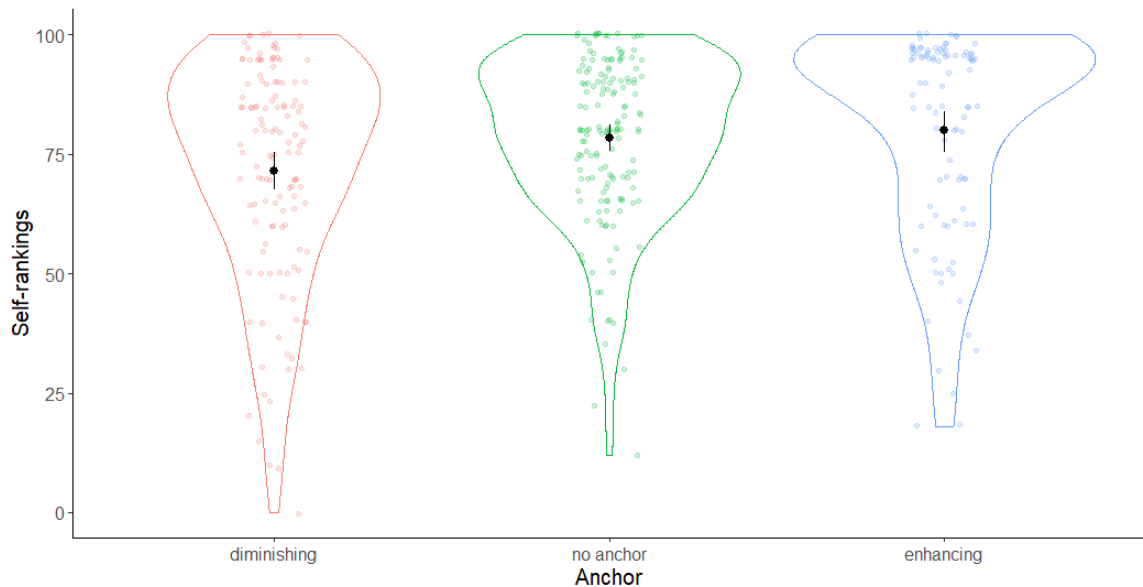


Figure 1. Violin plots for the self-rankings measure by anchor condition.

Anchoring aftereffect at the second stage of the experiment

To test whether *enhancing* and *diminishing* anchors had any *aftereffect* on donations in the Dictator Game (H3), we ran t-tests to compare the average donations in the *enhancing*, *diminishing* and the no anchor conditions (Table 2). The analysis showed that participants in the *enhancing* anchor condition ($M = 38.27$, $SD = 12.89$) donated more on average than participants in both the *diminishing* anchor ($M = 33.38$, $SD = 15.82$), *adjusted p* = .033, and no anchor conditions ($M = 34.24$, $SD = 10.90$), *adjusted p* = .033, Table 2. *Enhancing* anchors on self-rankings led to Dictator Game

donations that were nearly 15% higher than those following *diminishing* anchors. There was no significant difference between the average Dictator Game donations in the *diminishing* anchor ($M = 33.38$, $SD = 15.82$) and no anchor conditions ($M = 34.24$, $SD = 10.90$), $p = .601$, Table 2.

Table 2. T-tests comparing Dictator Game donations across anchoring conditions

	Condition	Condition	t-value	df	p-value	p-adjusted
Dictator	<i>enhancing</i>	<i>diminishing</i>	-2.57	222	0.011	0.033
Game	<i>enhancing</i>	no anchor	-2.57	225	0.011	0.033
Donations	<i>diminishing</i>	no anchor	-0.52	180	0.601	0.601

Number of observations: 386, $N = 193$

Exploratory analysis. The effect of the way the personality traits are phrased (positively or negatively)

In the first stage of the experiment, we employed pairs of the same personality traits, either positively or negatively phrased (*honest* or *dishonest* and *considerate* or *inconsiderate*), comprising desirable and undesirable traits respectively. As self-rankings on undesirable traits were reverse scored for the analyses (subtracting the provided self-rankings from 100), we expected similar average self-rankings on desirable traits (e.g., *honest*) and reverse-scored undesirable traits (e.g., *dishonest*) in the same experimental condition. We tested the data for systematic differences in self-rankings due to the way the personality traits were phrased (positively or negatively) by fitting a linear mixed model on the self-rankings measure provided in the first stage of the experiment. We included *enhancing* anchor, *diminishing* anchor, desirability as well as their interaction as fixed factors and participants as a random effect (Table 3).

Table 3. Estimated fixed effects for *anchor*, *desirability* and their interaction on self-rankings

	Estimate	SE	t-value	p-value
Intercept	84.67	2.63	32.20	< .001
<i>Enhancing</i> anchor	0.68	4.33	0.16	0.875
<i>Diminishing</i> anchor	-4.30	4.01	-1.07	0.285
Desirable traits	-13.04	3.79	-3.44	< .001
Interaction <i>Enhancing</i> anchor*Desirable trait	2.53	6.12	0.41	0.681
Interaction <i>Diminishing</i> anchor*Desirable trait	-3.95	5.64	-0.70	0.485

Baseline: control, undesirable traits, Number of observations: 386; grouped by participants, N=193

The effect of desirability was negative and highly significant ($p < .001$), showing that when reverse-scored the average self-rankings on traits phrased in a negative way were higher than the average self-rankings on positively phrased traits. To follow up on the significant effect of desirability we performed post-hoc binary comparisons. The data showed that in the control and the diminishing conditions, participants ranked themselves with 18-27% higher for negatively phrased traits (dishonest, inconsiderate) than for positively phrased traits (honest, considerate), please see Figure 2 for bar plots.

Specifically, in the no anchor condition, participants indicated higher self-rankings on reverse-scored undesirable traits ($M = 84.67$, $SD = 15.58$) than on desirable traits ($M = 71.63$, $SD = 16.53$), $p = .009$. In a similar vein, self-rankings were higher for *diminishing* anchor, undesirable traits ($M = 80.37$, $SD = 18.51$) than for *diminishing* anchor, desirable traits ($M = 63.38$, $SD = 23.78$), $p = .001$. However, self-rankings for *enhancing* anchor, undesirable traits ($M = 85.35$, $SD = 19.09$) were not significantly different than self-rankings for *enhancing* anchor, desirable traits ($M = 74.84$, $SD = 22.29$), $p = .250$. The phrasing effect in the no anchor and *diminishing* anchor conditions replicates our results from the exploratory analysis in Study 3 (Chapter 5).

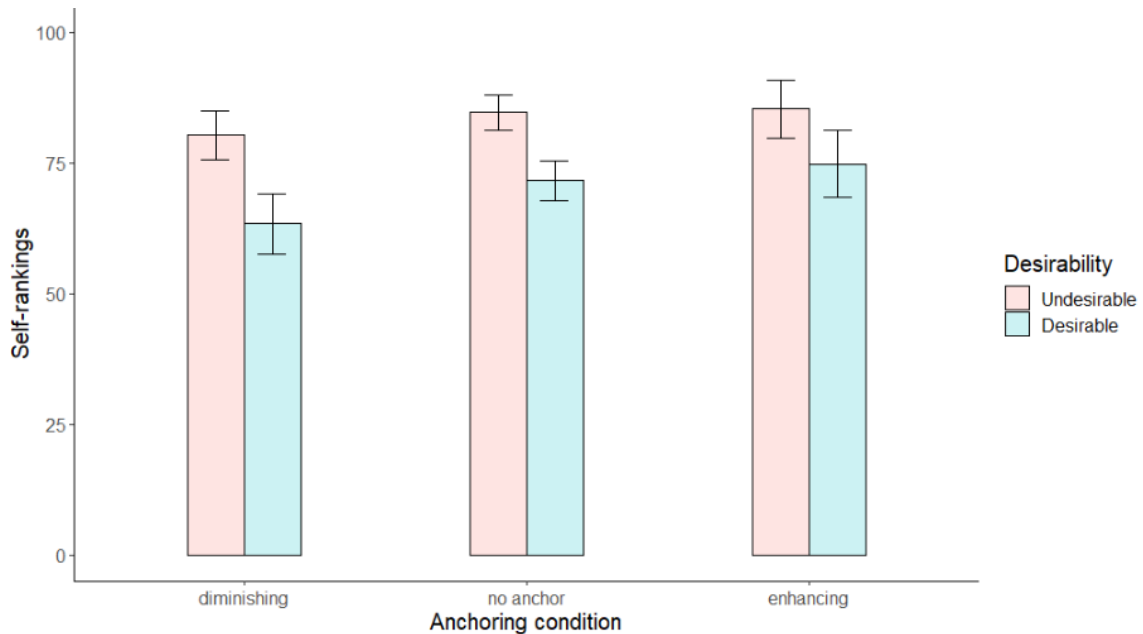


Figure 2. Bar plot of self-rankings by anchoring conditions.

Discussion

The anchoring effect is a robust and replicable cognitive phenomenon, documented across a variety of decision-making domains (Röseler & Schütz, 2022) and persists even when participants are presented with explicit instructions to correct for it (Wilson et al., 1996). Recent research shows that personality judgements are prone to anchoring when the anchors enhance the self-view (Ambrus et al., under review). This study investigated whether the self-serving anchoring effect on personality judgements affects subsequent prosocial choices. The results in the first stage of the experiment did not replicate the self-serving anchoring effect: the effect of *enhancing* anchors was not significant in our sample size, yet the coefficient was positive, which is consistent with our expectations. In line with our predictions, *diminishing* anchors led to lower self-rankings than in the *enhancing* anchor and no anchor conditions.

In the second stage of the experiment, participants in the *enhancing* anchor condition donated nearly 15% more than participants in the *diminishing* anchor condition. This result demonstrates an *aftereffect* of enhancing anchors on subsequent behaviour and seems promising as it relies on measuring revealed behaviour, which is

the most robust way to show the potential effect of a manipulation (Back & Vazire, 2012). *Diminishing* anchors did not lead to less prosocial behaviour in the Dictator Game, which is in line with literature showing the ineffectiveness of negative self-relevant information in influencing attitudes and beliefs (Joel et al., 2017; Möbius et al., 2022; Ritchie et al., 2017; Sedikides & Green, 2009; Sedikides et al., 2016).

At the first stage of the experiment, the data did not replicate the self-serving anchoring effect. As we based our predictions on our previous findings (Ambrus et al., under review), it is important to understand whether the results from the current study undermine findings on self-serving anchoring or whether there are other factors that might have interfered with our manipulation in the current study. First, due to the large percentage of data exclusions and the between-subject design we employed, we had a relatively small sample per condition (ranging from 49 to 79 participants per condition). Second, despite all the precautionary measures we took to ensure high quality data, the fact that we had to exclude a third of our data (our final sample comprised 193 out of 300 recruited respondents), somewhat undermines our confidence in the quality of the rest of the data too. Third, we could more effectively filter out inattentive participants from the anchoring conditions than from the control condition as we had both attention and comprehension checks in the former, while the control condition relied on the attention check only. Inattentive participants might have raised the average self-rankings in the control condition in a way that renders the coefficient for enhancing anchors non-significant. The study presented in Chapter 5 had an identical first stage and the average self-rankings in the no anchor condition were lower than the average self-rankings in the no anchor condition in the current study. Indeed, recent research has shown that excluding inattentive participants from the sample might substantially change the results (Sulik et al., 2023).

In addition, *enhancing* anchors led to more generous donations in the subsequent Dictator Game, suggesting that there was some elevating effect of the *enhancing* anchor on self-rankings and the general self-image, which was carried over to subsequent prosocial behaviour. The Dictator Game played at the second stage of the

study was incentivised, which might have also better engaged the attention of participants from the control condition, allowing us to distinguish the effect of *enhancing* anchors. This elevating effect of *enhancing* anchors on donations is in line with literature showing positive correlations between self-rankings on prosocial personality traits and prosocial behaviour (e.g., Zettler et al., 2020, cf. Tappin & McKay, 2019). Indeed, the traits we chose, honest and considerate, are correlated with prosocial choices as captured by the Dictator Game task (Thielmann et al., 2020). In retrospect, however, we realised the challenges involved in relying on self-assessments of how honest one is as dishonest individuals would also be dishonest about their self-rankings on honesty (Hilbig, 2022). Possibly, relying on different personality traits (e.g., generous) might have revealed even stronger aftereffect of enhanced self-rankings on subsequent donations.

Our findings also support research showing the ineffectiveness of anchoring self-relevant information in an undesirable direction (Ambrus et al., under review; Joel et al., 2017). At the second stage of the experiment, self-image concerns might have outweighed the effect of diminishing anchors, leading to the observed lack of change in subsequent prosocial behaviour in comparison with the no anchor condition. Although the current study required participants to indicate their personality judgements on two morally relevant personality traits only, *enhancing* anchors led to higher donations than *diminishing* anchors. It is possible that the *aftereffect* would have been even stronger if we employed eight personality traits (Ambrus et al., under review) or a set of nine central moral personality characteristics, which were shown to elicit association with the holistic moral identity (Aquino & Reed, 2002). It is an interesting question to explore whether enhancement of a few narrower traits, such as for instance “kind” and “generous” or a set of personality traits that would represent a more holistic self-view is needed to achieve stronger influence on one’s self-image and subsequent behaviour.

Furthermore, we conducted exploratory analyses on the effect of the way the personality traits are phrased (positive or negative) on personality judgements. As we reverse-scored the undesirable traits for the analysis, there should not have been

significant differences between average personality judgements on the same trait depending on the way it is phrased. However, we found highly significant differences between the average personality judgements on the same traits phrased in a positive or negative way in the no anchor and *diminishing* anchor conditions. These findings replicate the results from the corresponding analysis in Study 3 (Chapter 5). The phrasing effect seems similar to the loss aversion effect with losses having larger impact than gains (Kahneman & Tversky, 1979). The question of how “dishonest” one is, might be perceived as a potential “loss” in participants’ self-image, which triggers a stronger desire to state favourable self-rankings than when the question is phrased in terms of “honesty”. The overstatement of positive self-image might also stem from the stronger impact of self-protecting than self-enhancing motivation (Alicke & Sedikides, 2009). The phrasing effect is worthy of further exploration; for instance, future research might explore whether this effect is restricted to morally relevant traits only or whether it is also obtained with non-morally relevant traits.

In summary, we did not replicate the self-serving anchoring effect in this study: *enhancing* anchors had a positive coefficient, however it did not reach significance within our sample size, while *diminishing* anchors had a negative and significant coefficient. Nevertheless, we found evidence for more generous donations in the Dictator Game after participants were exposed to enhancing manipulations. Therefore, our results indicate not only that personality judgements are susceptible to anchoring, but that these quick changes to self-assessment are transferrable to subsequent behaviour. Although our results should be taken with caution due to the above-discussed limitations of the study (such as for example the massive loss of data we experienced), our findings are novel, interesting and worthy of further research. We also relied on self-rankings on two morally relevant personality traits only (in comparison with eight personality characteristics in our previous research), which suggests that the *aftereffect* we found might be even stronger should we employ a larger set of personality traits. Future research might explore whether targeting narrow traits related

to fairness and generosity or a broader range of personality traits is needed to facilitate more pronounced change in prosocial behaviour.

References

- Alicke, M. D., & Sedikides, C. (2009). Self-enhancement and self-protection: What they are and what they do. *European review of social psychology*, 20(1), 1-48.
<https://doi.org/10.1080/10463280802613866>
- Ashton, M. C., Lee, K., & de Vries, R. E. (2014). The HEXACO Honesty-Humility, agreeableness, and emotionality factors: A review of research and theory. *Personality and Social Psychology Review*, 18(2), 139–152.
<https://doi.org/10.1177/1088868314523838>
- Ashton MC, Lee K, De Vries RE: The HEXACO honestyhumility, agreeableness, and emotionality factors: a review of research and theory. *Pers Soc Psychol Rev* 2014, 18:139–152, <https://doi.org/10.1177/1088868314523838>
- Ambrus, E. Z., Hartig, B., & McKay, R. T. (2023). Self-serving anchoring of personality judgements. *Under review*
- Amir, O., & Rand, D. G. (2012). Economic games on the internet: The effect of \$1 stakes. *PloS one*, 7(2). <https://doi.org/10.1371/journal.pone.0031461>
- Aquino, K., & Reed, A. II. (2002). The self-importance of moral identity. *Journal of Personality and Social Psychology*, 83(6), 1423–1440.
<https://doi.org/10.1037/0022-3514.83.6.1423>
- Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. Retrieved from <https://arxiv.org/pdf/1506.04967.pdf>
- Beck, E. D., & Jackson, J. J. (2022). A mega-analysis of personality prediction: Robustness and boundary conditions. *Journal of Personality and Social Psychology*, 122(3), 523– 553. <https://doi.org/10.1037/pspp0000386>
- Bleidorn, W., Schwaba, T., Zheng, A., Hopwood, C. J., Sosa, S. S., Roberts, B. W., & Briley, D. A. (2022). Personality stability and change: A meta-analysis of longitudinal

- studies. *Psychological bulletin*, 148(7-8), 588
<https://doi.org/10.1037/bul0000365>
- Brown, J. D. (2012). Understanding the better than average effect: Motives (still) matter. *Personality and Social Psychology Bulletin*, 38(2), 209-219.
<https://doi.org/10.1177/0146167211432763>
- Camerer, C. F. (2003). Behavioural studies of strategic thinking in games. *Trends in cognitive sciences*, 7(5), 225-231. [https://doi.org/10.1016/S1364-6613\(03\)00094-9](https://doi.org/10.1016/S1364-6613(03)00094-9)
- Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, 114(3), 817-868. <https://doi.org/10.1086/519249>
- Forsythe, R., Horowitz, J. L., Savin, N. E., & Sefton, M. (1994). Fairness in simple bargaining experiments. *Games and Economic behavior*, 6(3), 347-369.
<https://doi.org/10.1006/game.1994.1021>
- Furnham, A., & Boo, H. C. (2011). A literature review of the anchoring effect. *The Journal of Socio-Economics*, 40(1), 35-42. <https://doi.org/10.1016/j.socec.2010.10.008>
- Glöckner, A., & Englich, B. (2015). When relevance matters. *Social Psychology*, 46(1), 4-12. <https://doi.org/10.1027/1864-9335/a000214>
- Heck, D. W., Thielmann, I., Moshagen, M., & Hilbig, B. E. (2018). Who lies? A large-scale reanalysis linking basic personality traits to unethical decision making. *Judgment and Decision making*, 13(4), 356-371.
<https://doi.org/10.1017/S1930297500009232>
- Hilbig, B. E. (2022). Personality and behavioral dishonesty. *Current Opinion in Psychology*, 101378. <https://doi.org/10.1016/j.copsyc.2022.101378>
- Hilbig, B. E., Zettler, I., Leist, F., & Heydasch, T. (2013). It takes two: Honesty–humility and agreeableness differentially predict active versus reactive

cooperation. *Personality and Individual Differences*, 54(5), 598–603.

<https://doi.org/10.1016/j.paid.2012.11.008>

Isler, O., & Gächter, S. (2022). Conforming with peers in honesty and cooperation. *Journal of Economic Behavior & Organization*, 195, 75-86.

<http://dx.doi.org/10.2139/ssrn.4114486>

Larney, A., Rotella, A., & Barclay, P. (2019). Stake size effects in ultimatum game and dictator game offers: A meta-analysis. *Organizational Behavior and Human Decision Processes*, 151, 61-72. <https://doi.org/10.1016/j.obhdp.2019.01.002>

Lucas, R. E., & Donnellan, M. B. (2011). Personality development across the life span: longitudinal analyses with a national sample from Germany. *Journal of Personality and Social Psychology*, 01(4), 847-861.

<https://doi.org/10.1037/a0024298>

Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1986). Fairness and the assumptions of economics. *Journal of business*, S285-S300.

<https://www.jstor.org/stable/2352761>

Knoch, D., & Fehr, E. (2007). Resisting the power of temptations: the right prefrontal cortex and self-control. *Annals of the New York Academy of Sciences*, 1104(1), 123-134. <https://doi.org/10.1196/annals.1390.004>

List, J. A. (2007). On the interpretation of giving in dictator games. *Journal of Political economy*, 115(3), 482-493. <https://doi.org/10.1086/519249>

McCrae, R. R., Costa, P. T. J., Jr., Ostendorf, F., Angleitner, A., Hrebícková, M., Avia, M. D., Sanz, J., Sánchez-Bernardos, M. L., Kusdil, M. E., Woodfield, R., Saunders, P. R., & Smith, P. B. (2000). Nature over nurture: Temperament, personality, and life span development. *Journal of Personality and Social Psychology*, 78(1), 173–186. <https://doi.org/10.1037/0022-3514.78.1.173>

- Malle, B. F. (2006). The actor-observer asymmetry in attribution: A (surprising) meta-analysis. *Psychological Bulletin*, 132(6), 895-919. <https://doi.org/10.1037/0033-2909.132.6.895>
- Möbius, M. M., Niederle, M., Niehaus, P., & Rosenblat, T. S. (2022). Managing self-confidence: Theory and experimental evidence. *Management Science* 68(11). <https://doi.org/10.1287/mnsc.2021.4294>
- Molouki, S., & Bartels, D. M. (2017). Personal change and the continuity of the self. *Cognitive Psychology*, 93, 1-17. <https://doi.org/10.1016/j.cogpsych.2016.11.006>
- Mussweiler, T. (2001). The durability of anchoring effects. *European Journal of Social Psychology*, 31, 431–442. <https://doi.org/10.1002/ejsp.52>
- Rand, D. G., Peysakhovich, A., Kraft-Todd, G. T., Newman, G. E., Wurzbacher, O., Nowak, M. A., & Greene, J. D. (2014). Social heuristics shape intuitive cooperation. *Nature Communications*, 5(1), 1-12. <http://dx.doi.org/10.2139/ssrn.2222683>
- Ritchie, T. D., Sedikides, C., & Skowronski, J. J. (2017). Does a person selectively recall the good or the bad from their personal past? It depends on the recall target and the person's favourability of self-views. *Memory*, 25(8), 934-944. <https://doi.org/10.1080/09658211.2016.1233984>
- Rhodes, R. E., & Dickau, L. (2012). Experimental evidence for the intention–behavior relationship in the physical activity domain: A meta-analysis. *Health Psychology*, 31(6), 724. <https://doi.org/10.1037/a0027290>
- Roberts, B. W. (2009). Back to the future: Personality and assessment and personality development. *Journal of research in personality*, 43(2), 137-145. <https://doi.org/10.1016/j.jrp.2008.12.015>
- Roberts, B. W., Kuncel, N. R., Shiner, R., Caspi, A., & Goldberg, L. R. (2007). The power of personality: The comparative validity of personality traits, socioeconomic status,

- and cognitive ability for predicting important life outcomes. *Perspectives on Psychological Science*, 2(4), 313-345. <https://www.jstor.org/stable/40212212>
- Röseler, L., & Schütz, A. (2022, March 9). Hanging the Anchor Off a New Ship: A Meta-Analysis of Anchoring Effects. PsyArXiv. <https://doi.org/10.31234/osf.io/wf2tn>
- Sedikides, C., & Green, J. D. (2009). Memory as a self-protective mechanism. *Social and Personality Psychology Compass*, 3(6), 1055–1068. <https://doi.org/10.1111/j.1751-9004.2009.00220.x>
- Sedikides, C., Green, J. D., Saunders, J., Skowronski, J. J., & Zengel, B. (2016). Mnemic neglect: Selective amnesia of one's faults. *European Review of Social Psychology*, 27, 1–62. <https://doi.org/10.1080/10463283.2016.1183913>
- Sheeran, P., & Webb, T. L. (2016). The intention–behavior gap. *Social and personality psychology compass*, 10(9), 503-518. <https://doi.org/10.1111/spc3.12265>
- Smith, A. R., & Windschitl, P. D. (2011). Biased calculations: Numeric anchors influence answers to math equations. *Judgment and Decision Making*, 6, 139–146. <https://doi.org/10.1017/S1930297500004083>
- Soraperra, I., Weisel, O., & Ploner, M. (2019). Is the victim Max (Planck) or Moritz? How victim type and social value orientation affect dishonest behavior. *Journal of Behavioral Decision Making*, 32(2), 168-178. <https://doi.org/10.1002/bdm.2104>
- Strack, F., Bahnik, Š., & Mussweiler, T. (2016). Anchoring: accessibility as a cause of judgmental assimilation. *Current Opinion in Psychology*, 12, 67-70. <https://doi.org/10.1016/j.copsyc.2016.06.005>
- Strack, F., & Mussweiler, T. (1997). Explaining the enigmatic anchoring effect: Mechanisms of selective accessibility. *Journal of Personality and Social Psychology*, 73, 437–446. <https://doi.org/10.1037/0022-3514.73.3.437>
- Strohinger, N. (2018). Identity is essentially moral. *Atlas of moral psychology*, 141-148. <https://doi.org/10.1016/j.cognition.2013.12.005>

- Strohminger, N., & Nichols, S. (2014). The essential moral self. *Cognition*, *131*(1), 159–171. <https://doi.org/10.1016/j.cognition.2013.12.005>
- Soto, C. J. (2021). Do links between personality and life outcomes generalize? Testing the robustness of trait–outcome associations across gender, age, ethnicity, and analytic approaches. *Social Psychological and Personality Science*, *12*(1), 118-130. <https://doi.org/10.1177/1948550619900572>
- Sulik, J., Ross, R. M., Balzan, R., & McKay, R. (2023). Delusion-like beliefs and data quality: Are classic cognitive biases artifacts of carelessness? *Journal of Psychopathology and Clinical Science*. Advance online publication. <https://doi.org/10.1037/abn0000844>
- Stieger, M., Wepfer, S., Rügger, D., Kowatsch, T., Roberts, B. W., & Allemand, M. (2020). Becoming more conscientious or more open to experience? Effects of a two-week smartphone-based intervention for personality change. *European Journal of Personality*, *34*(3), 345–366. <https://doi.org/10.1002/per.2267>
- Tappin, B. M., & McKay, R. T. (2019). Investigating the relationship between self-perceived moral superiority and moral behavior using economic games. *Social Psychological and Personality Science*, *10*(2), 135-143. <https://doi.org/10.1177/1948550617750736>
- Thielmann, I., Spadaro, G., & Balliet, D. (2020). Personality and prosocial behavior: A theoretical framework and meta-analysis. *Psychological Bulletin*, *146*(1), 30–90. <https://doi.org/10.1037/bul0000217>
- Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124-1131. <https://doi.org/10.1126/science.185.4157.1124>
- Wegener, D. T., Petty, R. E., Detweiler-Bedell, B. T., & Jarvis, W. B. G. (2001). Implications of attitude change theories for numerical anchoring: Anchor

plausibility and the limits of anchor effectiveness. *Journal of Experimental Social Psychology*, 37, 62–69. <https://doi.org/10.1006/jesp.2000.1431>

Wilson, T. D., Houston, C. E., Etling, K. M., & Brekke, N. (1996). A new look at anchoring effects: Basic anchoring and its antecedents. *Journal of Experimental Psychology: General*, 125(4), 387–402. <https://doi.org/10.1037/0096-3445.125.4.387>

Yamagishi, T., Mifune, N., Li, Y., Shinada, M., Hashimoto, H., Horita, Y., ... & Simunovic, D. (2013). Is behavioral pro-sociality game-specific? Pro-social preference and expectations of pro-sociality. *Organizational Behavior and Human Decision Processes*, 120(2), 260-271. <https://doi.org/10.1016/j.obhdp.2012.06.002>

Yoon, S., Fong, N. M., & Dimoka, A. (2019). The robustness of anchoring effects on preferential judgements. *Judgment & Decision Making*, 14(4), 470-487. <https://EconPapers.repec.org/RePEc:jdm:journl:v:14:y:2019:i:4:p:470-487>

Zettler, I., Thielmann, I., Hilbig, B. E., & Moshagen, M. (2020). The nomological net of the HEXACO model of personality: A large-scale meta-analytic investigation. *Perspectives on Psychological Science*, 15(3), 723-760. <https://doi.org/10.1177/1745691619895036>

Appendix 6A. Dictator Game

We will now ask you to make a decision affecting you and another participant. The decision involves **REAL** money (i.e., your decision is **NOT** hypothetical).

You will be paired at random with another person in this study. The other person cannot earn any extra money except from whatever you allocate to them. You will not learn their identity, and they will not learn yours.

We have allocated a budget of 75p (£0.75). Your task is to choose how to split this money between yourself and the other participant.

Whatever split you choose will be directly paid out to you and to the other participant, i.e., it is added as a bonus to the payment both of you receive at the end of the experiment.

Please proceed to the next screen to make your decision.



Please click directly on the line below to make the slider appear, then move the slider to indicate how you would like to split the 75p between yourself and the other participant.

___ p goes to you and ___ p goes to the other
participant.

Chapter 7. Study 5. A Blind Spot for Flattery: People are More Receptive to
Enhancing than Diminishing Manipulations of their Personal Qualities

Study 5. A Blind Spot for Flattery: People are More Receptive to Enhancing than
Diminishing Manipulations of their Personal Qualities

Elitza Ambrus

Bjoern Hartig

Petter Johansson

Ryan McKay

Submitted for publication at the Social Psychological and Personality Science

Word count: 5180 excluding references and appendices

Abstract

Motivation shapes the way people process self-relevant information, with previous research highlighting both self-consistency and self-enhancement motivations. Here, we employ *choice blindness* manipulations to explore how the interplay between these competing motivations affects personality judgements' flexibility. Participants (N = 535) ranked themselves relative to others on eight personality traits (0-100 slider). Participants then contemplated their self-rankings for four out of these traits, however two of the selected self-rankings were inconspicuously manipulated up or down by 20 units, yielding an enhanced or diminished self-evaluation. Participants then had the opportunity to revise their (ostensible) self-rankings. The data showed that the manipulation influenced the revised self-rankings, with most participants accepting the manipulated values or adjusting them only slightly. This effect, however, was moderated by the manipulation's nature – participants were less likely to correct flattering than diminishing manipulations. This self-serving flexibility of personality judgements cannot be easily accounted for by existing personality theories.

Keywords: Personality judgements; Choice Blindness; Self-Serving Bias; Self-enhancement; Self-consistency.

A Blind Spot for Flattery: People are More Receptive to Enhancing than Diminishing Manipulations of their Personal Qualities

Flattery, some say, will get you everywhere. The adage underscores how receptive we are to *enhancing* construals of our prowess and prospects (e.g., Sedikides et al., 2003). At the same time, however, there is merit in having a consistent self-concept, which allows predictability and facilitates social interactions (Epstein, 1973; Swann, 1983). But what happens when these two motives – self-enhancement and self-consistency – come into conflict? By leveraging the “choice blindness” paradigm (Johansson et al., 2005), which starkly illustrates how people strive to be consistent with their perceived earlier selves, we take a novel approach to this question. Specifically, we expose participants to enhancing or diminishing choice-blindness manipulations of their self-reported personality judgements and test whether the enhancing manipulations are accepted to a higher degree.

Self-consistency and Self-enhancement of Personality Judgements

The way we process self-relevant information is distorted by our motives (Sedikides & Strube, 1997; Vaughan-Johnston & Jacobson, 2020). Two prominent motives in the literature are self-consistency and self-enhancement, yet the dynamic interplay between them is not well understood (Sedikides & Strube, 1997; Szumowska et al., 2023; Westerwick et al., 2023). For instance, self-consistency motivation seems to prevail in some cases with individuals preferring feedback that is consistent with their existing self-views (Hixon & Swann, 1993; Swann & Pelham, 2002). At the same time, there is plentiful evidence of self-enhancement motivation shaping behaviour, resulting in self-serving biases, such as for instance the better-than-average effect (Alicke & Govorun, 2005).

The central assumption of the self-consistency theory is that individuals prefer information that is in line with their existing self-views. A stable self-concept ensures

predictability and controllability, facilitating safe social interactions (“self-verification” theory, Swann, 1983; Swann, 1987; Swann, 1990). Personality judgements are commonly considered to be stable and consistent from young adulthood onwards (Costa & McCrae, 1992). Even if some degree of flexibility of personality judgements throughout the lifespan is allowed for (Bleidorn et al., 2022; Roberts et al., 2007; Roberts & Yoon, 2022; Soto et al., 2021), personality traits are still defined as behavioural patterns that are largely consistent within a particular context (Roberts, 2009). Personality judgements seem to be subject to change only when participants themselves are motivated and ready to commit to weeks of psychological interventions to achieve the change they seek (Bleidorn et al., 2021; Olaru et al., 2022; Stieger et al., 2020, see Allemand & Flückiger, 2022 for a review).

In contrast, at the core of self-enhancement theory lies the assumption that individuals are motivated to elevate their self-view (Tesser, 1988). Holding a positive self-image is considered paramount for our wellbeing (Alicke et al., 2013; Leary, 2007; Taylor & Brown 1988; Taylor et al., 2003), and various cognitive processes may conspire to achieve such a self-image. For instance, autobiographical memory functions in a biased manner to ensure individuals perceive themselves as benevolent and effective (“beneffectance”, Greenwald, 1980): while positive self-relevant information is readily obtained and recalled, negative self-relevant information is avoided or forgotten (Alicke & Sedikides, 2009; Gaertner et al., 2012; Sedikides & Strube, 1997; Zhang et al., 2018).

In a similar vein, people find it easier to recall positive than negative previous behaviours (Ritchie et al., 2017; Sedikides & Green, 2009; Sedikides et al., 2016). Moreover, participants tend to misremember behaviours that fall short of their own fairness standards even when incentivised for accuracy of memory (Carlson et al., 2020). Individuals are also able to maintain a positive self-view by attributing their misdeeds to contextual factors (Malle et al., 2019). Moreover, Joel et al. (2017) showed that in making predictions about the likelihood of future life events or outcomes, individuals ignore or discount information suggesting a high probability of undesirable future events.

The above-discussed psychological mechanisms employed in constructing and maintaining a stable and positive self-view reveal self-serving attributions. Indeed, individuals' judgements and decisions are known to be susceptible to various self-serving biases, and people are even biased about their own susceptibility to such biases ("the bias blind spot", Pronin et al., 2002; Pronin & Hazel, 2023). In particular, the mechanism of self-image update is reminiscent of self-serving biases in belief formation, such as the asymmetric update of beliefs (Lefebvre et al., 2017; Sharot & Garrett, 2016, cf. Burton et al., 2022; Shah et al., 2016). For instance, Möbius et al. (2022) showed that while participants update their self-beliefs less than a perfect Bayesian would in response to both positive and negative feedback, the update is asymmetric: positive feedback is overrated, and negative feedback is underrated (see also Eil & Rao, 2011; Korn et al., 2022). Here, in an attempt to reconcile the potential conflict between self-consistency and self-enhancement motives in constructing and maintaining our self-image, we explore the possibility that the interplay between these two motives shapes personality judgements such that they are flexible enough to respond to adjustments elevating the self-view (self-enhancement motive), but rigid enough to resist adjustments that diminish the self-concept (self-consistency motive).

The Choice Blindness Paradigm

Research using the choice blindness paradigm (Johansson et al., 2005) highlights the lengths people will go to maintain consistency in their choices and judgements. In the original choice blindness study, participants were shown two face images and asked to indicate which one they found more attractive (Johansson et al., 2005). Respondents were then asked to justify the image they had selected, but unbeknownst to them this image was inconspicuously swapped for the image they had *not* selected. Most participants (74%) provided eloquent justification for a choice they had not actually made (Johansson et al., 2005). With some methodological modifications, the choice blindness concept has been replicated across a variety of domains, ranging from taste preferences (Hall et al., 2010) to eye-witness testimony (Sagana et al., 2016) and

financial decisions (McLaughlin & Somerville, 2013). Research has also provided evidence for choice blindness when it comes to moral issues such as political views and affiliations (Hall et al., 2013; Hall et al., 2012), and choice blindness manipulations have been found to produce a lasting change in political attitudes (Strandberg et al., 2018), an area commonly considered hard to influence (Druckman, 2004). Relying on a modified choice blindness paradigm, Strandberg et al. (2020) manipulated extreme political views towards a more neutral, open-minded position, both in face-to-face and online studies.

The essence of the choice blindness paradigm is to inconspicuously swap or manipulate participants' choices or judgements. The fact that participants accept and justify such manipulated outcomes reveals the extent to which self-consistency motives guide behaviour. Earlier research highlighted how people will justify their judgements and choices even when they could not have been aware of the exact reason that triggered their behaviour (Gazzaniga, 2000). For example, Ariely et al. (2003, 2006) showed that even though completely random factors can influence initial value judgement, individuals tend to infer value (utility) thereafter and maintain their initial choice over time revealing what looks like a stable behavioural pattern (self-herding, Ariely & Norton, 2008). Adapting behaviour to exhibit consistency might also stem from individuals' goals of constructing their lives in a sense-making manner (Chater & Loewenstein, 2016). These findings suggest that individuals strive to be consistent with their revealed attitudes and adjust their subsequent behaviour accordingly. The choice blindness paradigm tests whether people strive to be consistent with their (assumed) past selves. Contrasting enhancing and diminishing choice blindness manipulations, however, provides a novel way to explore the interaction between self-consistency and self-enhancing motives.

The Present Study

In the present study, we extended the choice blindness paradigm to the domain of the self, modifying it to implement manipulations that either enhanced or diminished the self-view. Due to the robustness of the choice blindness phenomenon and the importance of self-consistency motivation, we expected that most participants would be

vulnerable to choice blindness manipulations of their self-rankings. However, due to the self-enhancement motivation, we predicted that the basic choice blindness effect would be modulated by whether the manipulation is enhancing or diminishing, i.e., we hypothesised that individuals would be more susceptible to manipulations elevating their self-image.

Method

Overview

Participants compared themselves to others with respect to eight personality characteristics reflecting core dimensions of social perception (desirability and morality). They were then asked to reflect on four of their eight responses. Two of those four responses were manipulated so as to yield either enhanced or diminished (apparent) self-rankings. After answering a set of follow-up questions, participants were given the opportunity to revise their (ostensibly) original rankings for these four personality traits. Finally, they answered questions assessing their self-esteem and general self-view.

Participants

551 participants were recruited based on financial considerations and tested online via *Prolific* (www.prolific.co). All participants were pre-screened to hold UK nationality, be based in the UK, have English as a first language and have a *Prolific* approval rate of at least 90%. Our hypotheses, data collection and analysis plan were pre-registered (https://aspredicted.org/W5Z_9CS). Two exclusion criteria were specified in our pre-registration document: failing either or both of two attention check questions asking participants to move the slider bar to a specified point on the slider scale ($N = 10$); and completing the study in less than three minutes ($N = 0$). We also excluded six participants due to technical issues (the software failed to select a personality trait for the revision phase). The final sample comprised 535 participants (264 females, 264 males, four participants who selected “other” for the gender item and three who preferred not to specify their gender; Mean age = 40.4 years, $SD = 13.0$). Participants

were paid a flat participation fee (£1.00, the equivalent of £7.50 per hour). De-identified data and analysis scripts and study materials are available on the Open Science Framework: https://osf.io/9ke58/?view_only=f2bca85ae0fa43c3add0e9df47a382ff. The study was approved by the Royal Holloway, University of London Ethics Committee.

Materials and Procedure

In the first stage of the experiment, participants were asked to rank themselves relative to “a large number of other participants” on each of eight personality characteristics. These characteristics were either all *desirable* (honest, kind, trustworthy, considerate, intelligent, competent, hard-working, self-disciplined) or all *undesirable* (dishonest, unkind, manipulative, self-centred, irrational, incompetent, lazy, forgetful). As previous research has demonstrated the central role of morality in shaping individuals’ self-concept and their perceptions of self-continuity (Heiphetz et al., 2016; Molouki & Bartels, 2017; Stanley et al., 2019; Strohminger & Nichols, 2014; Strohminger, 2018), we chose four *moral* and four *non-moral traits* (desirable or undesirable). To ensure we could investigate whether *morality* had any effect on the acceptance of the choice blindness manipulations above and beyond the effect of desirability, we selected moral and non-moral personality traits with comparable average desirability ratings (Tappin & McKay, 2017; Ziano et al., 2021).

The personality traits were presented separately in randomized order. Participants indicated their self-rankings on a slider, with endpoints labelled “less than everyone else” and “more than everyone else” (the slider values ranged from 0 to 100, though no numeric values were displayed to participants).

Next, participants were encouraged to reflect on four of these eight self-rankings for a minimum of 20 seconds each (participants could not proceed to the next question earlier than that). Specifically, for each of the four selected traits, they were shown the slider bar with the slider ostensibly as they had positioned it for that trait. Two of those four responses had in fact been manipulated using a modified choice blindness

paradigm (Johansson et al., 2005) to yield either enhanced or diminished (apparent) self-rankings. Here, at the manipulation stage, *morality* was varied between-subjects, with either two *moral* or two *non-moral* traits being manipulated. As a reflection prompt, participants were asked to “think about instances where you have demonstrated (or not demonstrated) this trait and also about when others have demonstrated (or not demonstrated it). Have others ever commented about you being (or not being) _____?” (please see Supplementary material A for details).

The two manipulated and two non-manipulated traits were chosen as follows. First, the computer randomly chose either moral or non-moral traits for manipulation. Second, the computer selected two of the four traits from the chosen morality category. This selection was not purely random – a strict preference was given to traits which were ranked between 25 and 75. If two traits were found within this interval, the computer randomly decided to either increase or decrease both rankings by 20. If fewer traits were found within the interval, the computer also chose traits from outside the interval and then decreased (increased) their ranking by 20 if their rankings were above 75 (below 25). The two non-manipulated traits were chosen from the other moral category at random.

After answering the follow-up questions, respondents were again shown their (apparent) self-rankings and were given the opportunity to revise them by moving the slider from its (apparent) original position.

Next, participants answered questions assessing their self-esteem (using the single item self-esteem measure; Robins et al., 2001) and general self-view (using bespoke questions about how good, happy, and capable participants considered themselves to be; participants again responded to these on a 0-100 slider, with endpoints labelled “not very true of me” and “very true of me”; numeric values were not displayed to participants). Finally, to ascertain whether participants detected or suspected the choice-blindness manipulation, we concluded with: “This is a pilot study, please let us know if you have noticed anything unusual or if you have experienced any

technical issues". Participants had the opportunity to respond to the prompt with free text.

Motivation of the Manipulation Procedure

Individuals tend to think of themselves as better than the average person (the "better-than-average" effect; Alicke & Govorun, 2005; Brown, 1986) and this trend seems to be magnified for moral traits (Brown, 2012; Tappin & McKay, 2017). Hence, we expected that our participants would rank themselves relatively high on desirable traits and relatively low on undesirable traits, especially for moral traits. However, given that we manipulate rankings by 20 points, the consequence is that most rankings could probably only be manipulated in a "diminishing" way, i.e., in the direction which renders the self-rankings more negative. As a result, we reasoned that we would have more diminishing choice blindness manipulations than "enhancing" ones. It also meant that participants who receive a diminishing manipulation would have submitted a more positive original self-ranking on average. Both these issues could potentially bias a comparison of diminishing and enhancing manipulations.

In anticipation of this, we designed our mechanism to prefer traits ranked in the "middle zone", i.e., in the interval from 25 to 75 (slightly expanding the 20 and 80 limits to exclude values which if manipulated would result in sliders effectively on the end points), and to randomize 50-50 between diminishing and enhancing manipulations of traits in this interval. This procedure means that although, overall, we would likely have more diminishing manipulations than enhancing ones, within the middle zone we would have an equal number of diminishing and enhancing manipulations, allowing an unbiased comparison of the two conditions.⁴

Therefore, while we expect participants to be susceptible to the choice blindness manipulation in the total sample (H1), we pre-registered our hypothesis that enhancing manipulations will have stronger effect on personality judgements than diminishing

⁴ It turned out that these concerns were unfounded and the results for the full sample and the middle zone only were very similar.

manipulations in the middle zone only (H2). As secondary hypotheses, we expect enhancing manipulations to elevate measures of self-esteem and general self-view (indications on how good, happy, and capable one is) as well as that revisions in the two non-manipulated traits (if any) will be in an enhancing direction. We also explore whether enhancing and diminishing manipulations have a differential effect depending on the desirability and morality of the traits as well as test for potential effects on some additional measures of the general self-view (indications on how good, happy, and capable one is).

Results

Design and Analysis

We analysed the data from the manipulation stage of the experiment, i.e., we compared the revised and original self-rankings for the four traits (two manipulated and two non-manipulated) presented to participants for reflection, follow-up questions and second self-rankings⁵. In the manipulation stage the experiment employed a 2 (*morality*: moral, non-moral, between-participants) x 2 (*manipulation*: manipulated, non-manipulated, within-participants) x 2 (*manipulation type*: diminishing, enhancing, between-participants) mixed design.

For our analysis, we reverse-coded the self-rankings for undesirable traits (i.e., subtracting them from 100), so that for all rankings, a higher numerical value means a more positive self-judgment. In addition, we defined the following dummy variables: a dummy for *enhancing* manipulations, which is 1 if the participant's manipulated traits were moved in the direction of a more positive self-ranking and 0 otherwise; a dummy for *diminishing* manipulations, which is 1 if the participant's manipulated traits were

⁵ We have not specified hypotheses based on the data from the first stage of the experiment (the indicated self-rankings on eight personality traits), however we are enclosing the analysis in Supplementary material B for the sake of completeness.

moved in the direction of a more negative self-ranking and 0 otherwise; and a dummy for *moral* traits, which is 1 if the trait belongs to the *moral* category and 0 otherwise.

Main Hypotheses

Result 1: The Manipulation Is Effective

Our first main hypothesis was that most participants would be vulnerable to manipulations of their self-rankings, i.e., they would not notice or correct for the fact that some of their rankings had been manipulated. We investigate this in two ways.

First, when we asked whether they had encountered any technical issues with the study, only 13 participants (2.43% of the sample, 8 in the diminishing and 5 in the enhancing manipulation condition) reported that the self-rankings they were shown in the revision phase did not match their original self-rankings. Whereas in other choice blindness studies participants might have been reluctant to point out apparent mistakes, e.g., because they did not want to embarrass the experimenters, in our study pointing out such mistakes would have been a favour to the experimenter and was nearly costless. Hence, the low number of such reports suggest the vast majority of our participants were blind to the choice blindness manipulation. We present our subsequent analyses including these 13 participants who detected the manipulation, however excluding them yields virtually identical results (please see Supplementary material C).

Second, as per our pre-registration, we tested whether the revised self-rankings of manipulated traits deviated more from their original values than did those for non-manipulated traits. Specifically, we fitted linear mixed models on the difference between revised and original self-rankings (revised minus original; Table 1, Models 1-6). We ran separate regressions for the full sample (Models 1-3) and for observations in the middle zone, i.e., including only participants whose manipulated values originally lay in the 25-75 interval (Models 4-6). Enhancing manipulations, diminishing manipulations and morality were modelled as fixed effects while participants were modelled as

random effects.⁶ We also ran the analysis *including* interaction terms between the fixed effects (Table 1, Models 3 and 6).

The results of Models 1-6 show that the manipulation was successful. For example, the intercept in model 1 shows that the non-manipulated traits were on average revised slightly upwards (+1.00, $p < .001$). In contrast, revised rankings for traits that were manipulated in a diminishing way were on average -11.38 lower ($p < .001$), whereas revised rankings for traits manipulated in an enhancing way were on average 12.53 higher ($p < .001$) than the original rankings. Adding morality and its interactions to the models changes these coefficients only marginally and neither morality nor its interactions have a statistically significant effect. Likewise, the results for full sample and middle zone are qualitatively the same.

⁶ We pre-registered that we would include desirability and the type of manipulation as fixed effects, but we replaced these with the enhancing and diminishing manipulation dummies for ease of interpretation. Otherwise, we would have to work with interaction effects all the time, which are much less straightforward to interpret. We include the (equivalent) pre-registered analysis in Supplementary material D, which yielded identical results.

Table 1. Estimated fixed effects for the difference between revised and original self-rankings.

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
	total	total	total	middle	middle	middle
	sample	sample	sample	zone	zone	zone
Intercept (non-manipulated traits)	1.00*** (0.30)	1.17** (0.39)	1.01* (0.44)	1.44*** (0.38)	1.14* (0.48)	1.17* (0.53)
Enhancing manipulation	12.53*** (0.55)	12.52*** (0.55)	12.42*** (0.78)	12.33*** (0.62)	12.30*** (0.62)	12.66*** (0.88)
Diminishing manipulation	-11.38*** (0.43)	-11.38*** (0.43)	-10.83*** (0.64)	-9.86*** (0.61)	-9.84*** (0.61)	-10.21*** (0.81)
Morality (moral traits)		-0.31 (0.47)	-0.02 (0.60)		0.63 (0.60)	0.58 (0.77)
Interaction Enhancing*Morality (moral traits)			0.21 (1.10)			-0.69 (1.24)
Interaction Diminishing*Morality (moral traits)			-1.00 (0.87)			0.86 (1.23)

Baseline levels: non-manipulated, non-moral; Number of observations: total sample 2140, grouped by participants, N=535; middle zone: 1440; grouped by participants, N=360; Significance codes: '***' 0.001; '**' 0.01; '*' 0.05;

Figure 1 plots the histograms of the ranking differences (revised minus original) for non-manipulated and manipulated (enhanced or diminished) traits for the total sample (Panel A) and the middle zone (Panel B). The difference between revised and original self-rankings for non-manipulated traits is equally distributed around a high peak at 0,

i.e., most participants did not revise their original judgements much (± 5) when the trait was not manipulated. For diminished and enhanced traits, on the other hand, the distributions clearly shifted to the left and right with high peaks around -20 and +20, respectively, i.e., most participants largely accepted the manipulated values and adjusted them only slightly (± 5).

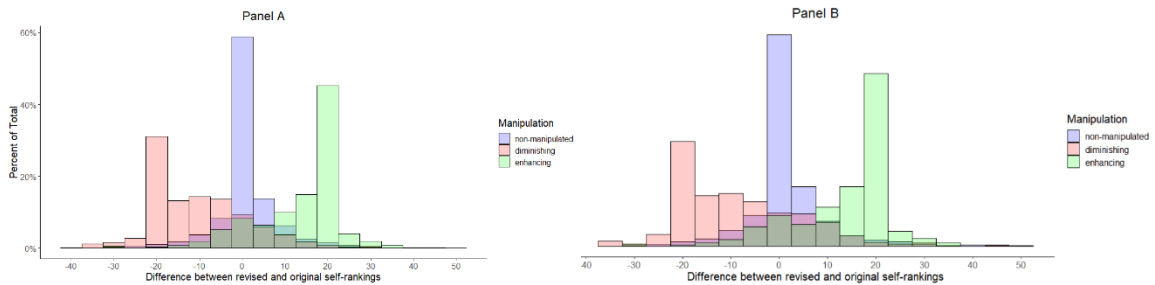


Figure 1. Histograms of the difference between revised and original self-rankings for non-manipulated and manipulated (enhanced or diminished) traits, total sample (Panel A) and middle zone (Panel B).

Result 2: Enhancing Manipulations Are Stronger Than Diminishing Manipulations.

In testing our second main hypothesis, that enhancing manipulations would be more readily accepted than diminishing manipulations, we could not directly compare the coefficients of the diminishing and enhancing conditions because they have different signs. To make the effects of the two conditions comparable, we flipped the sign of the difference between original and revised rankings in the diminishing condition. We call this *magnitudinal difference* to distinguish it from the difference used to test the first hypothesis⁷. As a result, a *magnitudinal difference* of +20 means that the participant fully accepted the manipulated ranking, be it diminished or enhanced. A

⁷ Note that flipping the sign is similar, but not exactly the same as taking the mathematical absolute value. To illustrate, consider a participant who received a diminishing manipulation, but whose revised ranking was more positive than their original ranking. Their *magnitudinal difference* would be negative since the manipulation had the opposite effect on this participant than was the norm. As the histogram in Figure 1 shows, such participants were small in number, but did exist.

magnitudinal difference of +10 means that the participant's revised ranking lay exactly between the original and the manipulated value.

We fitted linear mixed models on the *magnitudinal difference* with *enhancing* manipulations, *morality* and their interaction as fixed effects and participants as a random effect (Table 2, Models 7-12). As the models are fitted on the dataset for manipulated traits only (whether in enhancing or diminishing direction), the baseline is diminishing traits and the intercept reflects the coefficient for *diminishing* manipulations. Additionally, Figure 2 depicts the histogram as well as a bar plot of the average of the *magnitudinal difference* in diminishing and enhancing conditions for the middle zone (please see Supplementary material E for the histogram and bar chart for the total sample). All six models and the bar plot show that the effect of the *enhancing* manipulation was significantly stronger, i.e., it pulled participants further away from their original rankings than the *diminishing* manipulation did. Or, to put it differently, participants in the diminishing condition revised their rankings further back towards their original rankings. The histogram further reveals that in the enhancing condition a much larger proportion of the participants accepted the manipulation and revised it only marginally (± 5).

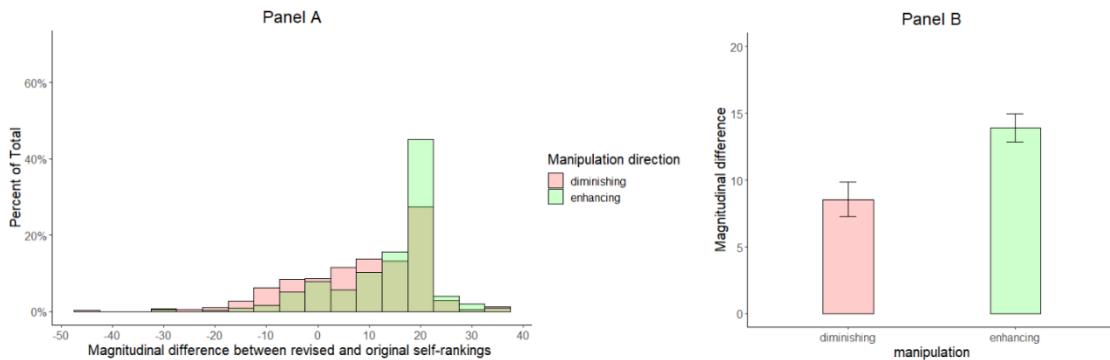


Figure 2. Histogram for the *magnitudinal difference* of enhancing and diminishing manipulations, middle zone (Panel A) and bar chart for the *magnitudinal difference* for enhancing and diminishing manipulations, middle zone (Panel B).

Table 2. Estimated fixed effects for enhancing, morality, and their interaction with respect to the *magnitudinal difference*.

	Model 7	Model 8	Model 9	Model 10	Model 11	Model 12
	total	total	total	middle	middle	middle
	sample	sample	sample	zone	zone	zone
Intercept	10.51***	10.05***	9.91***	8.54***	8.87***	9.14***
(diminishing manipulation)	(0.48)	(0.64)	(0.71)	(0.68)	(0.80)	(0.90)
Enhancing manipulation	3.26***	3.30***	3.68***	5.36***	5.43***	4.82***
	(0.81)	(0.81)	(1.16)	(0.98)	(0.98)	(1.36)
Morality (moral traits)		0.83	1.09		-0.75	-1.40
		(0.77)	(0.96)		(0.98)	(1.38)
Interaction			-0.72			1.30
Enhancing*Morality (moral traits)			(1.62)			(1.97)

*Baseline levels: diminishing, non-moral; Number of observations: total sample 1070; grouped by participants, N=535; middle zone: 720; grouped by participants, N=360; Significance codes: '***' 0.001; '**' 0.01; '*' 0.05;*

Secondary and Exploratory Hypotheses

Result 3: No effect on self-esteem and general self-image

To check whether manipulating traits in the middle zone in a diminishing or enhancing direction would affect subsequent self-judgements of self-esteem, goodness, happiness and capability, we ran additional regressions. However, none of them showed any significant effect of the type of manipulation on any subsequent self-assessment (Table G1, Supplementary material G).

Result 4: Revised self-rankings of non-manipulated traits tend to be more positive than the original self-rankings.

On average, revised rankings of non-manipulated traits were about 1 point higher than original rankings, a small but statistically significant difference (see Table G2, Figure G1, Supplementary material G). This is consistent with literature showing that second estimates tend to be more optimistic (Van der Leer & McKay, 2017). There was no difference between the diminishing and enhancing condition (please see Figure G2, Supplementary material G).

Result 5: Effects are not more pronounced for moral traits.

Contrary to our expectation, none of our analyses show any statistically significant effects of morality (cf. Models 1-12; Model G2).

Result 6: Enhancing (diminishing) manipulations are not stronger (weaker) for undesirable than desirable traits.

We found no evidence that the combination of desirability and manipulation type (enhancing or diminishing) mattered (please see Table G3, Supplementary material G for the analysis).

Discussion

The essential choice blindness finding has been replicated across various decision-making domains, ranging from eye-witness testimony (Sagana et al., 2016) to moral issues and political views (Hall et al., 2013; Hall et al., 2012; Stranberg et al., 2020). In the present study we extended the choice blindness paradigm to the domain of the self, exploring whether participants are “blind” to manipulations of their personality judgements. Our data evidenced a clear choice blindness effect on personality judgements: the majority of manipulated self-rankings were not revised

back to the original personality judgements. In addition, the analyses revealed a “self-serving” choice blindness effect on personality judgements, with enhancing manipulations being accepted to a higher degree than diminishing manipulations.

These findings are notable as personality judgements are commonly considered to be stable in adulthood (e.g., Costa & McCrae, 1992; McCrae et al., 2000). Even if some degree of malleability of personality traits throughout adulthood has been documented (e.g., Bleidorn et al., 2022), changes in personality judgements seem to be achieved only after weeks of nonclinical psychological interventions (e.g., Stieger et al., 2020). Our study, however, showed that self-rankings are responsive to a cognitive influence, which points towards somewhat different psychological mechanisms shaping the construction and flexibility of personality judgements than those currently theorised (Roberts & Yoon, 2022).

Our results support research showing that individuals can adjust their attitudes quickly to achieve consistency with their perceived revealed attitudes (Johansson et al., 2012). However, the data showed that personality judgements are not only susceptible to choice blindness manipulations, but that enhancing manipulations are accepted to a higher degree. If the sole motivation for the observed behavioural pattern was consistency with perceived previous behaviour, we should not have observed a difference in the choice blindness manipulation’s acceptance levels depending on the direction of the manipulation (enhancing or diminishing). Hence, besides self-consistency, self-enhancement motivation shapes personality judgements. The observed self-serving choice blindness of personality judgements is also in line with recent work showing that personality judgements are prone to anchoring when the anchors enhance the self-evaluation (Ambrus et al., under review).

The self-serving susceptibility of personality judgements to choice blindness manipulations also echoes findings on the range of psychological mechanisms employed in maintaining a positive and stable self-view, such as for instance the self-serving update of self-beliefs in response to personal feedback (Möbius et al. 2022). Moreover, individuals fail to recall previous unethical deeds even when incentivised for accuracy of

memory (Carlson et al., 2020). If previous unethical deeds are recalled, individuals dissociate from them and claim they perceive a self-change in themselves while if previous moral deeds are recollected, participants feel an association with their previous behaviours and report a perception of self-continuity (Stanley et al., 2019). Hence, individuals seem to exhibit flexibility about their perceived self-image stability as long as this aligns with their expectations and allows them to maintain a consistent, stable, and positive self-view.

Contrary to past findings on magnified cognitive effects in the moral domain, however (e.g., Meyers et al., 2019), the self-serving effect we found was not more pronounced for moral traits. As discussed above, to ensure a valid experimental design that could investigate the effect of morality above and beyond that of desirability, we chose moral traits that have a comparable average desirability ranking with the selected non-moral traits. The lack of a magnified effect for moral traits suggests that instead of morality per se, desirability determines the degree to which participants are susceptible to choice blindness manipulations of their personality characteristics. As participants rate moral traits as more desirable than non-moral traits (Tappin & McKay, 2017), it might be challenging to disentangle the effect of morality from that of desirability.

Summary and conclusions

In summary, our study extends the choice blindness paradigm to the domain of the self and reveals that most individuals are blind to subtle manipulations of their personality judgements. Humans are especially vulnerable to such “sleights of mind” (Macknik & Martinez-Conde, 2010; McKay et al., 2005), however, when the manipulations are flattering.

References

- Allemand, M., & Flückiger, C. (2022). Personality change through digital-coaching interventions. *Current Directions in Psychological Science*, 31(1), 41-48. <https://doi.org/10.1177/0963721421106778>
- Alicke, M. D. (1985). Global self-evaluation as determined by the desirability and controllability of trait adjectives. *Journal of Personality and Social Psychology*, 49(6), 1621–1630. <https://doi.org/10.1037/0022-3514.49.6.1621>
- Alicke, M. D., & Govorun, O. (2005). The better-than-average effect. *The Self in Social Judgment*, 1, 85-106. https://doi.org/10.1007/978-3-319-24612-3_300293
- Alicke, M., & Sedikides, C. (2009). Self-enhancement and self-protection: What they are and what they do. *European Review of Social Psychology* 20, 1-48. Self-Enhancement and Self-Protection: What They Are and What They Do. <https://doi.org/10.1080/10463280802613866>
- Alicke, M. D., Zell, E., & Guenther, C. L. (2013). Social self-analysis: Constructing, protecting, and enhancing the self. *Advances in Experimental Social Psychology*, 48, 173–234. <https://doi.org/10.1016/B978-0-12-407188-9.00004-1>
- Alicke, M. D., Vredenburg, D. S., Hiatt, M., & Govorun, O. (2001). The “better than myself effect”. *Motivation and Emotion*, 25(1), 7-22. <https://doi.org/10.1023/A:1010655705069>
- Ambrus, E. Z., Hartig, B., & McKay, R. T. (2023). Self-serving anchoring of personality judgements. *Under review*
- Ariely, D., & Norton, M. I. (2008). How actions create—not just reveal—preferences. *Trends in Cognitive Sciences*, 12(1), 13-16. <https://doi.org/10.1016/j.tics.2007.10.008>.
- Ariely, D., Loewenstein, G., & Prelec, D. (2003). “Coherent arbitrariness”: Stable demand curves without stable preferences. *The Quarterly Journal of Economics*, 118(1), 73-106. <https://doi.org/10.1162/00335530360535153>.

- Ariely, D., Loewenstein, G., & Prelec, D. (2006). Tom Sawyer and the construction of value. *Journal of Economic Behavior & Organization*, 60(1), 1-10.
<https://doi.org/10.1016/j.jebo.2004.10.003>
- Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. *arXiv preprint arXiv:1506.04967 [stat.ME]*.
<https://doi.org/10.48550/arXiv.1506.04967>
- Beck, E. D., & Jackson, J. J. (2022). A mega-analysis of personality prediction: Robustness and boundary conditions. *Journal of Personality and Social Psychology*, 122(3), 523– 553. <https://doi.org/10.1037/pspp0000386>
- Bleidorn, W., Hopwood, C. J., Back, M. D., Denissen, J. J., Hennecke, M., Hill, P. L., ... & Zimmermann, J. (2021). Personality trait stability and change. *Personality Science*, 2, 1-20. <https://doi.org/10.5964/ps.6009>
- Bleidorn, W., Schwaba, T., Zheng, A., Hopwood, C. J., Sosa, S. S., Roberts, B. W., & Briley, D. A. (2022). Personality stability and change: A meta-analysis of longitudinal studies. *Psychological bulletin*, 148(7-8), 588
<https://doi.org/10.1037/bul0000365>
- Brown, J. D. (1986). Evaluations of self and others: Self-enhancement biases in social judgement. *Social cognition*, 4(4), 353-376.
<https://doi.org/10.1521/SOCO.1986.4.4.353>
- Brown, J. D. (2012). Understanding the better than average effect: Motives (still) matter. *Personality and Social Psychology Bulletin*, 38(2), 209-219.
<https://doi.org/10.1177/0146167211432763>
- Burton, J. W., Harris, A. J., Shah, P., & Hahn, U. (2022). Optimism where there is none: asymmetric belief updating observed with valence-neutral life events. *Cognition*, 218, 104939.
<https://doi.org/10.1016/j.cognition.2021.104939>

- Carlson, R. W., Maréchal, M. A., Oud, B., Fehr, E., & Crockett, M. J. (2020). Motivated misremembering of selfish decisions. *Nature Communications*, *11*(1), 1-11. <https://doi.org/10.1038/s41467-020-15602-4>
- Chater, N., & Loewenstein, G. (2016). The under-appreciated drive for sense-making. *Journal of Economic Behavior & Organization*, *126*, 137-154. <https://doi.org/10.1016/j.jebo.2015.10.016>
- Costa Jr, P. T., & McCrae, R. R. (1992). Four ways five factors are basic. *Personality and Individual Differences*, *13*(6), 653-665. [https://doi.org/10.1016/0191-8869\(92\)90236-l](https://doi.org/10.1016/0191-8869(92)90236-l)
- Eil, D., & Rao, J. M. (2011). The good news-bad news effect: asymmetric processing of objective information about yourself. *American Economic Journal: Microeconomics*, *3*(2), 114-38. <https://doi.org/10.1257/mic.3.2.114>
- Epstein, S. (1973). The self-concept revisited: Or a theory of a theory. *American Psychologist*, *28*(5), 404–416. <https://doi.org/10.1037/h0034679>
- Druckman, J. N. (2004). Political preference formation: Competition, deliberation, and the (ir) relevance of framing effects. *American Political Science Review*, *98*(4), 671-686. <https://doi.org/10.1017/S0003055404041413>
- Dungan, J., & Young, L. (2015). Understanding the adaptive functions of morality from a cognitive psychological perspective. *Emerging Trends in the Social and Behavioral Sciences*. John Wiley & Sons (Wiley Online Library). <https://doi.org/10.1002/9781118900772.etrds0376>
- Dungan, J. A., & Young, L. (2019). Asking ‘why?’ enhances theory of mind when evaluating harm but not purity violations. *Social cognitive and affective neuroscience*, *14*(7), 699-708. <https://doi.org/10.1093/scan/nsz048>
- Gaertner, L., Sedikides, C., & Cai, H. (2012). Wanting to be great and better but not average: On the pancultural desire for self-enhancing and self-improving

- feedback. *Journal of CrossCultural Psychology*, 43, 521–526.
<https://doi.org/10.1177/0022022112438399>
- Garrett, N., & Sharot, T. (2017). Optimistic update bias holds firm: Three tests of robustness following Shah et al. *Consciousness and Cognition*, 50, 12–22.
<https://doi.org/10.1016/j.concog.2016.10.013>
- Gazzaniga, M. S. (2000). Cerebral specialization and interhemispheric communication: Does the corpus callosum enable the human condition?. *Brain*, 123(7), 1293–1326. <https://doi.org/10.1093/brain/123.7.1293>
- Greenwald, A. G. (1980). The totalitarian ego: Fabrication and revision of personal history. *American psychologist*, 35(7), 603. <https://doi.org/10.1037/0003-066X.35.7.603>
- Hall, L., Johansson, P., Tärning, B., Sikström, S., & Deutgen, T. (2010). Magic at the marketplace : Choice blindness for the taste of jam and the smell of tea. *Cognition*, 117(1), 54–61. <https://doi.org/10.1016/j.cognition.2010.06.010>
- Hall, L., Strandberg, T., Pärnamets, P., Lind, A., Tärning, B., & Johansson, P. (2013). How the polls can be both spot on and dead wrong: Using choice blindness to shift political attitudes and voter intentions. *PloS one*, 8(4), e60554.
<https://doi.org/10.1371/journal.pone.0060554>
- Hall, L., Johansson, P., & Strandberg, T. (2012). Lifting the veil of morality: Choice blindness and attitude reversals on a self-transforming survey. *PloS one*, 7(9), e45457. <https://doi.org/10.1371/journal.pone.0045457>
- Hirsch, J. B., Mar, R. A., & Peterson, J. B. (2013). Personal narratives as the highest level of cognitive integration. *Behavioral and Brain Sciences*, 36(3), 216–217.
<https://doi.org/10.1017/S0140525X12002269>
- Heiphetz, L., Strohminger, N., & Young, L. L. (2017). The role of moral beliefs, memories, and preferences in representations of identity. *Cognitive science*, 41(3), 744–767.
<https://doi.org/10.1111/cogs.12354>

- Hixon, J. G., & Swann, W. B. (1993). When does introspection bear fruit? Self-reflection, self-insight, and interpersonal choices. *Journal of Personality and Social Psychology*, 64(1), 35–43. <https://doi.org/10.1037/0022-3514.64.1.35>
- Joel, S., Spielmann, S. S., & MacDonald, G. (2017). Motivated use of numerical anchors for judgement relevant to the self. *Personality and Social Psychology Bulletin*, 43(7), 972-985. <https://doi.org/10.1177/014616721770261>
- Johansson, P., Hall, L., Sikstrom, S., & Olsson, A. (2005). Failure to detect mismatches between intention and outcome in a simple decision task. *Science*, 310(5745), 116-119. <https://doi.org/10.1126/science.1111709> PMID: 16210542
- Johansson, P., Hall, L., & Chater, N. (2012). Preference change through choice. In *Neuroscience of preference and choice* (pp. 121-141). Academic Press. <https://doi.org/10.1016/B978-0-12-381431-9.00006-1>
- Leary, M. R. (2007). Motivational and emotional aspects of the self. *Annual Review of Psychology*, 58, 317–344. <https://doi.org/10.1146/annurev.psych.58.110405.085658>
- Lefebvre, G., Lebreton, M., Meyniel, F., Bourgeois-Gironde, S., & Palminteri, S. (2017). Behavioural and neural characterization of optimistic reinforcement learning. *Nature Human Behaviour*, 1(4), 1-9. <https://doi.org/10.1038/s41562-017-0067>
- McCrae, R. R., Costa, P. T. J., Jr., Ostendorf, F., Angleitner, A., Hrebícková, M., Avia, M. D., Sanz, J., Sánchez-Bernardos, M. L., Kusdil, M. E., Woodfield, R., Saunders, P. R., & Smith, P. B. (2000). Nature over nurture: Temperament, personality, and life span development. *Journal of Personality and Social Psychology*, 78(1), 173–186. <https://doi.org/10.1037/0022-3514.78.1.173>
- Malle, B. F. (2006). The actor-observer asymmetry in attribution: A (surprising) meta-analysis. *Psychological Bulletin*, 132(6), 895-919. <https://doi.org/10.1037/0033-2909.132.6.895>

- Macknik, S., Martinez-Conde, S., & Blakeslee, S. (2010). *Sleights of mind: What the neuroscience of magic reveals about our everyday deceptions*. Henry Holt and Company.
- McKay, R., Langdon, R., & Coltheart, M. (2005). "Sleights of mind": Delusions, defences, and self-deception. *Cognitive Neuropsychiatry*, 10(4), 305–326. <https://doi.org/10.1080/13546800444000074>
- McLaughlin, O., & Somerville, J. (2013). Choice blindness in financial decision making. *Judgment and Decision Making*, 8(5), 577. <https://EconPapers.repec.org/RePEc:jdm:journl:v:8:y:2013:i:5:p:577-588>
- Meyers, E. A., Białek, M., Fugelsang, J. A., Koehler, D. J., & Friedman, O. (2019). Wronging past rights: The sunk cost bias distorts moral judgment. *Judgment and Decision Making*, 14(6), 721-727. <https://EconPapers.repec.org/RePEc:jdm:journl:v:15:y:2020:i:6:p:909-925>
- Molouki, S., & Bartels, D. M. (2017). Personal change and the continuity of the self. *Cognitive Psychology*, 93, 1-17. <https://doi.org/10.1016/j.cogpsych.2016.11.006>
- Möbius, M. M., Niederle, M., Niehaus, P., & Rosenblat, T. S. (2022). Managing self-confidence: Theory and experimental evidence. *Management Science* 68(11). <https://doi.org/10.1287/mnsc.2021.4294>
- Olaru, G., Stieger, M., Rügger, D., Kowatsch, T., Flückiger, C., Roberts, B. W., & Allemand, M. (2022). Personality change through a digital-coaching intervention: Using measurement invariance testing to distinguish between trait domain, facet, and nuance change. *European Journal of Personality*. <https://doi.org/10.1177/089020702211450>
- Pronin, E., & Hazel, L. (2023). Humans' Bias Blind Spot and Its Societal Significance. *Current Directions in Psychological Science*, 09637214231178745. <https://doi.org/10.1177/096372142311787>

- Pronin, E., Lin, D. Y., & Ross, L. (2002). The bias blind spot: Perceptions of bias in self versus others. *Personality and Social Psychology Bulletin*, 28(3), 369–381. <https://doi.org/10.1177/0146167202286008>
- R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>
- Ritchie, T. D., Sedikides, C., & Skowronski, J. J. (2017). Does a person selectively recall the good or the bad from their personal past? It depends on the recall target and the person's favourability of self-views. *Memory*, 25(8), 934-944. <https://doi.org/10.1080/09658211.2016.1233984>
- Roberts, B. W. (2009). Back to the future: Personality and assessment and personality development. *Journal of research in personality*, 43(2), 137-145. <https://doi.org/10.1016/j.jrp.2008.12.015>
- Roberts, B. W., Kuncel, N. R., Shiner, R., Caspi, A., & Goldberg, L. R. (2007). The power of personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. *Perspectives on Psychological Science*, 2(4), 313-345. <https://www.jstor.org/stable/40212212>
- Roberts, B. W., Luo, J., Briley, D. A., Chow, P. I., Su, R., & Hill, P. L. (2017). A systematic review of personality trait change through intervention. *Psychological Bulletin*, 143(2), 117-141. <https://doi.org/10.1037/bul0000088>
- Roberts, B. W., Walton, K. E., & Viechtbauer, W. (2006). Patterns of meanlevel change in personality traits across the life course: A meta-analysis of longitudinal studies. *Psychological Bulletin*, 132(1), 1–25. <https://doi.org/10.1037/0033-2909.132.1>
- Roberts, B. W., & Yoon, H. J. (2022). Personality psychology. *Annual Review of Psychology*, 73(1), 489–516. <https://doi.org/10.1146/annurevpsych-020821-114927>
- Robins, R. W., Hendin, H. M., & Trzesniewski, K. H. (2001). Measuring global self-esteem: Construct validation of a single-item measure and the Rosenberg Self-Esteem

- Scale. *Personality and social psychology bulletin*, 27(2), 151-161.
<https://doi.org/10.1177/0146167201272002>
- Sagana, A., Sauerland, M., & Merckelbach, H. (2016). The effect of choice reversals on blindness for identification decisions. *Psychology, Crime & Law*, 22(4), 303-314.
<https://doi.org/10.1080/1068316X.2015.1085984>
- Sanitioso, R., Kunda, Z., & Fong, G. T. (1990). Motivated recruitment of autobiographical memories. *Journal of Personality and Social psychology*, 59(2), 229-241.
<https://doi.org/10.1037/0022-3514.59.2.229>
- Sedikides, C., Gaertner, L., & Toguchi, Y. (2003). Pancultural self-enhancement. *Journal of Personality and Social Psychology*, 84(1), 60–79.
<https://doi.org/10.1037/0022-3514.84.1.60>
- Sedikides, C., Gaertner, L., & Vevea, J. L. (2005). Pancultural self-enhancement reloaded: A meta-analytic reply to Heine (2005). *Journal of Personality and Social Psychology*, 89(4), 539–551. <https://doi.org/10.1037/0022-3514.89.4.539>
- Sedikides, C., & Green, J. D. (2009). Memory as a self-protective mechanism. *Social and Personality Psychology Compass*, 3(6), 1055–1068.
<https://doi.org/10.1111/j.1751-9004.2009.00220.x>
- Sedikides, C., Green, J. D., Saunders, J., Skowronski, J. J., & Zengel, B. (2016). Mnemic neglect: Selective amnesia of one's faults. *European Review of Social Psychology*, 27, 1–62. <https://doi.org/10.1080/10463283.2016.1183913>
- Sedikides, C., & Strube, M. J. (1997). Self-evaluation: To thine own self be good, to thine own self be sure, to thine own self be true, and to thine own self be better. In *Advances in experimental social psychology* (Vol. 29, pp. 209-269). Academic Press. [https://doi.org/10.1016/S0065-2601\(08\)60018-0](https://doi.org/10.1016/S0065-2601(08)60018-0)
- Shafir, E. (1993). Choosing versus rejecting: Why some options are both better and worse than others. *Memory & cognition*, 21(4), 546-556.
<https://doi.org/10.3758/BF03197186>

- Sharot, T., & Garrett, N. (2016). Forming beliefs: Why valence matters. *Trends in Cognitive Sciences*, 20(1), 25–33. <https://doi.org/10.1016/j.tics.2015.11.002>
- Shah, P., Harris, A. J. L., Bird, G., Catmur, C., & Hahn, U. (2016). A pessimistic view of optimistic belief updating. *Cognitive Psychology*, 90, 71–127. <https://doi.org/10.1016/j.cogpsych.2016.05.004>
- Soto, C. J. (2021). Do links between personality and life outcomes generalize? Testing the robustness of trait–outcome associations across gender, age, ethnicity, and analytic approaches. *Social Psychological and Personality Science*, 12(1), 118–130. <https://doi.org/10.1177/1948550619900572>
- Stanley, M. L., Henne, P., & De Brigard, F. (2019). Remembering moral and immoral actions in constructing the self. *Memory & Cognition*, 47(3), 441–454. <https://doi.org/10.3758/s13421-018-0880-y>
- Stewart, N., & Brown, G. D. A. (2004). Sequence effects in categorizing tones varying in frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30(2), 416–430. <https://doi.org/10.1037/0278-7393.30.2.416>
- Stewart, N., Brown, G. D. A., & Chater, N. (2005). Absolute identification by relative judgment. *Psychological Review*, 112(4), 881–911. <https://doi.org/10.1037/0033-295X.112.4.881>
- Stieger, M., Wepfer, S., Rügger, D., Kowatsch, T., Roberts, B. W., & Allemand, M. (2020). Becoming more conscientious or more open to experience? Effects of a two-week smartphone-based intervention for personality change. *European Journal of Personality*, 34(3), 345–366. <https://doi.org/10.1002/per.2267>
- Strandberg, T., Sivén, D., Hall, L., Johansson, P., & Pärnamets, P. (2018). False beliefs and confabulation can lead to lasting changes in political attitudes. *Journal of Experimental Psychology: General*, 147(9), 1382. <https://doi.org/10.1037/xge0000489>

- Strandberg, T., Olson, J. A., Hall, L., Woods, A., & Johansson, P. (2020). Depolarizing American voters: Democrats and Republicans are equally susceptible to false attitude feedback. *Plos one*, *15*(2), e0226799.
<https://doi.org/10.1371/journal.pone.0226799>
- Strohming, N. (2018). Identity is essentially moral. *Atlas of moral psychology*, 141-148.
<https://doi.org/10.1016/j.cognition.2013.12.005>
- Strohming, N., & Nichols, S. (2014). The essential moral self. *Cognition*, *131*(1), 159–171. <https://doi.org/10.1016/j.cognition.2013.12.005>
- Strohming, N. (2018). Identity is essentially moral. *Atlas of moral psychology*, 141-148.
<https://doi.org/10.1016/j.cognition.2013.12.005>
- Strohming, N., & Nichols, S. (2014). The essential moral self. *Cognition*, *131*(1), 159–171. <https://doi.org/10.1016/j.cognition.2013.12.005>
- Swann WB Jr. 1983. Self-verification: bringing social reality into harmony with the self. In Psychological Perspectives on the Self, ed. J Suls, AG Greenwald, 2:33–66. Hillsdale, NJ: Erlbaum
- Swann, W. B. (1987). Identity negotiation: Where two roads meet. *Journal of Personality and Social Psychology*, *53*(6), 1038–1051. <https://doi.org/10.1037/0022-3514.53.6.1038>
- Swann, W. B., Jr. (1990). To be adored or to be known? The interplay of self-enhancement and self-verification. In E. T. Higgins & R. M. Sorrentino (Eds.), *Handbook of motivation and cognition: Foundations of social behavior*, Vol. 2, pp. 408–448). The Guilford Press.
- Swann, W. B., Jr., & Pelham, B. (2002). Who wants out when the going gets good? Psychological investment and preference for self-verifying college roommates. *Self and Identity*, *1*(3), 219–233. <https://doi.org/10.1080/152988602760124856>

- Szumowska, E., Szwed, P., Wójcik, N., & Kruglanski, A. W. (2023). The interplay of positivity and self-verification strivings: Feedback preference under increased desire for self-enhancement. *Learning and Instruction, 83*, 101715. <https://doi.org/10.1016/j.learninstruc.2022.101715>
- Tappin, B. M., & McKay, R. T. (2017). The illusion of moral superiority. *Social Psychological and Personality Science, 8*(6), 623-631. <https://doi.org/10.1177/1948550616673878>
- Taylor, S. E., & Brown, J. D. (1988). Illusion and well-being: A social psychological perspective on mental health. *Psychological Bulletin, 103*(2), 193–210. <https://doi.org/10.1037/0033-2909.103.2.193>
- Taylor, S. E., Lerner, J. S., Sherman, D. K., Sage, R. M., & McDowell, N. K. (2003). Portrait of the self-enhancer: Well adjusted and well liked or maladjusted and friendless? *Journal of Personality and Social Psychology, 84*(1), 165–176. <https://doi.org/10.1037/0022-3514.84.1.165>
- Tesser, A. (1988). Toward a self-evaluation maintenance model of social behavior. In *Advances in experimental social psychology* (Vol. 21, pp. 181-227). Academic Press. [https://doi.org/10.1016/S0065-2601\(08\)60227-0](https://doi.org/10.1016/S0065-2601(08)60227-0)
- Van der Leer, L., & McKay, R. (2017). The optimist within? Selective sampling and self-deception. *Consciousness and cognition, 50*, 23-29. <https://doi.org/10.1016/j.concog.2016.07.005>
- Vaughan-Johnston, T. I., & Jacobson, J. A. (2020). “Need” personality constructs and preferences for different types of self-relevant feedback. *Personality and Individual Differences, 154*, 109671. <https://doi.org/10.1016/j.paid.2019.109671>
- Westerwick, A., Sude, D., Brooks, D., Kaplan, B., & Knobloch-Westerwick, S. (2023). Self-Consistency and Self-Enhancement Motivation Impacts on Selective Exposure to Politics—A SESAM Model Application. *Mass Communication and Society, 26*(2), 300-325. <https://doi.org/10.1080/15205436.2022.2056854>

- Zhang, Y., Pan, Z., Li, K., & Guo, Y. (2018). Self-serving bias in memories. *Experimental Psychology*, 65(4), 236–244. <https://doi.org/10.1027/1618-3169/a000409>
- Ziano, I., Mok, P. Y., & Feldman, G. (2021). Replication and extension of Alicke (1985) better-than-average effect for desirable and controllable traits. *Social Psychological and Personality Science*, 12(6), 1005-1017. <https://doi.org/10.1177/vaughan1948550620948973>

Supplementary Material

A. Screenshots from the experiment

Please click directly on the line below to make the slider appear, then move the slider to indicate how **kind** you think you are, compared to the other people in the experiment.



This is how you ranked yourself compared to other people in the experiment for how **kind** you are.

Please take a moment to reflect on this ranking of yourself. Think about instances where you have demonstrated (or not demonstrated) this trait and also about when others have demonstrated (or not demonstrated) it. Have others ever commented about you being (or not being) **kind**? You will be able to proceed to the next question in a few seconds.



Could you think of previous situations in which you have demonstrated (or not demonstrated) this trait?

Yes, many situations.	<input type="radio"/>
Yes, some situations.	<input type="radio"/>
Yes, one situation.	<input type="radio"/>
No.	<input type="radio"/>

Could you think of previous situations where others have commented about you being (or not being) kind?

Yes, many situations.	<input type="radio"/>
Yes, some situations.	<input type="radio"/>
Yes, one situation.	<input type="radio"/>
No.	<input type="radio"/>

How easy was it for you to rank yourself on how kind you are?

Very easy.	<input type="radio"/>
Somewhat easy.	<input type="radio"/>
Somewhat difficult.	<input type="radio"/>
Very difficult.	<input type="radio"/>

Please make sure to answer all questions before proceeding.

Now that you have reflected on how **kind** you think you are, please indicate again your self-ranking on the slider below. Your previous self-ranking is shown on the slider line.

The "next page" button will appear after you have selected a value for the slider.



B. Analysis of the original personality judgements at the first stage of the experiment

To analyse the original self-rankings for the eight personality traits presented at the first stage of the experiment, we recoded the personality judgements by subtracting 50 from each (effectively recoding the 0 to 100 measurement scale to range from -50 to 50, with 0 as the midpoint). We then fitted a linear mixed model to the self-rankings with desirability and morality modelled as fixed effects and participants as a random effect, first relying on a baseline of undesirable, non-moral traits (Table B1, Model B1) and then changing the baseline to desirable, moral traits (Table B2, Model B2).

The intercept for Model B1 (Table B1) is negative and highly significant, showing that participants ranked themselves on average as possessing less the undesirable, non-moral traits than the average person, that is they considered themselves better-than-average. Fitting the model with a baseline of desirable and moral traits yields a positive and highly significant intercept (Table B2, Model B2), hence for desirable traits, respondents also considered themselves as better-than-average, which is in line with previous literature on the better-than-average effect (Alicke & Govorun, 2005; Brown, 1986).

The positive and highly significant coefficient of desirable traits ($p < .001$, Table B1, Model B1) as well as the negative and highly significant coefficient for undesirable traits ($p < .001$, Table B2, Model B2) show that self-rankings for desirable traits are higher on average than self-rankings for undesirable traits. Contrary to previous research (*Tappin & McKay, 2017*) however, we found no evidence of a magnified better-than-average effect for moral traits as the coefficients for both morality and the interaction between morality and desirability are non-significant (Table B1, Model B1 and Table B2, Model B2).

Table B1. Estimated fixed effects of desirability and morality on self-ranking at the first stage of the experiment

	Model B1
Intercept	-18.24*** (1.14)
Desirability (desirable traits)	31.98*** (1.61)
Morality (moral traits)	1.36 (1.58)
Interaction Desirable (desirable traits) * Morality (moral traits)	-1.35 (2.21)

*Baseline levels: undesirable, non-moral traits; Number of observations: 4176; grouped by participants, N=522 Significance codes: '***' 0.001 '**' 0.01 '*' 0.05;*

Table B2. Estimated fixed effects of desirability and morality on self-ranking at the first stage of the experiment

	Model B2
Intercept	13.75*** (1.04)
Desirability (undesirable traits)	-30.63*** (1.51)
Morality (non-moral traits)	-0.01 (1.54)
Desirability (undesirable traits) * Morality (non-moral traits)	-1.35 (2.21)

*Baseline levels: desirable, moral traits; Number of observations: 4176; grouped by participants, N=522; Significance codes: '***' 0.001; '**' 0.01; '*' 0.05; '.' 0.1*

C. Analysis excluding participants, who detected the manipulation

Table C1. Estimated fixed effects for the difference between revised and original self-rankings

	Model C1	Model C2	Model C3	Model C4
	total sample	total sample	middle zone	middle zone
Intercept (non-manipulated traits)	1.18** (0.39)	0.99* (0.44)	1.17*** (0.48)	1.14* (0.53)
Enhancing manipulation	12.74*** (0.55)	12.76*** (0.78)	12.55*** (0.62)	13.04*** (0.89)
Diminishing manipulation	-11.61*** (0.43)	-11.00*** (0.64)	-10.15*** (0.61)	-10.44*** (0.82)
Morality (moral traits)	-0.40 (0.47)	-0.04 (0.60)	0.49 (0.60)	0.55 (0.77)
Interaction Enhancing*Morality (moral traits)		-1.11 (0.87)		0.69 (1.23)
Interaction Diminishing*Morality (moral traits)		-0.004 (1.10)		-0.95 (1.24)

*Baseline levels: non-manipulated, non-moral; Number of observations: total sample 2088, grouped by participants, N=522; middle zone: 1392; grouped by participants, N=348; Significance codes: '***' 0.001; '**' 0.01; '*' 0.05;*

Table C2. Estimated fixed effects for enhancing and diminishing manipulations, morality, and their interaction on the *magnitudinal* difference between revised and original self-rankings.

	Model C5	Model C6	Model C7	Model C8
	total sample	total sample	middle zone	middle zone
Intercept (diminishing manipulation)	10.34*** (0.64)	10.13*** (0.71)	9.27*** (0.89)	9.44*** (0.91)
Enhancing manipulation	3.22*** (0.81)	3.80** (1.12)	5.29*** (0.99)	4.91*** (1.36)
Morality (moral traits)	0.77 (0.77)	1.16 (0.95)	-0.87 (0.99)	-1.26 (1.39)
Interaction Enhancing*Morality (moral traits)		-1.12 (1.61)		0.80 (1.97)

Baseline levels: *diminishing, non-moral*; Number of observations: total sample 1044; grouped by participants, N=522; middle zone: 696; grouped by participants, N=348; Significance codes: '***' 0.001; '**' 0.01; '*' 0.05;

D. Analysis, including desirability and type of manipulation as fixed effects.

As pre-registered, we have analysed with desirability, type of manipulation and morality as fixed effects and participants as a random effect (Table D1, Models D1 to D6).

Table D1. Estimated fixed effects on revised self-rankings, including desirability, manipulation up, manipulation down, morality and interactions between desirability and manipulation up and manipulation down as a fixed effects.

	Model D1 total sample	Model D2 total sample	Model D3 total sample	Model D4 middle zone	Model D5 total sample	Model D6 middle zone
Intercept	0.0006 (0.56)	0.49 (0.67)	1.21* (0.50)	0.57 (0.72)	-0.25 (0.83)	1.27 (0.65)
Desirable	1.96** (0.72)	1.98** (0.72)	-0.05 (0.60)	1.54 . (0.87)	1.58 . (0.87)	-0.31 (0.77)
Manipulation Up	-2.69*** (0.59)	-2.69*** (0.59)	- 12.40*** (0.60)	1.61* (0.75)	1.58* (0.75)	-12.84*** (0.92)
Manipulation Down	-3.41*** (0.62)	-3.42*** (0.62)	11.51*** (0.82)	0.36* (0.76)	0.39 (0.76)	11.37*** (0.94)
Moral		-0.95 (0.71)	-0.33 (0.47)		1.67 (0.86)	0.72 (0.59)
Interaction Desirable * Manipulation Up			25.70*** (0.95)			25.84*** (1.22)
Interaction Desirable * Manipulation Down			- 21.79*** (1.02)			-18.90*** (1.24)

Baseline levels: non-manipulated, undesirable non-moral; Number of observations: total sample 2140; grouped by participants, N=535; middle zone: 1440; grouped by participants, N=360; Significance codes: '***' 0.001; '**' 0.01; '*' 0.05; '.' 0.10

E. Result 2, total sample

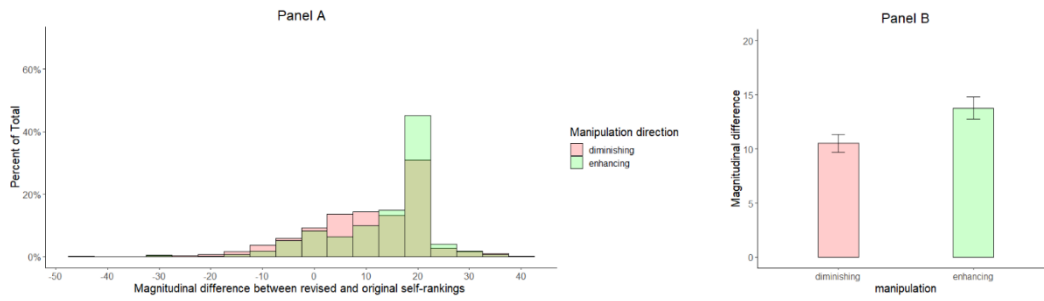


Figure E1. Histogram for the *magnitudinal difference* of enhancing and diminishing manipulation, total sample (Panel A) and bar chart for the *magnitudinal difference* for enhancing and diminishing manipulations, total sample (Panel B).

F. Analysis, including the initial self-rankings as predictor of the revised self-rankings.

As a robustness test, specified in our pre-registration document, we also fitted linear mixed models explaining the revised self-rankings with the original self-rankings, desirability, morality and the type of manipulation as fixed effects and participants as random effects (Table F1, Models F1 and F3). Models F2 and F4 built on Models F1 and F3 respectively by including an interaction term between desirability and type of manipulations (Table F1). The analysis conveyed the same general message as the main analysis: the manipulation is effective, and the observed effect is stronger for enhancing manipulations.

Table F1. Estimated fixed effects on revised self-rankings, including average self-rankings at the first stage of the study as a predictor.

	Model F1 total sample	Model F2 total sample	Model F3 middle zone	Model F4 middle zone
Intercept	10.29*** (1.10)	4.93*** (0.87)	0.73 (1.50)	3.50** (1.19)
Initial self-rankings	0.85*** (0.01)	0.94*** (0.01)	0.98*** (0.02)	0.96*** (0.02)
Desirable	1.32 . (0.68)	-0.33 (0.60)	1.61 . (0.86)	-0.28 (0.77)
Manipulation Up	-2.56*** (0.58)	-12.09*** (0.60)	1.51*** (0.76)	-13.06*** (0.92)
Manipulation Down	-3.65*** (0.61)	10.77*** (0.83)	0.31 (0.76)	11.24*** (0.94)
Moral	-0.78 (0.67)	-0.29 (0.47)	1.60 (0.86)	0.56 (0.60)
Interaction Desirable * Manipulation Up		25.01*** (0.95)		25.95 (1.22)
Interaction Desirable * Manipulation Down		-20.84*** (1.04)		-18.96*** (1.23)

*Baseline levels: non-manipulated, undesirable non-moral; Number of observations: total sample 2140; grouped by participants, N=535; middle zone: 1440; grouped by participants, N=360; Significance codes: '****' 0.001; '***' 0.01; '**' 0.05;*

Each of the Models F1 to F4 demonstrates the effect of the manipulation on personality judgements; For instance, in Model F4 (Table F1), self-rankings for non-

manipulated traits will add the value of the intercept (3.50, $p < .010$) and the initial self-rankings (0.96, $p < .001$). For manipulated traits the revised self-rankings will be either higher with 12.89 for desirable traits, manipulated up, $p < .001$) or lower with 7.72 (for desirable traits manipulations down ($p < .001$)).

To test whether enhancing manipulations have stronger effect than diminishing, we have fitted the above-specified models on manipulated traits only Table F2, Models F5 to F8). Running the analysis in the middle zone (as pre-specified) and in the total sample, shows that enhancing diminishing manipulations decrease revised self-rankings while enhancing increase them. For example, revised self-rankings are on average 8.24 lower for diminishing manipulation ($p < .001$) and higher with 23.17 ($p < .001$, Table F2, Model F8) for enhancing manipulations. To compare formally whether enhancing manipulations have a stronger effect, we have run a Wald test for equality of coefficients, which showed that the coefficients for enhancing manipulations is significantly higher than the one for diminishing manipulations (the intercept), $W(1) = 145.87$, $p < .001$.

Table F2. Estimated fixed effects for the difference between revised and original self-rankings

	Model F5	Model F6	Model F7	Model F8
	total sample	total sample	middle zone	middle zone
Intercept (diminishing manipulation)	-6.30*** (1.41)	-6.15*** (1.42)	-7.98*** (1.89)	-8.24*** (1.91)
Average self-rankings at the first stage	0.94*** (0.02)	0.94*** (0.02)	0.98*** (0.03)	0.98*** (0.03)
Enhancing manipulation	23.29*** (0.86)	22.74*** (1.18)	22.42*** (0.99)	23.17*** (1.37)
Morality (moral traits)	-0.07 (0.78)	-0.47 (0.98)	0.69 (0.99)	1.45 (1.39)
Interaction		1.09		-1.54
Enhancing*Morality (moral traits)		(1.62)		(1.97)

*Baseline levels: diminishing, manipulated down, non-moral; Number of observations: total sample 1070; grouped by participants, N=535; middle zone: 720; grouped by participants, N=360; Significance codes: '***' 0.001; '**' 0.01; '*' 0.05;*

G. Secondary and Exploratory Hypotheses

H3 (secondary). For self-rankings in the middle zone, enhancing manipulations will lead to higher subsequent self-esteem.

H4 (secondary). Revised self-rankings on non-manipulated traits will on average deviate from the original rankings in a self-enhancing direction. This effect will be more pronounced for moral traits

H5. Participants will be even more accepting of enhancing manipulations and less accepting of diminishing manipulations on undesirable than desirable traits: revised self-rankings for undesirable traits manipulated in a enhancing direction will deviate more (in the direction of the manipulation) from the original self-rankings than revised self-rankings for desirable traits manipulated in a enhancing direction. Vice versa for diminishing manipulations.

H7 (exploratory). For self-rankings in the middle zone, enhancing manipulations will lead to higher subsequent self-rankings for how good/happy/capable one is.

To test H3 and H7, we regressed the respective measures (self-esteem, good, happy, capable) on enhancing manipulations and morality. We found no support for an elevating effect of enhancing manipulations on self-esteem or on any of the other measures. If anything, it seems that enhancing manipulations were associated with lower subsequent self-rankings of how “good” one is (Table G1).

Table G1. Linear regressions on measures of self-esteem, goodness, happiness and capability

	Self-esteem	Good	Happy	Capable
Intercept	-1.84 (8.34)	25.48*** (6.26)	-1.31 (8.37)	22.78** (6.98)
Average initial self-rankings	0.72*** (0.13)	0.76*** (0.10)	0.86*** (0.13)	0.71*** (0.11)
Enhancing manipulation	-2.80 (2.70)	-4.25* (2.02)	-2.13 (2.70)	-0.82 (2.25)
Morality (moral traits)	2.27 (2.73)	-0.17 (2.05)	2.51 (2.74)	-1.38 (2.28)

Baseline levels: diminishing, non-moral; N=360, manipulations of self-rankings in the middle zone (between 25 and 75)

Significance codes: '****' 0.001; '***' 0.01; '**' 0.05;

H4 (secondary). Revised self-rankings on non-manipulated traits will on average deviate from the original rankings in a self-enhancing direction. This effect will be more pronounced for moral traits

Figure G1 shows that revisions in non-manipulated traits were always in a direction that elevated the self-image with enhancing manipulations having a slightly stronger effect than diminishing manipulations.

Table G2. Effect of the manipulation for non-manipulated traits (presented for revision along with the manipulated traits)

	Model G1	Model G2	Model G3	Model G4
	total sample	total sample	middle zone	middle zone
Intercept (diminishing manipulation)	0.47 (0.30)	0.44 (0.40)	0.82 (0.44)	0.62 (0.52)
Enhancing manipulation	1.53** (0.51)	1.54** (0.51)	1.28* (0.63)	1.24* (0.63)
Morality (moral traits)		0.06 (0.49)		0.47 (0.63)

Baseline levels: diminishing, non-moral; Number of observations: total sample 1070; grouped by participants, N=535; middle zone: 720; grouped by participants, N=360; Significance codes: '***' 0.001; '**' 0.01; '*' 0.05;

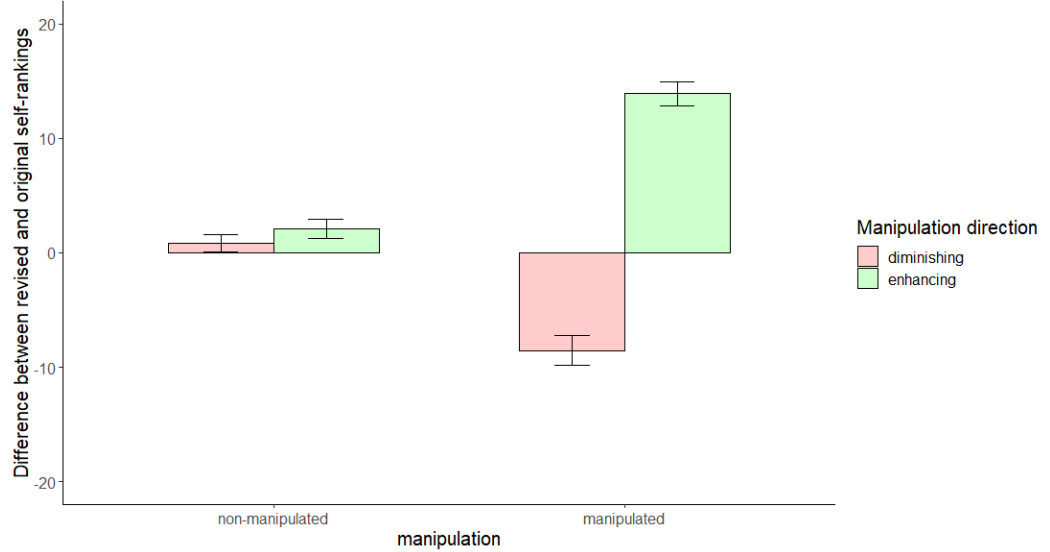


Figure G1. Difference between revised and original self-rankings for manipulated and non-manipulated traits (self-rankings for undesirable traits are reverse-coded), middle zone

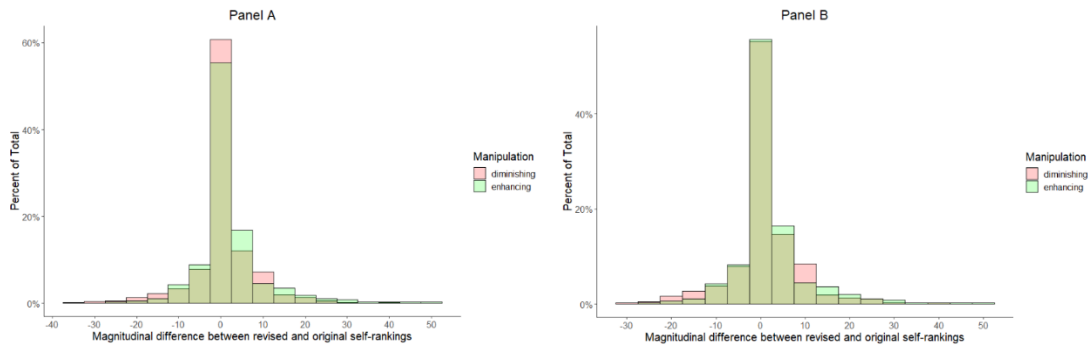


Figure G2. Histogram for the *magnitudinal difference* for non-manipulated traits in the enhancing and diminishing manipulation condition, total sample (Panel A) and middle zone (Panel B).

H5. Participants will be even more accepting of enhancing manipulations and less accepting of diminishing manipulations on undesirable than desirable traits: revised self-rankings for undesirable traits manipulated in an enhancing direction will deviate more (in the direction of the manipulation) from the original self-rankings than revised self-rankings for desirable traits manipulated in an enhancing direction. Vice versa for diminishing manipulations.

To investigate H5, we fitted a linear mixed model on the magnitudinal difference between the revised and original self-rankings for manipulated traits with desirability, morality and their interaction modelled as fixed effects and participants as a random effect, first on enhancing manipulations only (Model G5) and then on diminishing manipulations only Model G6 (Table G3).

The results did not provide support for H5: there was no differential effect on desirable and undesirable traits for neither enhancing manipulations (the coefficient for desirability is non-significant, $p = .639$, Table F3, Model F5), nor for diminishing manipulations (the coefficient for desirability is non-significant, $p = .093$, Table G3, Model G6).

Table G3. Estimated fixed effects of desirability and morality on the *magnitudinal* difference between revised and original self-rankings for enhancing and diminishing manipulations, middle zone

	Model G5	Model G6
	enhancing	diminishing
	manipulations	manipulations
Intercept	7.56***	5.47***
	(0.90)	(0.94)
Desirability (desirable traits)	0.56	-2.05
	(1.20)	(1.21)
Morality (moral traits)	0.47	-0.48
	(1.27)	(1.36)
Interaction Desirability (desirable traits) *	-0.41	-1.07
Morality (moral traits)	(1.66)	(1.83)

*Baseline levels: undesirable, non-moral; Model G5: number of enhancing observations: 704; grouped by participants, N=176; Model G6: number of diminishing observations: 736; grouped by participants, N=184; Significance codes: '***' 0.001; '**' 0.01; '*' 0.05;*

Chapter 8. Discussion

Discussion

This thesis investigated the malleability of individuals' self-concepts by extending anchoring (Tversky & Kahneman, 1974) and the choice blindness (Johansson et al., 2005) paradigms to the domain of the self. The paradigms I employed incorporated two essential influences, namely information about others' behaviour and individuals' own previous behaviour. I focussed on the interplay between cognitive influences and positive self-image considerations, exploring when personality traits exhibit flexibility and when they display rigidity. In addition, I investigated whether manipulating judgements of one's own moral qualities has any aftereffects on subsequent personality judgements or prosocial choices. Throughout the studies, I contrasted attitudes in the moral and non-moral domain to explore whether moral behaviours are more sensitive to cognitive influences.

Specifically, in Study 1, I explored whether choices with moral implications are *more malleable* than choices without moral implications; Study 2 extended the anchoring paradigm to the personality domain and investigated the effect of anchoring on personality judgements, testing in particular for a *self-serving bias* in individuals' susceptibility to anchoring; Study 3 explored potential *aftereffects* of anchored moral personality judgements on the general self-image by measuring subsequent rankings on a set of personality traits; Study 4 investigated potential *aftereffects* of anchored moral personality judgements on the general self-image as measured by donations in subsequent prosocial choices. Study 5 relied on the choice blindness manipulation to investigate the effect of one's own (or alleged) previous behaviour on personality judgements. The rest of this chapter is organised as follows: (i) Summary of the main results; (ii) Critical discussion of the results: strengths and limitations; (iii) Implications and future research directions and (iv) Summary and conclusions.

Summary of the main results

In Study 1, I utilised a newly developed task ("wheel of fortune" task, Appendix 3B, Chapter 3) to contrast choices with and without moral implications. The task

presented participants with a risky choice, designed in a way that precludes a straightforward interpretation of what constitutes prosocial or selfish behaviour. Subsequently, I exposed participants in the experimental condition to the influence of anchoring (Tversky & Kahneman, 1974). In addition, I incorporated social information into the anchoring paradigm by presenting the anchor value as the average behaviour others had exhibited in a previous round of the study. The main finding was that both choices with and without moral implications are susceptible to anchoring. However, I found no evidence for a differential social anchoring effect depending on whether the choice entailed moral implications or not.

In Study 2, I extended the anchoring framework to the personality domain and exposed personality judgements to anchors that either enhanced or diminished individuals' self-rankings. The central finding was that personality judgements are susceptible to anchoring when the anchors elevate the self-image while diminishing anchors were found to have little or no effect on personality judgements. Participants' personality judgements also showed evidence for the "better-than-average" effect (Alicke & Govrun, 2005). Although participants ranked themselves more positively on moral than non-moral traits, neither the self-serving anchoring effect nor the better-than-average effect were particularly exacerbated for moral personality judgements.

In Study 3, I investigated whether anchoring morally relevant personality judgements would affect the general self-image. Following the anchoring phase, I asked participants to indicate their self-rankings on a new set of personality traits (desirable and undesirable, moral and non-moral). I replicated the self-serving anchoring effect we observed in Study 2, however there was no evidence for an aftereffect on subsequent self-rankings. Morality was found to have a general enhancing effect on personality judgements regardless of the anchoring condition; although we selected moral and non-moral traits with comparable average desirability rankings, the data revealed substantially higher self-rankings on moral than non-moral traits in the control condition, which prevented me from contrasting the aftereffect on moral and non-moral traits. In an exploratory analysis, I also found evidence for a "phrasing effect" with

participants indicating higher self-rankings, on average, for negatively than positively phrased traits.

Study 4 also tested for a potential aftereffect of anchored morally relevant personality judgements on the general self-image, however in this study I measured overt behaviour in a subsequent prosocial choice (Dictator Game). I did not replicate the self-serving anchoring effect in this study. However, there was evidence for nearly 15% more generous donations in the Dictator Game after participants were exposed to the influence of enhancing anchors. The phrasing effect observed in Study 3 was replicated in the no anchor and diminishing anchor conditions.

In Study 5, I extended the choice blindness framework (Johansson et al., 2005) to the personality domain and investigated whether personality judgements are flexible in response to information about one's own (alleged) previous behaviour. I found that individuals' personality judgements are susceptible to choice blindness manipulations with enhancing manipulations being accepted to a higher degree than diminishing ones. I found no evidence of a magnified choice blindness effect for moral personality traits. Nevertheless, there was again a general effect of morality with participants ranking themselves more positively on moral than non-moral personality characteristics.

Critical evaluation of findings: strengths and limitations

As the results from each study were already discussed independently in each respective chapter, the discussion here is structured by topics, linking findings from different studies. The rest of the subsection is organised as follows: (i) Flexibility of choices and the effect of morality on decisions and personality judgements; (ii) Flexibility of personality judgements; (iii) Repercussions for the general self-image and (iv) Limitations of the research.

Flexibility of choices and the effect of morality on choices and personality judgements

In line with literature showing the robustness and replicability of the anchoring phenomenon across decision-making domains (e.g., Röseler & Schütz, 2022), Study 1 showed that anchoring influenced behaviour in risky choices both for participants that

played for themselves and for those that took decisions on behalf of a chosen charity. The anchor value I employed was imbued with social meaning, which likely exacerbated the impact of the anchor. Although from a standard economic point of view, social reference points such as others' decisions, should not influence behaviour in risky choices, research has shown that social reference points impact individuals' behaviour (Gamba et al., 2017; Schwerter. 2023). As participants in Study 1 were provided with social information, it might have served as a social reference point, magnifying the effect of the anchor value.

Nevertheless, I found no evidence for an exacerbated social anchoring effect on choices with moral implications. To contrast moral and non-moral choices, I relied on a newly developed task (the "wheel of fortune" task, please see Appendix 3A, Chapter 3). My supervisors and I invested a lot of time and effort designing a task suitable for comparing moral and non-moral choices in an experimentally valid way. To be able to differentiate the effect of our manipulation from the potential influence of social norms (Andreoni & Bernheim, 2009), the task should be devoid of contextual social norms and of a straightforward interpretation of what constitutes a fair or selfish choice. In retrospect, however, we realised that in aiming to construct an experimentally valid task, we designed a choice with moral implications rather than a classic moral choice, such as for example the trolley dilemma (Greene et al., 2001). It turned out to be challenging (if possible) to construct an experimentally valid non-moral equivalent of a classic moral choice.

Previous research shows the paramount importance of morality both as guiding values and as shaping our self-concept (Schwartz & Cieciuch, 2022; Strohminger, 2018). Moral norms seem to also anchor individuals' behaviours when they navigate the social world (Ellemers et al., 2013; Ellemers & van Nunspeet, 2020). Judgements and choices in the moral domain are also associated with much stronger emotional responses (Greene et al., 2001; Rozin et al., 1999). Certain cognitive effects, such as the sunk cost effect and the illusion of moral superiority are magnified in the moral domain (Meyers et al., 2019; Tappin & McKay, 2017). Recent research has also shown that comparisons in the

moral domain differ from other social comparisons (Fleisgmann et al., 2021). As there is both theoretical and empirical evidence suggesting that the anchoring effect should be more pronounced in the moral domain, the lack of differential anchoring effect in Study 1 may have been due to the task I employed, which might not have captured the essence of a classic moral choices.

I continued my investigation on the effect of morality by contrasting moral and non-moral personality judgements in Study 2, Study 3 and Study 5. I focussed on the personality domain as it offers an experimentally valid way to contrast moral and non-moral attitudes. In addition, morality has been outlined as central for the self-concept with moral traits defining to a large extent how we perceive ourselves (Heiphetz et al., 2016; Strohminger & Nichols, 2014). I selected personal characteristics that had a clear categorisation as moral or non-moral (for e.g., “kind” and “intelligent”) and tested for differential cognitive effects on the respective personality judgements. I also controlled for desirability as research has outlined its impact on personality judgements (e.g., Santioso et al., 1990; Ziano et al., 2021). Hence, I chose moral and non-moral traits that had comparable average desirability self-rankings (Tappin & McKay, 2017; Ziano et al., 2021).

In Study 2, I found that both moral and non-moral traits are susceptible to self-serving anchoring, however there was no magnified effect for moral traits. In Study 3, contrasting the self-serving anchoring effect on moral and non-moral traits was precluded by the substantially higher self-rankings on moral traits in the control condition. Moral personality judgements were also not more susceptible to the choice blindness manipulations in Study 5 than non-moral personality judgements. Nevertheless, there was a general effect of morality in Studies 2, 3 and 5 with individuals ranking themselves more positively on moral than non-moral personality traits, which is in line with literature outlining the importance of morality to the self-concept (e.g., Strohminger, 2018).

A potential explanation for the lack of a magnified effect for moral personality traits might stem from the fact that I might not have managed to select a set of moral

personality traits that would evoke individuals' moral identity (Aquino & Reed, 2002). Our primary selection criterion was to control for desirability by selecting moral and non-moral traits with comparable average desirability self-rankings. Desirability had a strong and pronounced effect on personality judgements in all the studies, which interacted with the respective cognitive influences I employed and resulted in the observed self-serving anchoring (Studies 2 and 3) and higher acceptance of choice blindness manipulations that enhance the self-image (Study 5). It might be challenging to elicit an association with individuals' moral identity controlling for the effects of desirability; the effect of morality and desirability might be hard to disentangle if moral traits are intrinsically highly desirable.

In addition, I found an interesting phrasing effect for morally relevant personality judgements in Study 3 and Study 4. Both studies shared the same first stage with participants indicating personality judgements on two morally relevant traits, either desirable or undesirable. I chose personality traits reflecting the same quality phrased in a positive ("honest", "considerate") or negative way ("dishonest", "inconsiderate"). Participants' personality judgements were anchored either in an enhancing or diminishing direction. The results showed that participants were more sensitive to the negative phrasing of personality traits, such that they indicated more positive self-rankings on negatively than positively phrased personality traits. These findings are reminiscent of the loss aversion phenomenon (Kahneman & Tversky, 1979) – the "loss" of being more dishonest might have loomed larger than the "gain" of being more honest. Furthermore, phrasing the personality traits in a negative way might have triggered self-protection motivation, which has been shown to be stronger than self-enhancing motivation (Alicke & Sedikides, 2009).

To sum up, although I did not find a more pronounced self-serving anchoring effect or choice blindness manipulation for moral traits, the data showed evidence for a general effect of morality when it comes to self-appraisals of personality traits. Participants ranked themselves more positively on moral than non-moral personality traits. It is an intriguing research question to disentangle the effect of desirability and

morality and test whether moral traits are intrinsically more desirable (or undesirable) than non-moral traits (please see the “Implications and future directions” below for a discussion).

Flexibility of personality judgements

The interplay of cognitive influences and positive self-image concerns and how they determine the flexibility (or rigidity) of personality judgements is the focus of this thesis. I extended two cognitive paradigms, anchoring (Study 2, Study 3, and Study 4) and choice blindness (Study 5) to the domain of the self. Taken together, the results provide evidence that personality traits can exhibit both stability and flexibility in response to cognitive influences that incorporate either benchmarks of others’ or one’s own (alleged) previous behaviour. The message emerging across studies is that maintaining a positive self-image is an important factor for the construction and adjustment of personality judgements with personality traits’ stability or flexibility depending on the repercussions the respective influence has for the self-view. Personality judgements seem to quickly adjust in response to cognitive influences when the potential revisions elevate the self-view while remaining relatively rigid in response to diminishing manipulations.

The simultaneous flexibility and stability of personality judgements has several interesting aspects. First, personality traits are theorised as consistent behavioural patterns across contexts (Roberts, 2009). Although recent research allows some degree of flexibility of personality traits throughout the lifespan (e.g., Ferguson, 2010), personality judgements are supposed to be stable enough to resist cognitive influences. Our results however showed that personality judgements can adjust instantaneously in response to cognitive influences, and thus point towards a somewhat different psychological mechanism shaping the construction and adjustment of personality judgements than that currently theorised (Roberts & Yoon, 2022).

Dweck (2017) suggests that there are three basic needs (competence, predictability, and acceptance) and that personality traits are constructed with the goal to fulfil these needs. Such a theoretical perspective seems to allow certain flexibility of

personality traits to ensure that goals stemming from the three basic needs are achieved across contexts and over time. This might explain why personality judgements were readily adjusted only when the revisions were in an enhancing direction: acceptance and maintaining a positive self-image are important personal goal (e.g., Alicke, 2013), thus enhancing manipulations were embraced as fulfilling the need of acceptance, diminishing manipulations however were resisted as they threaten the positive self-image. Therefore, the stability or flexibility of personality judgements might be a function of its ability to achieve the goal of ensuring a stable and positive self-view.

A related, but somewhat different account of our findings stems from the literature on self-protection and self-enhancing motivation, which outlines both motives as influencing behaviour with self-protection having a stronger impact (Alicke & Sedikides, 2009; Sedikides & Strube, 1997). Indeed, self-enhancing motivation would exacerbate the effect of enhancing anchors and lead to higher levels of acceptance of flattering choice blindness manipulations. At the same time, self-protecting motivation would render diminishing anchors less effective and would result in lower levels of acceptance of diminishing choice blindness manipulations. Furthermore, research on the stability of the self-concept and self-continuity has shown that while individuals are averse to diminishing changes in their moral traits, enhancing changes in moral traits are embraced. Although stable, the self-concept seems flexible to the degree that it integrates certain expectations for natural improvement over time (Molouki & Bartels, 2017; Newmann et al., 2014; Newmann et al., 2015).

To return to the flexibility of judgements, another way to think of our results is that instead of stemming from a notion of an absolute true self, personality judgements are constructed in relative terms, in comparison to others (Crusius et al., 2022; Festinger, 1954). However, if a comparison process shapes the construction of personality judgements, it might not necessarily be that others are the adopted standard for comparison (although they often are). Research has shown that salient reference points influence behaviour (Kahneman & Tversky, 1979; Kahneman, 1992). Naturally, others' attitudes are a salient reference point in self-evaluation. However, if

individuals are exposed to other salient standards for comparison, they might adopt them instead in constructing their personality judgements. Indeed, while Study 2 relied on anchoring that evoked natural comparison with others, Study 5 highlighted a very different potential standard for comparison – one’s own previous behaviour.

Nevertheless, in both Study 2 and Study 5, there was another factor that seemed to influence the adjustments of personality judgements, namely the positive view participants hold of themselves. The positive self-image might be considered as a salient benchmark that defines the degree of flexibility of personality traits – any adjustments that potentially threaten individuals’ positive self-image are resisted. People often make judgements in an environment that presents them with multiple influential factors or standards for comparison (Kahneman, 1992; Schwerter, 2013), however the interplay among different factors and the exact mechanism through which one or another standard for comparison prevails in shaping behaviour is yet to be well understood. In any case, our research points towards positive self-image as a potential salient factor that can outweigh the effect of other influences such as anchors, others’ behaviour or our own (alleged) previous behaviour.

Indeed, if the anchoring effect I observed in Study 2 was influenced only by the salient anchor value, the enhancing or diminishing nature of that anchor should not have mattered. Yet, enhancing anchors had a very different impact on personality judgements than diminishing anchors. In a similar vein, if the acceptance of the choice blindness manipulation in Study 5 was due only to self-consistency motivation, I should not have observed higher acceptance levels for enhancing manipulations. The way I have defined “enhancing” and “diminishing” anchors or choice blindness manipulations incorporates the effect of the manipulation on the self-image. Therefore, it seems that maintaining a positive self-image facilitates or restricts the effect of cognitive influences on individuals’ personality judgements.

The observed importance of preserving a positive self-image might also shed additional light on the psychological processes shaping anchoring. The prevailing model explaining the mechanism of anchoring, the Selective Accessibility Model, theorises

anchoring as a hypothesis-confirmatory search rendering evidence in line with the hypothesis more accessible (Mussweiler & Strack, 1999a, 1999b, 2000a, 2000b, 2001; Bahník & Strack, 2016; Mussweiler, 2003; Strack & Mussweiler, 1997). The premises of the Selective Accessibility Model (Mussweiler & Strack, 1999a) have recently been questioned (Bahník, 2021; Harris et al., 2019). Harris and colleagues (2019) reasoned that if selective accessibility is increased, agents should be faster to recognise anchor-relevant information when it is consistent with the anchor. Thus, the signature test for the model has been to analyse response speed in a categorisation task following the comparative judgement in the anchoring paradigm (Mussweiler & Strack, 2000a). In a series of experiments, Harris et al. (2019) registered strong anchoring effects, however, there were no significant differences in the response speed of anchor-consistent and anchor-inconsistent information recognition. Yet as stated by these researchers (Bahník, 2021; Harris et al., 2019), it is methodologically challenging to reliably measure the presumed semantic priming and provide decisive evidence for or against the Selective Accessibility Model.

Currently, no other theory of anchoring, such as numeric priming (Wong & Kwong, 2000), insufficient adjustment (Tversky & Kahneman, 1974), scale distortion (Frederick & Mochon, 2012) or resource-rational anchoring and adjustment (Lieder et al., 2018), can explain and credibly predict when and to what degree anchoring will occur. Research on individual differences and anchoring has also yielded contradictory findings (Cheek & Norem, 2020). For instance, Bergman et al. (2010) showed that higher cognitive abilities are associated with weaker anchoring effects while Teovanović (2019) found no evidence that cognitive ability moderates individuals' susceptibility to anchoring. Moreover, relying on a within-subject design, researchers have shown lack of consistency in susceptibility to anchoring at an individual level as measured by cognitive ability and have called for rethinking of the potential moderators of anchoring (Röseler et al., 2019).

Although Study 2 was not designed with the aim of differentiating between theoretical accounts of anchoring, in retrospect it seems that our findings can provide

additional insight into the psychological mechanisms shaping anchoring. If anchoring results from adjustment of the anchor value (Tversky & Kahneman, 1974) or from numeric priming (Wong & Kwong, 2000), individuals' personality judgements should be influenced to the same degree by the numeric value of the anchor (low or high), regardless of the anchor's repercussions for the self-image (enhancing or diminishing). If, however, the anchor values render evidence in line with the anchor more accessible (Mussweiler & Strack, 1999a), this anchor-consistent information search, coupled with the higher accessibility of positive self-relevant information (Ritchie et al., 2017; Sedikides & Green, 2009; Santioso et al., 1990; Sedikides et al., 2016; Zhang et al., 2018) would produce the observed self-serving anchoring effect in Studies 2 and 3.

Furthermore, a process of assimilation towards the anchor value underlies all current explanations of the anchoring mechanism (Strack et al., 2016). However, enhancing and diminishing anchors might trigger different psychological mechanisms depending on their implications for the self-image. Pinter et al. (2011) suggest that only positive self-relevant information is assimilated into self-knowledge while negative self-relevant information is separated or contrasted away from stored self-knowledge. In a similar vein, enhancing anchors might elicit assimilation processes, facilitated also by the high accessibility of positive self-relevant information while diminishing anchors might evoke contrasting away processes as negative self-relevant information is not assimilated into the self-image.

To return to the Selective Accessibility Model, instead of different psychological processes being triggered by enhancing and diminishing anchors, the anchoring process might rely on a unified process. The underlying mechanism could be hypothesis-confirmatory search, however, enhancing and diminishing anchors might induce selection of opposite hypotheses. Researchers have theorised comparison as a two-stage process: first, individuals conduct a quick, holistic assessment of their similarity with the proposed standard for comparison (the anchor value) and decide whether they are similar or dissimilar to it (Hanko et al., 2010; Mussweiler, 2003; Mussweiler, 2001a;

Mussweiler & Strack, 2000). Subsequently, a hypothesis-confirmatory search is elicited, which – depending on the selected hypothesis – will either search for evidence for similarity or dissimilarity. As research has shown that individuals hold a positive image of themselves and even believe they rank higher than the average person on desirable traits and lower than the average person on undesirable traits (the better-than-average effect, Alicke & Govorun, 2005), the initial assessment stage for enhancing anchors would result in choosing a similarity hypothesis while diminishing anchors would induce selection of a dissimilarity hypothesis.

Thus, for enhancing anchors, a search for evidence showing similarity with an elevated self-image will be evoked, which would tap into a pool of highly accessible positive self-relevant information (e.g., Ritchie et al., 2017). In contrast, diminishing anchors will elicit a dissimilarity search, which would tap into the same pool of highly accessible *positive* self-relevant information, however this evidence would result in contrasting away rather than assimilation to the diminishing anchor. The Selective Accessibility Model seems to offer the most plausible framework to explain our results: anchor-consistent information is more accessible, however positive self-relevant information is also selectively accessible, which enhances the recall of desirable behaviours and magnifies the effect of self-serving anchors while negative self-relevant information is separated from self-knowledge, which restricts its accessibility during the hypothesis-confirmatory search, rendering self-diminishing anchors ineffective.

In a similar vein, our findings contribute to the understanding of the psychological mechanisms shaping the choice blindness effect. If individuals' acceptance of choice blindness manipulations stems from self-consistency motivation only, there should not have been higher acceptance levels for enhancing manipulations. Hence, besides self-consistency, self-enhancement motivation is shaping individuals' susceptibility to the choice blindness effect. Again, hypothesis-confirmatory search and the selective accessibility of positive self-relevant information might explain the observed stronger influence of enhancing choice blindness manipulations.

Although there is no explicit comparison phase in the choice blindness framework, one's own previous behaviour might be a natural standard for comparison due to individuals' wish to behave in a consistent manner (Gazzaniga, 2000; Johansson et al., 2012; Swann, 1987). In such cases, the most plausible hypothesis to be selected in view of the salience of one's own previous behaviour is that of similarity. Once the similarity hypothesis is adopted, a hypothesis-confirmatory search will be initiated. As positive self-relevant information is highly accessible, participants could easily justify enhanced personality judgements and thus accept enhancing choice blindness manipulations. Due to the variety of psychological mechanisms ensuring negative self-relevant information is separated from self-knowledge (e.g., Möbius et al., 2022) however, it would be more challenging to find support for self-rankings manipulated in a diminishing direction. Therefore, a psychological process of comparison with a salient benchmark, followed by a hypothesis-confirmatory search might explain both the anchoring and the choice blindness effects on personality traits.

Repercussions for the general self-image

I investigated whether anchoring morally relevant personality judgements influences the general self-image as measured by subsequent personality judgements (Study 3) and prosocial choices (Study 4). In Study 5, I measured potential repercussions for self-esteem, employing the single item self-esteem measure (Robins et al., 2001) as well as collecting data about how good, happy, and capable participants believed they are.

In Study 3, I found no evidence for an aftereffect of enhancing manipulations on subsequent personality judgements. Diminishing anchors also did not have an aftereffect on subsequent personality judgements, however this finding fits well with the results from Study 2, Study 4 and Study 5 showing that individuals resist diminishing manipulations, that is personality traits are rigid in response to influences that threaten to deteriorate their self-image. These findings are also in line with literature showing the

importance of self-enhancement and self-protection motivations with individuals exhibiting even stronger motivation to protect their self-image from negative self-relevant information than to embrace elevating feedback on their self-view (Alicke & Sedikides, 2009; Gaertner et al., 2012; Sedikides & Strube, 1997; Tesser, 1988; Zhang et al., 2018).

In Study 4, I found evidence that enhancing anchors led to more generous donations in the Dictator Game. Diminishing anchors did not influence behaviour in the Dictator Game, which is in line with individuals' resistance to potential deterioration of their self-concept. Given the discussed limitations of Study 4, the obtained results should be taken with caution, yet there is evidence that self-judgements are not only susceptible to enhancing anchors, but that these quick changes to self-assessment are transferrable to subsequent behaviour. These results also seem promising as they rely on measuring overt behaviour (donations in an incentivised Dictator Game), which is the gold standard in demonstrating the potential effect of a manipulation (Back & Vazire, 2012). In addition, Studies 3 and 4 provided important insights into the most suitable way to explore the potential repercussions of self-serving manipulations to the general self-image from that could serve as a basis for future research (please see the "Implications and future directions" subsection below for a discussion).

In Study 5, I did not find support for an elevating aftereffect of enhancing manipulations on self-esteem or on the measures of how good, happy and capable participants considered themselves to be. Revised rankings of non-manipulated traits however tended to be more positive than the original rankings. This replicates earlier work by van der Leer and McKay (2017), showing that second estimates tend to be more optimistic. In retrospect, I also realised that the self-esteem measure might not be the most suitable to capture potential enhancement of the general self-image as excessively high self-esteem is not perceived favourably by others (Paulhus, 1998). Future research with a carefully designed measure of the potential aftereffect on the general self-image may provide more decisive evidence as to whether the observed quick adjustments in

personality judgements have repercussions for the general self-concept and transfer to subsequent attitudes.

Limitations of the research

All the studies in this thesis were conducted using online platforms due to several reasons, such as the necessity of relatively large sample sizes (please see Chapter 2 for a detailed discussion). I had good reason to believe that conducting the studies online would not compromise the quality of the data. Indeed, previous research has shown that participants' behaviour in various decision domains including personality measures and prosocial behaviour is comparable between online and lab-based experiments (Amir et al, 2012; Clifford et al., 2015; Hergueux & Jacquemet, 2015; Hilbig, 2016; Manago et al., 2021). I also relied on the online platform *Prolific* (www.prolific.co) which is claimed to provide higher quality data than other online platforms such as AMT, CloudResearch and panels, Qualtrics and Dynata, based on naivety of respondents, attention, comprehension, and reliability (Peer et al., 2021). In addition, conducting the studies on *Prolific* allowed me to apply pre-screening criteria, including approval rate, which should result in high-quality data.

Nevertheless, I lost a substantial amount of data due to participants failing the attention and/or comprehension checks. The loss of data affected the analysis of Study 4 the most: I had six between subject conditions and the exclusions led to relatively small numbers of participants per condition (ranging from 49 to 79 participants per condition in Study 4). Although the remaining data should be of high quality, the fact that I lost such a large proportion of our participants raises questions about the quality of the rest of the data. It would be interesting to conduct similar studies, testing whether personality judgements have any repercussions for the general self-image, in a lab environment, where I could better control for participants' comprehension as well as attentiveness to the tasks.

Another limitation of Study 3 and Study 4 is that they relied on two morally relevant traits to activate the moral dimension of the self-concept and thus influence the general self-view. The self-concept is defined by its important attributes (Alicke,

1985). Researchers have claimed that most people share a common set of moral traits that characterise their moral self-concept and that evoking this set would trigger salience of the holistic moral concept (Aquino & Reed, 2002). Aquino and Reed (2002) outlined nine moral personal characteristics as essential for eliciting moral identity and these traits are assumed to be associated with a larger set of traits. Hence, relying on a set of morally relevant traits instead of employing two traits only might have shown an aftereffect in Study 3 as well as even stronger aftereffect on subsequent donations in Study 4.

Although I selected moral and non-moral traits with comparable average desirability self-rankings for the second stage of Study 3 (Tappin & McKay, 2017), the data showed that participants ranked themselves substantially higher on the set of moral traits. This precluded exploring a potential differential self-serving aftereffect for moral traits. Moreover, interindividual differences in how important morality is to the self and how relevant morality is to the task predict moral behaviours (Bartels, 2008). My experimental design however did not measure to what degree the selected personality traits were considered important or central for participants' self-view (regarding both moral and non-moral traits). For instance, I chose "honesty", however for some individuals "loyalty" might be more central to the moral self-concept. In a similar vein, how forgetful one is might be a less important non-moral dimension of the self-image for some individuals and thus activate a smaller set of related personality traits. While research has shown that self-enhancement is a universal motivation, individuals self-enhance on personality dimensions they perceive as important (Sedikides et al., 2003; Sedikides et al., 2005). Future work should control for the importance of the selected traits and the centrality of morality to the self-concept at the interindividual level when investigating both the flexibility of personality traits to cognitive influences and the potential transfer of manipulated self-rankings to the general self-image.

Furthermore, a potential limitation of Studies 2 to 5 is that they relied on self-assessment while the most robust way to demonstrate the effect of a manipulation is to

measure revealed behaviour (Back & Vazire, 2012). Despite this potential objection I chose to rely on self-assessment as personality research employs self-reports in general (Paulhus & Vazire, 2007) and self-rankings are shown to be a valid predictor of life outcomes (Beck & Jackson, 2022; Heckman & Kautz, 2012; Kolar et al., 1996; Ozer & Benet-Martinez, 2006; Paunonen & Ashton, 2001; Roberts et al., 2007; Vazire & Mehl, 2008; Zell & Lesick, 2022). Moreover, I exposed self-assessment to certain cognitive influences and measured the differences between conditions so potentially overstated or understated self-evaluations should not interfere with our results. Yet, if one can measure the self-serving cognitive effects on personality judgements with overt behaviour, similar to Study 4, that would provide even more decisive evidence for the effectiveness of the manipulation. For example, after enhancing someone's judgement of how honest they are, they can be presented with a game that provides them with a costless opportunity to cheat, thus enabling a more objective measure of any changes in the levels of honesty (please see the "Implications and future directions" subsection below for a discussion).

Implications and future directions

The stability of patterns in judgements and choices as well as the consistency in individuals' personality judgements have important repercussions at many levels, from informing macro policies (e.g., modelling human behaviour in macroeconomic models) to designing successful interventions to aid the achievement of individuals' goals (e.g., how to save more money or become more conscientious). For instance, a central policy goal is to increase positive life outcomes and improve well-being (OECD, 2020). Research has long shown evidence that personality traits can predict life outcomes (e.g., Ozer & Benet-Martinez, 2006); allowing the possibility of flexibility of personality traits throughout the lifespan, however (Ferguson et al., 2010; Roberts et al., 2006), has provided a basis for designing and implementing nonclinical psychological interventions shaping personality traits (Bleidorn et al., 2019). Our research shows novel evidence that personality judgements can adjust instantaneously to cognitive influence provided they elevate the self-view. These findings shed additional light on the psychological

factors influencing the adjustments in personality judgements and could further facilitate the recent surge of research effort in shaping personality traits (Allemand & Flückiger, 2022).

The central finding of this thesis is that personality judgements are susceptible to cognitive biases (anchoring and choice blindness) in a self-serving manner. It would be interesting to replicate Study 2 and Study 5 in a way that explores whether self-serving anchoring always implies higher susceptibility to anchors that point in the direction that strengthens participants' current self-image. For instance, if I recruit a sample of participants suffering from clinical depression, I can test whether the observed self-serving pattern of anchoring is reversed, i.e., personality judgements exhibit flexibility and adjust in response to diminishing manipulations while remaining rigid in view of enhancing manipulations. Indeed, research has shown that people with negative self-views favour negative feedback due to self-consistency motivation (North & Swann, 2009; Swann et al., 1992). In a similar vein, the imposter syndrome (disregarding positive evidence on one's own abilities while accepting negative evidence) is a form of self-deception (Gadsby, 2022). In a collaboration project with Stephan Gadsby, my supervisors and I plan to explore whether individuals high in impostor phenomenon will be especially susceptible to manipulations that accord with the imposter syndrome construct (e.g., making people less intelligent, more hardworking). Such investigations would provide further insight into the most effective way to communicate information during clinical interventions influencing individuals' self-evaluations.

Our findings also shed additional light on the psychological processes shaping both anchoring and choice blindness. Anchoring has been theorised as resulting from various psychological mechanisms with ongoing debates about the existence of a unified mechanism that could explain all the observed anchoring effects (e.g., Harris et al., 2019). The results of Study 2 show that the specific case of anchoring personality judgements might provide more decisive evidence for selective accessibility of anchor-consistent information shaping the anchoring effect. Therefore, further exploring anchoring on personality judgements by conducting studies similar to Study 2 would

also be interesting and important from a theoretical point of view. Moreover, the current experimental paradigm of Study 2 could be enriched by including suitable reflection questions that would provide further insight and additional evidence on the mechanism shaping individuals' self-evaluations.

As discussed above, however, the investigation of potential aftereffects of the self-serving anchoring of personality traits suffered from certain limitations (Study 3). Moreover, in Study 5, I might not have chosen the most appropriate measure to capture the potential aftereffect on the self-image. Nevertheless, the results of Study 4 showed that enhancing anchors led to nearly 15% more generous donations, which seems like promising evidence for a transfer of self-serving anchoring of personality judgements to the general self-image. Therefore, an important future direction for research is to further investigate whether the observed instantaneous adjustments in personality judgements affect the general self-view and persist over task contexts and time. In doing so, a larger set of moral personality traits should be used to trigger individuals' moral identity and interindividual differences in how central morality is to the self-concept should be controlled for (Aquino & Reed, 2002). Previous research has shown that choices induce changes to subsequent preferences (Ariely et al., 2003, 2006; Ariely & Norton, 2008; Enisman et al., 2021; Sharot et al., 2009; Sharot et al., 2012). In addition, both the anchoring effect (Adomavicius et al., 2023; Mussweiler, 2001) and the choice blindness effect (Johansson et al., 2008; Johansson et al., 2014; cf. Taya et al., 2014) persist over time. These findings suggest that a well thought out experimental design, building on the studies in this thesis, might reveal even stronger aftereffect of manipulated personality judgements on the general self-image and subsequent behaviour than the one observed in Study 4.

Another potential extension of our research is to conduct experiments similar to Study 3 and Study 4, i.e., manipulating participants' personality judgements, followed by measuring subsequent revealed behaviour with suitable tasks. For example, I could investigate whether enhancing participants' judgements on how honest they are would transfer to more honest behaviour in a task involving, for example, self-reports of

correctly solved matrices (Mazar et al., 2008) or reporting the outcome of a die throw in the so called probabilistic-cheating paradigm (Fischbacher & Föllmi-Heusi, 2013), which was recently outlined as the most suitable task for measuring honesty (Hilbig, 2022).

Furthermore, an interesting implication of our findings is that anchor values might be perceived as the acceptable social norm with respect to the estimate in question. For example, underreporting alcohol consumption is a well-known issue (Stockwell et al., 2004), in a recent study (Rostekova et al., in prep.), participants were asked to report their alcohol consumption under high, low and no anchor conditions. The results show that participants reported higher consumption in the high anchor condition (Rostekova et al., in prep.). In this context, the reported higher alcohol consumption in the high anchor condition might be closer to the true consumption, which has important implications for prescribing more effective individual treatment.

Another interesting avenue for future research is to disentangle the effect of morality from that of desirability. For instance, a study similar to Study 3 and Study 4 might be conducted, but anchoring half of the participants on morally relevant traits and half on non-morally relevant traits, phrased either in a positive or negative manner (for e.g., “honest” and “dishonest” or “competent” and “incompetent”). In this way one could test whether individuals are more sensitive to negative phrasing of morally or non-morally relevant traits. Again, an important potential caveat is that one needs to contrast morally and non-morally relevant traits with comparable desirability self-rankings, i.e., one could use “competent”/“incompetent” as long as their desirability rankings are comparable to those for “honest”/“dishonest”.

Finally, at a more abstract level, knowing more about the principles of social cognition also has implications for the growing field of AI ethics. Humans’ social environment is increasingly virtual, and our judgements and choices have an additional effect by contributing to the vast dataset used to train AI. The patterns and biases in our behaviours are naturally reflected in the AI’s recommendations. Unsurprisingly, research has shown that AI decisions suffer gender and race biases that affect recommendations and decisions in healthcare and employment (Hall & Ellis, 2023; Peng et al., 2022;

Schwartz et al., 2022). More relevant to the current thesis is the question of how AI chatbots such as ChatGPT perceive themselves and if they could be defined as having consciousness, emotions, and a self-concept. It would be an interesting extension of the current thesis to ask future iterations of ChatGPT to rank itself on different personality traits, in comparison to other AI chatbots or humans. Moreover, I could investigate whether AI chatbots are susceptible to anchoring. In general, anchoring could not be tested on AI chatbots as they have access to the true estimates on trivia questions. Extending the anchoring paradigm to the personality domain, however, opens the intriguing opportunity to test whether AI chatbots' personality judgements could be anchored and whether there would be a self-serving anchoring effect on their personality judgements.

Summary and conclusions

The results of the present thesis have shown that personality judgements are selectively flexible – flexible in a self-serving way. From a theoretical point of view, these findings show that both anchoring, and the choice blindness paradigm can be extended to the domain of the self, revealing that personality judgements are susceptible to cognitive influences and that they are constructed in a way that is somewhat different than previously thought. Personality judgements seem to be flexible enough to accommodate elevating adjustments, but also rigid enough to resist revisions that would diminish the self-view. Our findings also illuminate the psychological mechanisms shaping anchoring and the choice blindness effect, suggesting that a unified process of comparison (utilising different standards for comparisons) might underly both cognitive effects in the domain of the self. Although, the results did not yield a unanimous support for a transfer of the manipulated personality judgements to the general self-image and subsequent personality judgements, the data provided evidence for elevating aftereffect of enhancing anchors on subsequent donations. I observed nearly 15% increase in donations in an incentivised Dictator Game, which is a promising result worthy of further investigation. I also gained important insights about the experimental design needed to show more decisively that the changes in personality traits due to

cognitive influences persist over task contexts and time. Self-rankings on morally relevant traits also revealed a “phrasing effect” with participants ranking themselves higher, on average, for negatively than positively phrased traits. There was no evidence for a magnified anchoring or choice blindness effect for moral traits. Nevertheless, the data showed that morality has a general elevating effect on self-assessments with individuals ranking themselves more positively on moral than on non-moral traits.

References

- Acciarini, C., Brunetta, F., & Boccardelli, P. (2021). Cognitive biases and decision-making strategies in times of change: a systematic literature review. *Management Decision*, 59(3), 638-652. <https://doi.org/10.1108/MD-07-2019-1006>
- Aquino, K., & Reed, A. II. (2002). The self-importance of moral identity. *Journal of Personality and Social Psychology*, 83(6), 1423–1440. <https://doi.org/10.1037/0022-3514.83.6.1423>
- Adomavicius, G., Bockstedt, J., Curley, S., & Zhang, J. (2023). Persistence of Recommender-Induced Biases in Consumer Preference Ratings. Available at SSRN 4416597. <https://ssrn.com/abstract=3346686>
- Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, 50(2), 179-211. [https://doi.org/10.1016/0749-5978\(91\)90020-T](https://doi.org/10.1016/0749-5978(91)90020-T)
- Akerlof, G. A., & Kranton, R. E. (2000). Economics and identity. *The Quarterly Journal of Economics*, 115(3), 715-753. <https://doi.org/10.1162/003355300554881>
- Alicke, M. D. (1985). Global self-evaluation as determined by the desirability and controllability of trait adjectives. *Journal of Personality and Social Psychology*, 49(6), 1621–1630. <https://doi.org/10.1037/0022-3514.49.6.1621>
- Alicke, M. D., & Govorun, O. (2005). The better-than-average effect. *The Self in Social Judgment*, 1, 85-106. https://doi.org/10.1007/978-3-319-24612-3_300293
- Alicke, M., & Sedikides, C. (2009). Self-enhancement and self-protection: What they are and what they do. *European Review of Social Psychology* 20, 1-48. Self-Enhancement and Self-Protection: What They Are and What They Do. <https://doi.org/10.1080/10463280802613866>

- Alicke, M. D., Zell, E., & Guenther, C. L. (2013). Social self-analysis: Constructing, protecting, and enhancing the self. *Advances in Experimental Social Psychology*, 48, 173–234. <https://doi.org/10.1016/B978-0-12-407188-9.00004-1>
- Allemand, M., & Flückiger, C. (2022). Personality change through digital-coaching interventions. *Current Directions in Psychological Science*, 31(1), 41-48. <https://doi.org/10.1177/09637214211106778>
- Amir, O., & Rand, D. G. (2012). Economic games on the internet: The effect of \$1 stakes. *PloS one*, 7(2). <https://doi.org/10.1371/journal.pone.0031461>
- Andreoni, J., & Bernheim, B. D. (2009). Social image and the 50–50 norm: A theoretical and experimental analysis of audience effects. *Econometrica*, 77(5), 1607-1636. <https://www.jstor.org/stable/25621371>
- Anusic, I., & Schimmack, U. (2016). Stability and change of personality traits, self-esteem, and well-being: Introducing the meta-analytic stability and change model of retest correlations. *Journal of Personality and Social Psychology*, 110(5), 766. <https://doi.org/10.1037/pspp0000066>
- Anwyl-Irvine, A., Dalmaijer, E. S., Hodges, N., & Evershed, J. K. (2021). Realistic precision and accuracy of online experiment platforms, web browsers, and devices. *Behavior Research Methods*, 53(4), 1407-1425. <https://doi.org/10.3758/s13428-020-01501-5>
- Ariely, D., Loewenstein, G., & Prelec, D. (2003). “Coherent arbitrariness”: Stable demand curves without stable preferences. *The Quarterly Journal of Economics*, 118(1), 73-106. <https://doi.org/10.1162/00335530360535153>.
- Ariely, D., Loewenstein, G., & Prelec, D. (2006). Tom Sawyer and the construction of value. *Journal of Economic Behavior & Organization*, 60(1), 1-10. <https://doi.org/10.1016/j.jebo.2004.10.003>

- Ariely, D., & Norton, M. I. (2008). How actions create—not just reveal—preferences. *Trends in Cognitive Sciences*, 12(1), 13-16.
<https://doi.org/10.1016/j.tics.2007.10.008>.
- Ashton, M. C., & Lee, K. (2020). Objections to the HEXACO model of personality structure—And why those objections fail. *European Journal of Personality*, 34(4), 492-510. <https://doi.org/10.1002/per.2242>
- Back, M. D. (2012). Knowing our personality (S. Vazire, Ed.). In S. Vazire & T. D. Wilson (Eds.), *Handbook of self-knowledge* (pp. 131–156). The Guilford Press.
- Bahník, Š., & Strack, F. (2016). Overlap of accessible information undermines the anchoring effect. *Judgment and Decision Making*, 11(1), 92-98.
<https://doi.org/10.1017/S1930297500007610>
- Bartels, D. M. (2008). Principled moral sentiment and the flexibility of moral judgment and decision making. *Cognition*, 108(2), 381-417.
<https://doi.org/10.1016/j.cognition.2008.03.001>
- Beck, E. D., & Jackson, J. J. (2022). A mega-analysis of personality prediction: Robustness and boundary conditions. *Journal of Personality and Social Psychology*, 122(3), 523– 553. <https://doi.org/10.1037/pspp0000386>
- Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon. com's Mechanical Turk. *Political Analysis*, 20(3), 351-368. <https://doi.org/10.1093/pan/mpr057>
- Blanken, I., van de Ven, N., & Zeelenberg, M. (2015). A meta-analytic review of moral licensing. *Personality and Social Psychology Bulletin*, 41(4), 540–558. <https://doi.org/10.1177/0146167215572134>
- Blanken, I., van de Ven, N., Zeelenberg, M., & Meijers, M. H. C. (2014). Three attempts to replicate the moral licensing effect. *Social Psychology*, 45(3), 232–238.
<https://doi.org/10.1027/1864-9335/a000189>

- Bleidorn, W., Hill, P. L., Back, M. D., Denissen, J. J. A., Hennecke, M., Hopwood, C. J., Jokela, M., Kandler, C., Lucas, R. E., Luhmann, M., Orth, U., Wagner, J., Wrzus, C., Zimmermann, J., & Roberts, B. (2019). The policy relevance of personality traits. *American Psychologist*, 74(9), 1056-0067. <https://doi.org/10.1037/amp0000503>
- Bleidorn, W., Hopwood, C. J., Back, M. D., Denissen, J. J. A., Hennecke, M., Jokela, M., Kandler, C., Lucas, R. E., Luhmann, M., Orth, U., Roberts, B. W., Wagner, J., Wrzus, C., & Zimmermann, J. (2020). Longitudinal Experience-Wide Association Studies—A framework for studying personality change. *European Journal of Personality*, 34(3), 285–300. <https://doi.org/10.1002/per.2247>
- Bleidorn, W., Hopwood, C. J., Back, M. D., Denissen, J. J., Hennecke, M., Hill, P. L., Jokela, M., Kandler, C., Lucas, R. E., Luhmann, M., Orth, U., Roberts, B. W., Wagner, J., Wrzus, C., & Zimmermann, J. (2021). Personality trait stability and change. *Personality Science*, 2, 1–20. <https://doi.org/10.5964/ps.6009>
- Bleidorn, W., Schwaba, T., Zheng, A., Hopwood, C. J., Sosa, S. S., Roberts, B. W., & Briley, D. A. (2022). Personality stability and change: A meta-analysis of longitudinal studies. *Psychological bulletin*, 148(7-8), 588-619. <https://doi.org/10.1037/bul0000365>
- Brown, J. D. (2012). Understanding the better than average effect: Motives (still) matter. *Personality and Social Psychology Bulletin*, 38(2), 209-219. <https://doi.org/10.1177/0146167211432763>
- Buhrmester, M., Kwang, T., Gosling, S. D. (2011). Amazon’s Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data? *Perspectives on Psychological Science*, 6(1), 3-5. <https://doi.org/10.1177/1745691610393980>
- Burnham, M. J., Le, Y. K., & Piedmont, R. L. (2018). Who is Mturk? Personal characteristics and sample consistency of these online workers. *Mental Health*,

Religion and Culture, 21(9–10), 934– 944.

<https://doi.org/10.1080/13674676.2018.1486394>

Byun, T. M., Halpin, P. F., & Szeredi, D. (2015). Online crowdsourcing for efficient rating of speech: A validation study. *Journal of Communication Disorders*, 53, 70-83.

<https://doi.org/10.1016/j.jcomdis.2014.11.003>

Camerer, C. F. (2003). Behavioural studies of strategic thinking in games. *Trends in Cognitive Sciences*, 7(5), 225-231. [https://doi.org/10.1016/S1364-](https://doi.org/10.1016/S1364-6613(03)00094-9)

[6613\(03\)00094-9](https://doi.org/10.1016/S1364-6613(03)00094-9)

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

<https://doi.org/10.1038/s41467-020-15602-4>

Cascio, J., & Plant, E. A. (2015). Prospective moral licensing: Does anticipating doing good later allow you to be bad now? *Journal of Experimental Social Psychology*, 56, 110-116. <https://doi.org/10.1016/j.jesp.2014.09.009>

Casler, K., Bickel, L., & Hackett, E. (2013). Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Computers in Human Behavior*, 29(6), 2156-2160.

<https://doi.org/10.1016/j.chb.2013.05.009>

Chambers, C. D., Feredoes, E., Muthukumaraswamy, S. D., & Etchells, P. (2014). Instead of "playing the game" it is time to change the rules: Registered Reports at AIMS Neuroscience and beyond. *AIMS Neuroscience*, 1(1), 4-17.

<http://doi.org/10.3934/Neuroscience.2014.1.4>

Chandler, J., Mueller, P., & Paolacci, G. (2014). Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods*, 46(1), 112-130. <https://doi.org/10.3758/s13428-013-0365-7>

- Chapman, B. P., Hampson, S., & Clarkin, J. (2014). Personality-informed interventions for healthy aging: conclusions from a National Institute on Aging work group. *Developmental psychology, 50*(5), 1426.
<https://doi.org/10.1037/a0034135>
- Chater, N., & Loewenstein, G. (2016). The under-appreciated drive for sense-making. *Journal of Economic Behavior & Organization, 126*, 137-154.
<https://doi.org/10.1016/j.jebo.2015.10.016>
- Cheek, N. N., Coe-Odess, S., & Schwartz, B. (2015). What have I just done? Anchoring, self-knowledge, and judgements of recent behavior. *Judgment and Decision Making, 10*, 76–85.
<https://EconPapers.repec.org/RePEc:jdm:journl:v:10:y:2015:i:1:p:76-85>
- Clifford, S., Jewell, R. M., & Waggoner, P. D. (2015). Are samples drawn from Mechanical Turk valid for research on political ideology?. *Research & Politics, 2*(4), 2053168015622072. <https://doi.org/10.1177/2053168015622072>
- Costa Jr, P. T., & McCrae, R. R. (1992). Four ways five factors are basic. *Personality and Individual Differences, 13*(6), 653-665. [https://doi.org/10.1016/0191-8869\(92\)90236-I](https://doi.org/10.1016/0191-8869(92)90236-I)
- Crusius, J., Corcoran, K., & Mussweiler, T. (2022). Social Comparison: Theory, Research, and Applications. *Theories in social psychology, 165*. <https://www.persistent-identifier.nl/urn:nbn:nl:ui:12-78d48390-505e-4e6d-a159-a045e4b9cc88>
- Curry, O. S., Chesters, M. J., & Van Lissa, C. J. (2019). Mapping morality with a compass: Testing the theory of ‘morality-as-cooperation’ with a new questionnaire. *Journal of Research in Personality, 78*, 106-124.
<https://doi.org/10.1016/j.jrp.2018.10.008>
- Curry, O. S., Mullins, D. A., & Whitehouse, H. (2019). Is it good to cooperate? Testing the theory of morality-as-cooperation in 60 societies. *Current anthropology, 60*(1), 47-69. <https://doi.org/10.1086/701478>

- Cushman, F., & Young, L. (2011). Patterns of moral judgment derive from nonmoral psychological representations. *Cognitive Science*, 35, 1052–1075.
<https://doi.org/10.1111/j.1551-6709.2010.01167.x>
- Ditto, P. H., Pizarro, D. A., & Tannenbaum, D. (2009). Motivated moral reasoning. In B. H. Ross (Ed.), *Psychology of learning and motivation* (Vol. 50, pp. 307–338). Academic Press. [https://doi.org/10.1016/S0079-7421\(08\)00410-6](https://doi.org/10.1016/S0079-7421(08)00410-6)
- Doğruyol, B., Alper, S., & Yilmaz, O. (2019). The five-factor model of the moral foundations theory is stable across WEIRD and non-WEIRD cultures. *Personality and Individual Differences*, 151, 109547.
<https://doi.org/10.1016/j.paid.2019.109547>
- Dunning, D., & Hayes, A. F. (1996). Evidence for egocentric comparison in social judgment. *Journal of Personality and Social Psychology*, 71(2), 213–229.
<https://doi.org/10.1037/0022-3514.71.2.213>
- Dweck, C. S. (2017). From needs to goals and representations: Foundations for a unified theory of motivation, personality, and development. *Psychological Review*, 124(6), 689–719. <https://doi.org/10.1037/rev0000082>
- Eil, D., & Rao, J. M. (2011). The good news-bad news effect: asymmetric processing of objective information about yourself. *American Economic Journal: Microeconomics*, 3(2), 114–38. <https://doi.org/10.1257/mic.3.2.114>
- Enisman, M., Shpitzer, H., & Kleiman, T. (2021). Choice changes preferences, not merely reflects them: A meta-analysis of the artifact-free free-choice paradigm. *Journal of Personality and Social Psychology*, 120(1), 16–29. <https://doi.org/10.1037/pspa0000263>
- Ellemers, N., & van Nunspeet, F. (2020). Neuroscience and the social origins of moral behavior: How neural underpinnings of social categorization and conformity affect

- everyday moral and immoral behavior. *Current Directions in Psychological Science*, 29(5), 513–520. <https://doi.org/10.1177/0963721420951584>
- Ellemers, N., Pagliaro, S., & Barreto, M. (2013). Morality and behavioural regulation in groups: A social identity approach. *European Review of Social Psychology*, 24(1), 160-193. <https://doi.org/10.1080/10463283.2013.841490>
- Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, 114(3), 817-868. <https://doi.org/10.3758/s13428-021-01694-3>
- Festinger, L. (1954). A theory of social comparison processes. *Human Relations*, 7(2), 117-140. <https://doi.org/10.1177/001872675400700202>
- Ferguson, C. J. (2010). A meta-analysis of normal and disordered personality across the life span. *Journal of Personality and Social Psychology*, 98(4), 659–667. <https://doi.org/10.1037/a0018770>
- Fischbacher, U., & Föllmi-Heusi, F. (2013). Lies in disguise—an experimental study on cheating. *Journal of the European Economic Association*, 11(3), 525-547. <https://doi.org/10.1111/jeea.12014>
- Fleischmann, A., Lammers, J., Diel, K., Hofmann, W., & Galinsky, A. D. (2021). More threatening and more diagnostic: How moral comparisons differ from social comparisons. *Journal of Personality and Social Psychology*, 121(5), 1057–1078. <https://doi.org/10.1037/pspi0000361>
- Forsythe, R., Horowitz, J. L., Savin, N. E., & Sefton, M. (1994). Fairness in simple bargaining experiments. *Games and Economic Behavior*, 6(3), 347-369. <https://doi.org/10.1006/game.1994.1021>
- Frederick, S. W., & Mochon, D. (2012). A scale distortion theory of anchoring. *Journal of Experimental Psychology: General*, 141(1), 124-133. <https://doi.org/10.1037/a0024006>

- Gadsby, S. (2022). Imposter syndrome and self-deception. *Australasian Journal of Philosophy*, 100(2), 247-261. <https://doi.org/10.1080/00048402.2021.1874445>
- Gadsby, S., & Hohwy, J. (2022). Incentivising accuracy reduces bias in the imposter phenomenon. *Current Psychology*, 1-9. <https://doi.org/10.1007/S12144-022-03878-2>
- Gaertner, L., Sedikides, C., & Cai, H. (2012). Wanting to be great and better but not average: On the pancultural desire for self-enhancing and self-improving feedback. *Journal of CrossCultural Psychology*, 43, 521–526. <https://doi.org/10.1177/0022022112438399>
- Gamba, A., Manzoni, E., & Stanca, L. (2017). Social comparison and risk taking behavior. *Theory and Decision*, 82, 221-248. <https://dx.doi.org/10.1007/s11238-016-9562-z>
- Gazzaniga, M. S. (2000). Cerebral specialization and interhemispheric communication: Does the corpus callosum enable the human condition?. *Brain*, 123(7), 1293-1326. <https://doi.org/10.1093/brain/123.7.1293>
- Gilbert, D. T., Giesler, R. B., & Morris, K. A. (1995). When comparisons arise. *Journal of Personality and Social Psychology*, 69(2), 227–236. <https://doi.org/10.1037/0022-3514.69.2.227>
- Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2013). Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in experimental social psychology* (Vol. 47, pp. 55-130). Academic Press. <https://doi.org/10.1016/B978-0-12-407236-7.00002-4>
- Graham, J., Haidt, J., Motyl, M., Meindl, P., Iskiwitsch, C., & Mooijman, M. (2018). Moral foundations theory: On the advantages of moral pluralism over moral monism. In K. Gray & J. Graham (Eds.), *Atlas of moral psychology* (pp. 211–222). The Guilford Press. <https://doi.org/10.1016/B978-0-12-407236-7.00002-4>

- Greene, J., & Haidt, J. (2002). How (and where) does moral judgment work?. *Trends in Cognitive Sciences*, 6(12), 517-523. [https://doi.org/10.1016/S1364-6613\(02\)02011-9](https://doi.org/10.1016/S1364-6613(02)02011-9)
- Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRWE investigation of emotional engagement in moral judgment. *Science*, 293(5537), 2105-2108. <https://www.science.org/doi/10.1126/science.1062872>
- Goodwin, G. P., Piazza, J., & Rozin, P. (2014). Moral character predominates in person perception and evaluation. *Journal of Personality and Social Psychology*, 106(1), 148–168. <https://doi.org/10.1037/a0034726>
- Grimmer, M., & Miles, M. P. (2017). With the best of intentions: a large sample test of the intention-behaviour gap in pro-environmental consumer behaviour. *International Journal of Consumer Studies*, 41(1), 2-10. <https://doi.org/10.1111/ijcs.12290>
- Gosling, S. D., Vazire, S., Srivastava, S., & John, O. P. (2004). Should I trust web-based studies? A comparative analysis of six preconceptions about internet questionnaires. *American Psychologist*, 59(2), 93. <https://doi.org/10.1037/0003-066X.59.2.93>
- 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>
- Haidt, J. (2007). The new synthesis in moral psychology. *Science*, 316(5827), 998-1002. <https://www.science.org/doi/10.1126/science.1137651>
- Hall, P., & Ellis, D. (2023). A systematic review of socio-technical gender bias in AI algorithms. *Online Information Review*. <https://publons.com/publon/10.1108/OIR-08-2021-0452>

- Hall, L., Johansson, P., & Strandberg, T. (2012). Lifting the veil of morality: Choice blindness and attitude reversals on a self-transforming survey. *PloS one*, 7(9), e45457. <https://doi.org/10.1371/journal.pone.0045457>
- Hall, L., Johansson, P., Tärning, B., Sikström, S., & Deutgen, T. (2010). Magic at the marketplace: Choice blindness for the taste of jam and the smell of tea. *Cognition*, 117(1), 54-61. <https://doi.org/10.1016/j.cognition.2010.06.010>
- Hall, L., Strandberg, T., Pärnamets, P., Lind, A., Tärning, B., & Johansson, P. (2013). How the polls can be both spot on and dead wrong: Using choice blindness to shift political attitudes and voter intentions. *PloS one*, 8(4), e60554. <https://doi.org/10.1371/journal.pone.0060554>
- Harris, A. J., Blower, F. B., Rodgers, S. A., Lagator, S., Page, E., Burton, A., Urlichich, D. & Speekenbrink*, M. (2019). Failures to replicate a key result of the selective accessibility theory of anchoring. *Journal of Experimental Psychology: General*, 148(9), e30-e50. <https://doi.org/10.1037/xge0000644>
- Heckman, J. J., & Kautz, T. (2012). Hard evidence on soft skills. *Labour economics*, 19(4), 451-464. <https://doi.org/10.1016/j.labeco.2012.05.014>
- Heintz, C., Karabegovic, M., & Molnar, A. (2016). The co-evolution of honesty and strategic vigilance. *Frontiers in Psychology*, 7, 1503. <https://doi.org/10.3389/fpsyg.2016.01503>
- Heiphetz, L., Strohminger, N., & Young, L. L. (2017). The role of moral beliefs, memories, and preferences in representations of identity. *Cognitive science*, 41(3), 744-767. <https://doi.org/10.1111/cogs.12354>
- Hergueux, J., & Jacquemet, N. (2015). Social preferences in the online laboratory: a randomized experiment. *Experimental Economics*, 18(2), 251-283. <https://doi.org/10.1007/s10683-014-9400-5>

- Hilbig, B. E. (2022). Personality and behavioral dishonesty. *Current Opinion in Psychology*, 101378. <https://doi.org/10.1016/j.copsyc.2022.101378>
- Hilbig, B. E. (2016). Reaction time effects in lab-versus Web-based research: Experimental evidence. *Behavior Research Methods*, 48(4), 1718-1724. <https://doi.org/10.3758/s13428-015-0678-9>
- Hoffman, E., McCabe, K., & Smith, V. L. (1996). Social distance and other-regarding behavior in dictator games. *The American economic review*, 86(3), 653-660. <https://www.jstor.org/stable/2118218>
- Isler, O., & Gächter, S. (2022). Conforming with peers in honesty and cooperation. *Journal of Economic Behavior & Organization*, 195, 75-86. <http://dx.doi.org/10.2139/ssrn.4114486>
- Iyer, R., Koleva, S., Graham, J., Ditto, P. H., & Haidt, J. (2012). Understanding libertarian morality: The psychological roots of an individualist ideology. *PLoS One*, 7(8), e42366. <https://doi.org/10.1371/journal.pone.0042366>
- Joel, S., Spielmann, S. S., & MacDonald, G. (2017). Motivated use of numerical anchors for judgements relevant to the self. *Personality and Social Psychology Bulletin*, 43(7), 972-985. <https://doi.org/10.1177/0146167217702613>
- Johansson, P., Hall, L., & Chater, N. (2012). Preference change through choice. In *Neuroscience of preference and choice* (pp. 121-141). Academic Press. <https://doi.org/10.1016/B978-0-12-381431-9.00006-1>
- Johansson, P., Hall, L., Tärning, B., Sikström, S., & Chater, N. (2014). Choice blindness and preference change: You will like this paper better if you (believe you) chose to read it! *Journal of Behavioral Decision Making*, 27(3), 281–289. <https://doi.org/10.1002/bdm.1807>
- Johansson, P., Hall, L., & Sikström, S. (2008). From change blindness to choice blindness. *Psychologia: An International Journal of Psychology in the Orient*, 51(2), 142–155. <https://doi.org/10.2117/psysoc.2008.142>

- Johansson, P., Hall, L., Sikstrom, S., & Olsson, A. (2005). Failure to detect mismatches between intention and outcome in a simple decision task. *Science*, 310(5745), 116-119. <https://doi.org/10.1126/science.1111709> PMID: 16210542
- Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1986). Fairness and the assumptions of economics. *Journal of Business*, S285-S300. <https://doi.org/10.1086/296367>
- Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47, 263-291. <https://doi.org/10.2307/1914185>
- Kahneman, D. (1992). Reference points, anchors, norms, and mixed feelings. *Organizational Behavior and Human Decision Processes*, 51(2), 296–312. [https://doi.org/10.1016/0749-5978\(92\)90015-Y](https://doi.org/10.1016/0749-5978(92)90015-Y)
- Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, 2(3), 196-217. https://doi.org/10.1207/s15327957pspr0203_4
- Keuschnigg, M., Bader, F., & Bracher, J. (2016). Using crowdsourced online experiments to study context-dependency of behavior. *Social Science Research*, 59, 68-82. <https://doi.org/10.1016/j.ssresearch.2016.04.014>
- Kimbrough, E. O., & Vostroknutov, A. (2016). Norms make preferences social. *Journal of the European Economic Association*, 14(3), 608-638. <https://doi.org/10.1111/jeea.12152>
- Kirk, R. E. (2012). *Experimental design: Procedures for the behavioral sciences*. Sage Publications. <https://dx.doi.org/10.4135/9781483384733>
- Kolar, D.W., Funder, D.C., & Colvin, C.R. (1996). Comparing the accuracy of personality judgements by the self and knowledgeable others. *Journal of Personality*, 64, 311-337. <https://doi.org/10.1111/j.1467-6494.1996.tb00513.x>

- Korn, C. W., Prehn, K., Park, S. Q., Walter, H., & Heekeren, H. R. (2012). Positively biased processing of self-relevant social feedback. *The Journal of Neuroscience*, 32(47), 16832–16844. <https://doi.org/10.1523/JNEUROSCI.3016-12.2012>
- Krupka, E. L., & Weber, R. A. (2013). Identifying social norms using coordination games: Why does dictator game sharing vary? *Journal of the European Economic Association*, 11(3), 495-524. <https://doi.org/10.1111/jeea.12006>
- Landau, M. J., Vess, M., Arndt, J., Rothschild, Z. K., Sullivan, D., & Atchley, R. A. (2011). Embodied metaphor and the “true” self: Priming entity expansion and protection influences intrinsic self-expressions in self-perceptions and interpersonal behavior. *Journal of Experimental Social Psychology*, 47(1), 79–87. <https://doi.org/10.1016/j.jesp.2010.08.012>
- Larney, A., Rotella, A., & Barclay, P. (2019). Stake size effects in ultimatum game and dictator game offers: A meta-analysis. *Organizational Behavior and Human Decision Processes*, 151, 61-72. <https://doi.org/10.1016/j.obhdp.2019.01.002>
- Leary, M. R. (2007). Motivational and emotional aspects of the self. *Annual Review of Psychology*, 58, 317–344. <https://doi.org/10.1146/annurev.psych.58.110405.085658>
- Lefebvre, G., Lebreton, M., Meyniel, F., Bourgeois-Gironde, S., & Palminteri, S. (2017). Behavioural and neural characterization of optimistic reinforcement learning. *Nature Human Behaviour*, 1(4), 1-9. <https://doi.org/10.1038/s41562-017-0067>
- Lindsay, D. S. (2017). Sharing data and materials in psychological science. *Psychological Science Editorial*. <https://doi.org/10.1177/0956797617704015>
- List, J. A. (2007). On the interpretation of giving in dictator games. *Journal of Political Economy*, 115(3), 482-493. <https://www.jstor.org/stable/10.1086/519249>

- Logg, J. M., & Dorison, C. A. (2021). Pre-registration: Weighing costs and benefits for researchers. *Organizational Behavior and Human Decision Processes*, 167, 18-27. <https://doi.org/10.1016/j.obhdp.2021.05.006>
- Malle, B. F. (2006). The actor-observer asymmetry in attribution: A (surprising) meta-analysis. *Psychological Bulletin*, 132(6), 895-919. <https://doi.org/10.1037/0033-2909.132.6.895>
- Manago, B., Mize, T. D., & Doan, L. (2021). Can You Really Study an Army on the Internet? Comparing How Status Tasks Perform in the Laboratory and Online Settings. *Sociological Methodology*, 51(2), 319-347. <https://doi.org/10.1177/00811750211014242>
- Mazar, N., Amir, O., & Ariely, D. (2008). The dishonesty of honest people: A theory of self-concept maintenance. *Journal of Marketing Research*, 45(6), 633–644. <https://doi.org/10.1509/jmkr.45.6.633>
- McLaughlin, O., & Somerville, J. (2013). Choice blindness in financial decision making. *Judgment and Decision Making*, 8(5), 577. <https://EconPapers.repec.org/RePEc:jdm:journl:v:8:y:2013:i:5:p:577-588>
- Meyers, E. A., Białek, M., Fugelsang, J. A., Koehler, D. J., & Friedman, O. (2019). Wronging past rights: The sunk cost bias distorts moral judgment. *Judgment and Decision Making*, 14(6), 721-727. <https://EconPapers.repec.org/RePEc:jdm:journl:v:15:y:2020:i:6:p:909-925>
- Midanik, L. (1982). The validity of self-reported alcohol consumption and alcohol problems: A literature review. *British Journal of Addiction*, 77(4), 357–382. <https://doi.org/10.1111/j.1360-0443.1982.tb02469.x>
- Molouki, S., & Bartels, D. M. (2017). Personal change and the continuity of the self. *Cognitive Psychology*, 93, 1-17. <https://doi.org/10.1016/j.cogpsych.2016.11.006>

- Möbius, M. M., Niederle, M., Niehaus, P., & Rosenblat, T. S. (2022). Managing self-confidence: Theory and experimental evidence. *Management Science* 68(11). <https://doi.org/10.1287/mnsc.2021.4294>
- Morse, S., & Gergen, K. J. (1970). Social comparison, self-consistency, and the concept of self. *Journal of Personality and Social Psychology*, 16(1), 148–156. <https://doi.org/10.1037/h0029862>
- Mroczek, D. K. (2014). Personality plasticity, healthy aging, and interventions. *Developmental Psychology*, 50(5), 1470–1474. <https://doi.org/10.1037/a0036028>
- Mullen, E., & Monin, B. (2016). Consistency versus licensing effects of past moral behavior. *Annual Review of Psychology*, 67, 363–385. <https://doi.org/10.1146/annurev-psych-010213-115120>
- Mussweiler, T. (2003). Comparison processes in social judgment: mechanisms and consequences. *Psychological review*, 110(3), 472. <https://doi.org/10.1037/0033-295X.110.3.472>
- Mussweiler, T. (2001). The durability of anchoring effects. *European Journal of Social Psychology*, 31, 431–442. <https://doi.org/10.1002/ejsp.52>
- Mussweiler, T., & Strack, F. (1999a). Hypothesis-consistent testing and semantic priming in the anchoring paradigm: A selective accessibility model. *Journal of Experimental Social Psychology*, 35(2), 136–164. <https://doi.org/10.1006/jesp.1998.1364>
- Mussweiler, T., & Strack, F. (1999b). Comparing is believing: A selective accessibility model of judgmental anchoring. *European review of social psychology*, 10(1), 135–167. <https://doi.org/10.1080/14792779943000044>
- Mussweiler, T., & Strack, F. (2000a). The use of category and exemplar knowledge in the solution of anchoring tasks. *Journal of Personality and Social Psychology*, 78, 1038–1052. <https://doi.org/10.1037/0022-3514.78.6.1038>

- Mussweiler, T., & Strack, F. (2000b). Numeric judgements under uncertainty: The role of knowledge in anchoring. *Journal of Experimental Social Psychology*, 36(5), 495-518. <https://doi.org/10.1006/jesp.1999.1414>
- Mussweiler, T., & Strack, F. (2001). The semantics of anchoring. *Organizational Behavior and Human Decision Processes*, 86(2), 234-255. <https://doi.org/10.1006/obhd.2001.2954>
- Newman, G. E., Bloom, P., & Knobe, J. (2014). Value judgements and the true self. *Personality and Social Psychology Bulletin*, 40(2), 203–216. <https://doi.org/10.1177/0146167213508791>
- Newman, G. E., De Freitas, J., & Knobe, J. (2015). Beliefs about the true self explain asymmetries based on moral judgment. *Cognitive science*, 39(1), 96-125. <https://doi.org/10.1111/cogs.12134>
- Nickerson, D. W., & Rogers, T. (2010). Do you have a voting plan? Implementation intentions, voter turnout, and organic plan making. *Psychological Science*, 21(2), 194-199. <https://doi.org/10.1177/0956797609359326>
- Nosek, B. A., Beck, E. D., Campbell, L., Flake, J. K., Hardwicke, T. E., Mellor, D. T., ... & Vazire, S. (2019). Preregistration is hard, and worthwhile. *Trends in Cognitive Sciences*, 23(10), 815-818. <https://doi.org/10.1016/j.tics.2019.07.009>
- Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11), 2600-2606. <https://doi.org/10.1073/pnas.1708274114>
- Nosek, B. A., & Lakens, D. (2014). A method to increase the credibility of published results. *Social Psychology*, 45(3), 137-141. <https://doi.org/10.1027/1864-9335/a000192>

- North, R. J., & Swann, W. B., Jr. (2009). Self-verification 360°: Illuminating the light and dark sides. *Self and Identity*, 8(2-3), 131–146. <https://doi.org/10.1080/15298860802501516>
- OECD (2020), *How's Life? 2020: Measuring Well-being*, OECD Publishing, Paris, <https://doi.org/10.1787/9870c393-en>.
- Oltmanns, J. R., Jackson, J. J., & Oltmanns, T. F. (2020). Personality change: Longitudinal self-other agreement and convergence with retrospective-reports. *Journal of Personality and Social Psychology*, 118(5), 1065. <https://doi.org/10.1037/pspp0000238>
- Olaru, G., Stieger, M., Rügger, D., Kowatsch, T., Flückiger, C., Roberts, B. W., & Allemand, M. (2022). Personality change through a digital-coaching intervention: Using measurement invariance testing to distinguish between trait domain, facet, and nuance change. *European Journal of Personality*. <https://doi.org/10.1177/089020702211450>
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251). <https://doi.org/10.1126/science.aac4716>
- Ozer, D. J., & Benet-Martinez, V. (2006). Personality and the prediction of consequential outcomes. *Annual Review of Psychology*, 57, 401-421. <https://doi.org/10.1146/annurev.psych.57.102904.190127>
- Palan, S., & Schitter, C. (2018). Prolific. ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17, 22-27. <https://doi.org/10.1016/j.jbef.2017.12.004>
- Paunonen, S. V., & Ashton, M. C. (2001). Big Five factors and facets and the prediction of behavior. *Journal of Personality and Social Psychology*, 81, 524–539. <https://doi.org/10.1037/0022-3514.81.3.524>

- Paulhus, D. L. (1998). Interpersonal and intrapsychic adaptiveness of trait self-enhancement: A mixed blessing? *Journal of Personality and Social Psychology*, 74(5), 1197–1208. <https://doi.org/10.1037/0022-3514.74.5.1197>
- Paulhus, D. L., & Vazire, S. (2007). The self-report method. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 224–239). The Guilford Press.
- Payne, J. W., Bettman, J. R., & Johnson, E. J. (1992). Behavioral decision research: A constructive processing perspective. *Annual Review of Psychology*, 43(1), 87-131. <https://doi.org/10.1146/annurev.ps.43.020192.000511>
- Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, 70, 153-163. <https://doi.org/10.1016/j.jesp.2017.01.006>
- Peer, E., Rothchild, D., Gordon A., Everdeen, Z., & Damer, E. (2021). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods*, 1-20. <https://doi.org/10.3758/s13428-021-01694-3>
- Peng, A., Nushi, B., Kiciman, E., Inkpen, K., & Kamar, E. (2022, June). Investigations of performance and bias in human-AI teamwork in hiring. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 36, No. 11, pp. 12089-12097). <https://doi.org/10.48550/arXiv.2202.11812>
- Rai, T. S., & Holyoak, K. J. (2010). Moral principles or consumer preferences? Alternative framings of the trolley problem. *Cognitive Science*, 34, 311–321. <https://doi.org/10.1111/j.1551-6709.2009.01088.x>
- Rathbone, C. J., Moulin, C. J., & Conway, M. A. (2008). Self-centered memories: The reminiscence bump and the self. *Memory & Cognition*, 36(8), 1403-1414. <https://doi.org/10.3758/MC.36.8.1403>

- Rhodes, R. E., & Dickau, L. (2012). Experimental evidence for the intention–behavior relationship in the physical activity domain: A meta-analysis. *Health Psychology, 31*(6), 724–727. <https://doi.org/10.1037/a0027290>
- Ritchie, T. D., Sedikides, C., & Skowronski, J. J. (2017). Does a person selectively recall the good or the bad from their personal past? It depends on the recall target and the person’s favourability of self-views. *Memory, 25*(8), 934-944. <https://doi.org/10.1080/09658211.2016.1233984>
- Roberts, B. W. (2009). Back to the future: Personality and assessment and personality development. *Journal of research in personality, 43*(2), 137-145. <https://doi.org/10.1016/j.jrp.2008.12.015>
- Roberts, B. W., & DelVecchio, W. F. (2000). The rank-order consistency of personality traits from childhood to old age: A quantitative review of longitudinal studies. *Psychological Bulletin, 126*(1), 3–25. <https://doi.org/10.1037/0033-2909.126.1.3>
- Roberts, B. W., Kuncel, N., Shiner, R. N., Caspi, A., & Goldberg, L. R. (2007). The power of personality: The comparative validity of personality traits, socio-economic status, and cognitive ability for predicting important life outcomes. *Perspectives in Psychological Science, 2*, 313–345. <https://doi.org/10.1111/j.1745-6916.2007.00047.x>
- Roberts, B. W., Walton, K. E., & Viechtbauer, W. (2006). Patterns of meanlevel change in personality traits across the life course: A meta-analysis of longitudinal studies. *Psychological Bulletin, 132*(1), 1–25. <https://doi.org/10.1037/0033-2909.132.1.1>
- Roberts, B. W., & Yoon, H. J. (2022). Personality psychology. *Annual Review of Psychology, 73*(1), 489–516. <https://doi.org/10.1146/annurevpsych-020821-114927>
- Röseler, L., & Schütz, A. (2022, March 9). Hanging the Anchor Off a New Ship: A Meta-Analysis of Anchoring Effects. PsyArXiv. <https://doi.org/10.31234/osf.io/wf2tn>

- Rostekova, A., Pearson, J., Allen, E., Ambrus, E., Hartig, B. & McKay, R. (in prep). Using the Anchoring Effect to Boost Self-reported Alcohol Consumption.
- Rozin, P. (1999). The process of moralization. *Psychological science*, 10(3), 218-221.
<https://doi.org/10.1111/1467-9280.0013>
- Sagana, A., Sauerland, M., & Merckelbach, H. (2016). The effect of choice reversals on blindness for identification decisions. *Psychology, Crime & Law*, 22(4), 303-314.
<https://doi.org/10.1080/1068316X.2015.1085984>
- Sagiv, L., & Roccas, S. (2021). How do values affect behavior? Let me count the ways. *Personality and Social Psychology Review*, 25(4), 295-316.
<https://doi.org/10.1177/10888683211015975>
- Sagiv, L., Roccas, S., Cieciuch, J., & Schwartz, S. H. (2017). Personal values in human life. *Nature human behaviour*, 1(9), 630-639.
<https://doi.org/10.1177/10888683211015975>
- Schlegel, R. J., Hicks, J. A., Davis, W. E., Hirsch, K. A., & Smith, C. M. (2013). The dynamic interplay between perceived true self-knowledge and decision satisfaction. *Journal of Personality and Social Psychology*, 104(3), 542–558. <https://doi.org/10.1037/a0031183>
- Schwartz, S. H., & Bardi, A. (2001). Value hierarchies across cultures: Taking a similarities perspective. *Journal of Cross-Cultural Psychology*, 32(3), 268–290. <https://doi.org/10.1177/0022022101032003002>
- Schwartz, S. H., & Cieciuch, J. (2022). Measuring the refined theory of individual values in 49 cultural groups: Psychometrics of the Revised Portrait Value Questionnaire. *Assessment*, 29(5), 1005–1019. <https://doi.org/10.1177/1073191121998760>

- Schwerter, F. (2023). Social reference points and risk taking. *Management Science*.
<https://doi.org/10.1287/mnsc.2023.4698>
- Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A., & Hall, P. (2022). Towards a standard for identifying and managing bias in artificial intelligence. *NIST Special Publication*, 1270, 1-77. <https://doi.org/10.6028/NIST.SP.1270>
- Sedikides, C., Gaertner, L., & Toguchi, Y. (2003). Pancultural self-enhancement. *Journal of Personality and Social Psychology*, 84(1), 60–79.
<https://doi.org/10.1037/0022-3514.84.1.60>
- Sedikides, C., Gaertner, L., & Vevea, J. L. (2005). Pancultural self-enhancement reloaded: A meta-analytic reply to Heine (2005). *Journal of Personality and Social Psychology*, 89(4), 539–551. <https://doi.org/10.1037/0022-3514.89.4.539>
- Sedikides, C., & Green, J. D. (2009). Memory as a self-protective mechanism. *Social and Personality Psychology Compass*, 3(6), 1055–1068.
<https://doi.org/10.1111/j.1751-9004.2009.00220.x>
- Sedikides, C., & Strube, M. J. (1997). Self-evaluation: To thine own self be good, to thine own self be sure, to thine own self be true, and to thine own self be better. In *Advances in experimental social psychology* (Vol. 29, pp. 209-269). Academic Press. [https://doi.org/10.1016/S0065-2601\(08\)60018-0](https://doi.org/10.1016/S0065-2601(08)60018-0)
- Sedikides, C., Green, J. D., Saunders, J., Skowronski, J. J., & Zengel, B. (2016). Mnemic neglect: Selective amnesia of one's faults. *European Review of Social Psychology*, 27, 1–62. <https://doi.org/10.1080/10463283.2016.1183913>
- Sharot, T., De Martino, B., & Dolan, R. J. (2009). How choice reveals and shapes expected hedonic outcome. *The Journal of Neuroscience*, 29(12), 3760–3765. <https://doi.org/10.1523/JNEUROSCI.4972-08.2009>
- Sharot, T., Fleming, S. M., Yu, X., Koster, R., & Dolan, R. J. (2012). Is choice-induced preference change long lasting? *Psychological Science*, 23(10), 1123–1129. <https://doi.org/10.1177/0956797612438733>

- Soraperra, I., Weisel, O., & Ploner, M. (2019). Is the victim Max (Planck) or Moritz? How victim type and social value orientation affect dishonest behavior. *Journal of Behavioral Decision Making*, 32(2), 168-178. <https://doi.org/10.1002/bdm.2104>
- Soto, C. J. (2021). Do links between personality and life outcomes generalize? Testing the robustness of trait–outcome associations across gender, age, ethnicity, and analytic approaches. *Social Psychological and Personality Science*, 12(1), 118-130. <https://doi.org/10.1177/1948550619900572>
- Stanley, M. L., Henne, P., & De Brigard, F. (2019). Remembering moral and immoral actions in constructing the self. *Memory & Cognition*, 47(3), 441-454. <https://doi.org/10.3758/s13421-018-0880-y>
- Stewart, N., Chandler, J., & Paolacci, G. (2017). Crowdsourcing samples in cognitive science. *Trends in Cognitive Sciences*, 21(10), 736-748. <https://doi.org/10.1016/j.tics.2017.06.007>
- Stieger, M., Wepfer, S., Rügger, D., Kowatsch, T., Roberts, B. W., & Allemand, M. (2020). Becoming more conscientious or more open to experience? Effects of a two-week smartphone-based intervention for personality change. *European Journal of Personality*, 34(3), 345–366. <https://doi.org/10.1002/per.2267>
- Stockwell, T., Donath, S., Cooper-Stanbury, M., Chikritzhs, T., Catalano, P., & Mateo, C. (2004). Under-reporting of alcohol consumption in household surveys: a comparison of quantity–frequency, graduated–frequency and recent recall. *Addiction*, 99(8), 1024-1033. <https://doi.org/10.1111/j.1360-0443.2004.00815.x>
- Strack, F., & Mussweiler, T. (1997). Explaining the enigmatic anchoring effect: Mechanisms of selective accessibility. *Journal of Personality and Social Psychology*, 73, 437–446. <https://doi.org/10.1037/0022-3514.73.3.437>

- Strandberg, T., Olson, J. A., Hall, L., Woods, A., & Johansson, P. (2020). Depolarizing American voters: Democrats and Republicans are equally susceptible to false attitude feedback. *Plos one*, *15*(2), e0226799.
<https://doi.org/10.1371/journal.pone.0226799>
- Strohminger, N. (2018). Identity is essentially moral. *Atlas of moral psychology*, 141-148.
<https://doi.org/10.1016/j.cognition.2013.12.005>
- Strohminger, N., & Nichols, S. (2014). The essential moral self. *Cognition*, *131*(1), 159–171. <https://doi.org/10.1016/j.cognition.2013.12.005>
- Sulik, J., Ross, R. M., Balzan, R., & McKay, R. (2023). Delusion-like beliefs and data quality: Are classic cognitive biases artifacts of carelessness? *Journal of Psychopathology and Clinical Science*. Advance online publication.
<https://doi.org/10.1037/abn0000844>
- Swann, W. B. (1987). Identity negotiation: Where two roads meet. *Journal of Personality and Social Psychology*, *53*(6), 1038–1051. <https://doi.org/10.1037/0022-3514.53.6.1038>
- Swann, W. B., Wenzlaff, R. M., & Tafarodi, R. W. (1992). Depression and the search for negative evaluations: More evidence of the role of self-verification strivings. *Journal of Abnormal Psychology*, *101*(2), 314–317. <https://doi.org/10.1037/0021-843X.101.2.314>
- Tappin, B. M., & McKay, R. T. (2017). The illusion of moral superiority. *Social Psychological and Personality Science*, *8*(6), 623-631.
<https://doi.org/10.1177/1948550616673878>
- Tappin, B. M., & McKay, R. T. (2019). Investigating the relationship between self-perceived moral superiority and moral behavior using economic games. *Social Psychological and Personality Science*, *10*(2), 135-143.
<https://doi.org/10.1177/1948550617750736>

- Tassy, S., Oullier, O., Mancini, J., & Wicker, B. (2013). Discrepancies between judgment and choice of action in moral dilemmas. *Frontiers in Psychology, 4*, 250.
<https://doi.org/10.3389/fpsyg.2013.00250>
- Taya, F., Gupta, S., Farber, I., & Mullette-Gillman, O. D. A. (2014). Manipulation detection and preference alterations in a choice blindness paradigm. *PloS one, 9*(9), e108515. <https://doi.org/10.1371/journal.pone.0108515>
- Taylor, S. E., & Brown, J. D. (1988). Illusion and well-being: A social psychological perspective on mental health. *Psychological Bulletin, 103*(2), 193–210.
<https://doi.org/10.1037/0033-2909.103.2.193>
- Teovanović, P. (2019). Individual differences in anchoring effect: Evidence for the role of insufficient adjustment. *Europe's Journal of Psychology, 15*(1), 8.
<https://doi.org/10.5964/ejop.v15i1.1691>
- Tesser, A. (1988). Toward a self-evaluation maintenance model of social behavior. In *Advances in experimental social psychology* (Vol. 21, pp. 181-227). Academic Press. [https://doi.org/10.1016/S0065-2601\(08\)60227-0](https://doi.org/10.1016/S0065-2601(08)60227-0)
- Thalmayer, A. G., & Saucier, G. (2014). The questionnaire big six in 26 nations: Developing cross-culturally applicable big six, big five and big two inventories. *European Journal of Personality, 28*(5), 482-496.
<https://doi.org/10.1002/per.2094>
- Thielmann, I., Moshagen, M., Hilbig, B., & Zettler, I. (2022). On the comparability of basic personality models: Meta-analytic correspondence, scope, and orthogonality of the Big Five and HEXACO dimensions. *European Journal of Personality, 36*(6), 870-900. <https://doi.org/10.1177/08902070211026793>
- Thielmann, I., Zimmermann, J., Leising, D., & Hilbig, B. E. (2017). Seeing is knowing: On the predictive accuracy of self- and informant reports for prosocial and moral

- behaviours. *European Journal of Personality*, 31, 404–418.
<http://dx.doi.org/10.1002/per.2112>
- Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157), 1124-1131.
<https://doi.org/10.1126/science.185.4157.1124>
- Van der Leer, L., & McKay, R. (2017). The optimist within? Selective sampling and self-deception. *Consciousness and cognition*, 50, 23-29.
<https://doi.org/10.1016/j.concog.2016.07.005>
- Vazire, S. (2010). Who knows what about a person? The self-other knowledge asymmetry (SOKA) model. *Journal of Personality and Social Psychology*, 98, 281-300. <https://doi.org/10.1037/a0017908>
- Vazire, S., & Mehl, M.R. (2008). Knowing me, knowing you: The accuracy and unique predictive validity of self-ratings and other-ratings of daily behavior. *Journal of Personality and Social Psychology*, 95, 1202-1216.
<https://doi.org/10.1037/a0013314>
- Von Neumann, J. & Morgenstern, O. (1947). *Theory of Games and Economic Behavior*. Princeton University Press. Retrieved from
<http://press.princeton.edu/chapters/i7802.pdf>
- Wagenmakers, E. J., Wetzels, R., Borsboom, D., van der Maas, H. L., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, 7(6), 632-638.
<https://doi.org/10.1177/1745691612463078>
- Wilson, T. D., Houston, C. E., Etling, K. M., & Brekke, N. (1996). A new look at anchoring effects: Basic anchoring and its antecedents. *Journal of Experimental Psychology: General*, 125(4), 387–402. <https://doi.org/10.1037/0096-3445.125.4.387>

- Wong, K. F. E., & Kwong, J. Y. Y. (2000). Is 7300 m equal to 7.3 km? Same semantics but different anchoring effects. *Organizational Behavior and Human Decision Processes*, 82(2), 314-333. <https://doi.org/10.1006/obhd.2000.2900>
- Yoon, S., Fong, N. M., & Dimoka, A. (2019). The robustness of anchoring effects on preferential judgements. *Judgment & Decision Making*, 14(4), 470-487. <https://doi.org/10.1017/S1930297500006148>
- Zell, E., & Lesick, T. L. (2022). Big five personality traits and performance: A quantitative synthesis of 50+ meta-analyses. *Journal of Personality*, 90(4), 559-573. <https://doi.org/10.1111/jopy.12683>
- Zettler, I., Thielmann, I., Hilbig, B. E., & Moshagen, M. (2020). The nomological net of the HEXACO model of personality: A large-scale meta-analytic investigation. *Perspectives on Psychological Science*, 15(3), 723-760. <https://doi.org/10.1177/1745691619895036>
- Zhang, Y., Pan, Z., Li, K., & Guo, Y. (2018). Self-serving bias in memories. *Experimental Psychology*, 65(4), 236–244. <https://doi.org/10.1027/1618-3169/a000409>
- Ziano, I., Mok, P. Y., & Feldman, G. (2021). Replication and extension of Alicke (1985) better-than-average effect for desirable and controllable traits. *Social Psychological and Personality Science*, 12(6), 1005-1017. <https://doi.org/10.1177/1948550620948973>