UNIVERSITÄT AUGSBURG
Universitätsbibliothek

OPEN ACCESS
OPUS AUGSBURG

A model of moral balancing under motivated reasoning

Daniela Flörchinger

Angaben zur Veröffentlichung / Publication details:

Flörchinger, Daniela. 2026. "A model of moral balancing under motivated reasoning." *Journal of Economic Behavior & Organization* 242: 107393. https://doi.org/10.1016/j.jebo.2025.107393.

Nutzungsbedingungen / Terms of use:

CC BY 4.0

Dieses Dokument wird unter folgenden Bedingungen zur Verfügung gestellt: / This document is made available under these conditions:
CC-BY 4.0: Creative Commons: Namensnennung
Weitere Informationen finden Sie unter: / For more information see:
https://creativecommons.org/licenses/by/4.0/deed.de

Journal of Economic Behavior and Organization 242 (2026) 107393

Contents lists available at ScienceDirect

Journal of Economic Behavior and Organization

journal homepage: www.elsevier.com/locate/jebo

A model of moral balancing under motivated reasoning

Daniela Flörchinger

RWI – Leibniz Institute for Economic Research, Hohenzollernstr. 1-3, 45128 Essen, Germany

ARTICLE INFO

JEL classification:
D91
D62

Keywords:
Behavioral economics
Theoretical model
Moral balancing
Motivated reasoning
Self-signaling
Prosocial behavior

ABSTRACT

While there is abundant empirical evidence of individuals switching between selfish and prosocial behavior, few economic models formalize these findings. This paper presents a novel model that jointly analyzes three key concepts for understanding prosocial behavior: moral balancing, self-signaling, and motivated reasoning. Individuals maximize material utility under the constraint of maintaining a minimum level of self-image (moral balancing), where self-image depends on signals extracted from past behavior (self-signaling). The processing of these signals is biased toward arriving at a positive self-image (motivated reasoning). The time horizon for which the self-image constraint is active depends on individuals' intrinsic motivation and their awareness of self-image relevant choices. Selfish behavior tends to be higher when the constraint is only active in the long term and it increases with the tendency toward motivated reasoning. These results are partially consistent with experimental results. The model suggests that voluntary prosocial behavior is unlikely to be sustained and thus calls for adequate policy measures.

1. Introduction

Research in the fields of social psychology and economics provides evidence that individuals tend to display prosocial behavior in certain situations whilst exhibiting selfish conduct in other circumstances. This phenomenon is known as moral balancing (Nisan and Horenczyk, 1990). For example, individuals cheat or lie to obtain a higher pay-off in one round of an experiment, but donate in another round (Gneezy et al., 2014; Ploner and Regner, 2013). Similarly, they may travel by train on one occasion and by plane on another, where traveling by train is the prosocial action because it produces fewer negative externalities.

Although empirically well documented (see Jordan et al., 2011; Merritt et al., 2010, 2012; Monin and Miller, 2001; Sachdeva et al., 2009; Brañas-Garza et al., 2013; Gneezy et al., 2014; Ho et al., 2016; List and Momeni, 2021; Ploner and Regner, 2013), only few work exists to formalize the theory of moral balancing. Notable exceptions are Brañas-Garza et al. (2013), Gneezy et al. (2014), and List and Momeni (2017).

This paper adds to the theoretical literature on moral balancing by combining three behavioral concepts that are interrelated and thus essential to understanding prosocial behavior, but have not yet been jointly analyzed: First, individuals in the model do not have a complete understanding of their true moral values. Instead, they infer these values and construct their self-image from their past actions and, when making a decision, already take into account its implications for their future self-image, i. e. they use self-signaling (see, for example, Bénabou and Tirole, 2004, 2006, 2011; Bodner and Prelec, 2003; Flörchinger et al., 2025; Dubé et al., 2017; Grossman, 2014; Grossman and van der Weele, 2017; Gneezy et al., 2012; Falk, 2021). Second, individuals use moral balancing, meaning that they balance moral and immoral actions against each other and strive to achieve a certain level of self-image that they consider satisfactory (Ploner and Regner, 2013). I call this satisfactory level of self-image the self-image constraint. Third, they tend toward motivated reasoning, which is a combination of individuals placing less weight on negative signals, i. e. signals that

E-mail address: floerchinger@rwi-essen.de

https://doi.org/10.1016/j.jebo.2025.107393
Received 5 April 2024; Received in revised form 9 December 2025; Accepted 15 December 2025
Available online 8 January 2026
0167-2681/© 2025 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

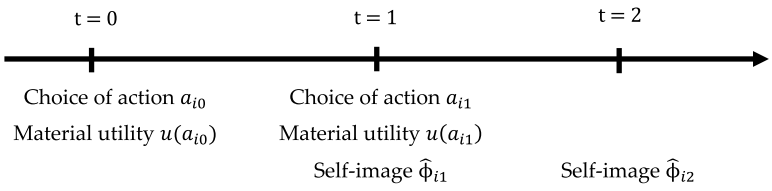


Fig. 1. Timing in the three-period model.

diminish their self-image, and negative signals being less persistent than positive signals (Druckman and McGrath, 2019; Bénabou and Tirole, 2002; Bénabou, 2015; Chew et al., 2020; Zimmermann, 2020).

The model comprises three periods and there are two versions of the model that differ by the time horizon for which individuals have to achieve their satisfactory level of self-image. In other words, the self-image constraint is either active in both the short and the long run or only in the long run. The reasons for an active self-image constraint in both the short and the long run could either be an intrinsic motivation to maintain a satisfactory self-image at all times, i. e. narrow choice bracketing (Read et al., 1999), or individuals could be unaware of another upcoming self-image-relevant choice (see Gneezy et al., 2014; Khan and Dhar, 2007).

The main results are that, first, when the self-image constraint is only active in the long run, individuals display more selfish behavior and selfish behavior is more concentrated in the first model period. This is in contrast to previous models, in which individuals have no incentive to advance selfish behavior. Second, using comparative statics, I show that selfish behavior is more prevalent the stronger the tendency toward motivated reasoning. The effects differ depending on which component of motivated reasoning increases – reduced weight of negative signals or lower persistence of negative signals – and depending on the time horizon for which the self-image constraint is active. These predictions are consistent with data from laboratory experiments by Khan and Dhar (2007) and Gneezy et al. (2014), and partially consistent with data on travel mode choice that was collected in an incentivized survey experiment among German household heads in 2019.

The subsequent section introduces the theoretical model. Section 3 compares predicted behavior in the two versions of the model. Section 4 analyzes how behavior changes with increased motivated reasoning. Section 5 calibrates the model to data from an experiment on travel mode choice, and Section 6 concludes.

2. Model

I develop a model of moral balancing to theoretically analyze why individuals fluctuate between selfish and prosocial actions and the role of motivated reasoning therein. Modeling discrete decisions between prosocial and selfish actions, where prosocial actions have positive implications and selfish actions have negative implications for the individuals' self-image, the model builds on Brañas-Garza et al. (2013).

The model has three periods (see Fig. 1). In the first two periods ($t = 0, 1$) individuals choose an action a_{it}, where i stands for the individual and t for the time period. Individuals immediately obtain utility. Future utility is discounted at discount rate $r > 0$. For simplicity, I assume that the utility function is the same for all individuals, such that utility evaluated at $t = 0$ is given by

$$U(a_{i0}, a_{i1}) \equiv u(a_{i0}) + \frac{1}{1+r} u(a_{i1}). \tag{1}$$

Individuals decide between a selfish action and a prosocial action, $a_{it} \in \{S, P\}$, where S stands for the selfish action and P for the prosocial action. There are four possible combinations of actions in $t = 0$ and $t = 1$:

$$(a_{i0}, a_{i1}) \in \Big\{ (S,S), (S,P), (P,S), (P,P) \Big\}.$$

In many cases, choosing a prosocial action over a selfish action implies a loss of convenience. Consider, for example, the decision of whether to travel by train or by plane (see also Flörchinger et al., 2025). Traveling by train corresponds to the prosocial action since it produces less negative externalities but at the same time it is much more time consuming. Traveling by plane is usually more convenient and thus corresponds to the selfish action. By definition, the selfish action yields larger material utility, $u(S) > u(P)$, and individuals in the model do not gain affective utility from prosocial actions. Thus, individuals maximizing their utility without constraints would choose $a_{it} = S$ in both periods $t = 0, 1$. However, individuals' pursuit of self-interest is constrained by the moral standards they seek to comply with (Rabin, 1995): They strive for a certain self-image that meets their standards, but experience no additional benefits from exceeding them. I call this the self-image constraint. Each action produces a signal $\hat{v}(a_{it})$ that individuals use to construct their self-image in subsequent periods. The signal derived from the prosocial action is positive, $\hat{v}(P) = v_H > 0$, i. e. it increases self-image. The signal derived from the selfish action is negative, $\hat{v}(S) = v_L < 0$, i. e. it diminishes self-image. Self-image $\hat{\phi}_{it}$ is evaluated in $t = 1, 2$ and depends on all actions taken in the previous periods. Temporally more distant actions are less important for current self-image because they are harder to recall (see Gneezy et al., 2014). Formally, in period t, signals from an action in period $s < t$ receive weight γ^{t-1-s}, with $\gamma \in (0, 1)$. This means that in time period t, signals from an action in time period $t-2$ receive weight $\gamma^{t-1-(t-2)} = \gamma$, whereas signals from period $t-1$ receive weight $\gamma^{t-1-(t-1)} = \gamma^0 = 1$.

What is perceived as a satisfactory level of self-image may differ between individuals. Individuals with low moral standards have low self-image constraints and individuals with high moral standards have high self-image constraints. There is a mass 1 of individuals

with their self-image constraints, $\bar{\phi}_i$, distributed over $\left[-v_H, v_H\right]$ with density function $f(\bar{\phi})$. The bounds of the distribution are set such that it is symmetric around zero and such that the individual with the highest self-image constraint can fulfill it by acting prosocially in both periods $t = 0, 1$. Since $v_L < 0$, self-image can be negative, and for some individuals it may be sufficient to achieve a negative self-image. To account for the case in which some individuals behave selfishly in both periods $t = 0, 1$, I assume that $-v_H < (1 + \gamma')v'_L$.

In line with empirical evidence (see, for example, Kahan, 2013; Möbius et al., 2022; Munro and Stansbury, 2009; Zimmermann, 2020), individuals tend toward motivated reasoning. That is, they prefer to have a positive self-image and consider information less credible if it contradicts their desired self-view (Druckman and McGrath, 2019; Kunda, 1990). For simplicity, I assume that everyone has the same tendency toward motivated reasoning. Specifically, when evaluating their self-image, individuals reduce the weight they place on negative signals by $\delta \in (0, 1)$ compared to positive signals. In addition, they tend to forget negative signals more quickly (Chew et al., 2020; Zimmermann, 2020), i. e. negative signals are less persistent than positive signals by $\epsilon \in (0, 1)$.

An individual's self-image starts at 0, and in period t it is given by

$$\hat{\phi}_{it} \equiv \sum_{s=0}^{t-1} \underbrace{\left[\left(1 - \underbrace{\epsilon^{\mathbb{1}[a_{is}=S]}}_{\substack{\text{reduced persistence}\\ \text{of negative signal}}}\right)\gamma\right]^{(t-1-s)}}_{\text{persistence of signal}} \left(1 - \underbrace{\delta^{\mathbb{1}[a_{is}=S]}}_{\substack{\text{reduced weight}\\ \text{of negative}\\ \text{signal}}}\right) \underbrace{\hat{v}(a_{is})}_{\text{signal}}. \tag{2}$$

To ease notation, I define the negative signal under motivated reasoning as

$$v'_L \equiv (1 - \delta)v_L, \tag{3}$$

and the persistence of negative signals under motivated reasoning as

$$\gamma' \equiv (1 - \epsilon)\gamma. \tag{4}$$

If δ and ϵ approach zero, there is virtually no motivated reasoning:

$$\lim_{\delta \to 0} v'_L(\delta) = v_L \text{ and} \tag{5}$$

$$\lim_{\epsilon \to 0} \gamma'(\epsilon) = \gamma. \tag{6}$$

If δ is close to one, individuals completely ignore negative signals, and if ϵ is close to one, they immediately forget them:

$$\lim_{\delta \to 1} v'_L(\delta) = 0 \text{ and} \tag{7}$$

$$\lim_{\epsilon \to 1} \gamma'(\epsilon) = 0. \tag{8}$$

Individuals are sophisticated in the sense that they know that they will place less weight on negative signals and that negative signals will be less persistent than positive signals. For the moment, the degree of motivated reasoning is taken as given, whereas Section 4 uses comparative statics to investigate how behavior varies with changes in δ and ϵ.

The model has two versions. In the first version, individuals have to fulfill their self-image constraint in both the short run and the long run (i. e. in $t = 1$ and in $t = 2$) – either because they are initially unaware of the second self-image relevant decision or because of narrow choice bracketing (Read et al., 1999). In the second version, the self-image constraint is only active in the long run (i. e. in $t = 2$), because individuals are aware in $t = 0$ that they will have to take a second self-image relevant decision and only care about their long-run self-image. Getting back to the example of travel mode choice, individuals may either plan from the start to take two trips in a given year (constraint active only in the long run), or they may initially plan only one trip and later see the necessity to additionally take a business trip – i. e. they are initially unaware of the second self-image relevant choice and the constraint is active in both the short and long run. However, the application of the model is not limited to decisions within the same domain. Empirical studies have shown that moral balancing also operates across different domains (Mazar and Zhong, 2010; Ho et al., 2016; Jordan et al., 2011).

2.1. Self-image constraint active in the short and long run ($t = 1, 2$)

In this version of the model, individuals have to meet their self-image constraint in both the short run and the long run, i. e. in $t = 1$ and in $t = 2$. This means that utility maximization is subject to two constraints. Total utility is given by Expression (1), but individuals maximize utility in each period separately, because, first, in $t = 0$ they might not be aware of the upcoming choice in $t = 1$ and, second, utility in each period is subject to a different constraint.[1] The utility maximization problem in $t = 0$ is

$$\max_{a_{i0} \in \{S,P\}} u(a_{i0}) \qquad \text{subject to} \qquad \hat{\phi}_{i1} = \left(1 - \delta^{\mathbb{1}[a_{i0}=S]}\right)\hat{v}(a_{i0}) \geq \bar{\phi}_i.$$

[1] Appendix A proves that behavior does not differ between myopic and forward-looking agents when the self-image constraint is active in $t = 1, 2$.

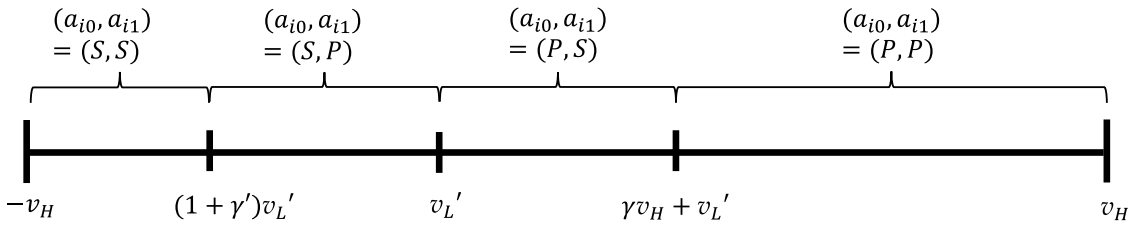


Fig. 2. Mapping of self-image constraints and choices with active self-image constraint in $t = 1$ and $t = 2$.

The utility maximization problem in $t = 1$ is

$$\max_{a_{i1}\in\{S,P\}} u(a_{i1}) \quad \text{subject to}$$

$$\hat{\phi}_{i2} = \left(1-\epsilon^{\mathbb{1}[a_{i0}=S]}\right)\gamma\left(1-\delta^{\mathbb{1}[a_{i0}=S]}\right)\hat{v}(a_{i0}) + \left(1-\delta^{\mathbb{1}[a_{i1}=S]}\right)\hat{v}(a_{i1}) \geq \bar{\phi}_i.$$

Proposition 1. *If the self-image constraint is active in both $t = 1$ and $t = 2$, there are four types of individuals that are characterized by their self-image constraint, which in turn determines their choice of (a_{i0}, a_{i1}). Fig. 2 maps self-image constraints and the corresponding actions.*[2]

*a) Individuals with **very low self-image constraints** $\bar{\phi}_i \in \left[-v_H, (1+\gamma')v_L'\right]$ choose $(a_{i0}, a_{i1}) = (S,S)$.*

*b) Individuals with **low self-image constraints** $\bar{\phi}_i \in \left((1+\gamma')v_L', v_L'\right]$ choose $(a_{i0}, a_{i1}) = (S,P)$.*

*c) Individuals with **high self-image constraints** $\bar{\phi}_i \in \left(v_L', \gamma v_H + v_L'\right]$ choose $(a_{i0}, a_{i1}) = (P,S)$.*

*d) Individuals with **very high self-image constraints** $\bar{\phi}_i \in \left(\gamma v_H + v_L', v_H\right]$ choose $(a_{i0}, a_{i1}) = (P,P)$.*

Proof. *Individuals have no incentive to overfulfill their self-image constraint, because $u(S) > u(P)$ and because they do not derive utility from a higher self-image. Individuals choose $a_{i0} = S$ whenever this choice satisfies their self-image constraint in $t = 1$, i. e. if $v_L' \geq \bar{\phi}_i$. Otherwise, they choose $a_{i0} = P$. In $t = 1$ the same logic applies. That is, individuals choose $a_{i1} = S$ if this choice satisfies their self-image constraint in $t = 2$. For individuals who have chosen $a_{i0} = S$, this is the case if $(1+\gamma')v_L' \geq \bar{\phi}_i$. For individuals who have chosen $a_{i0} = P$, the self-image constraint in $t = 2$ can be fulfilled by choosing $a_{i1} = S$ if $\gamma v_H + v_L' \geq \bar{\phi}_i$. Otherwise, individuals choose $a_{i1} = P$.*

□

2.2. Self-image constraint active only in the long run ($t = 2$)

In this version of the model, individuals maximize their material utility, $U(a_{i0}, a_{i1})$, under the constraint of achieving their satisfactory level of self-image, $\bar{\phi}_i$, only in $t = 2$:

$$\max_{(a_{i0},a_{i1})\in\{S,P\}^2} U(a_{i0}, a_{i1}) = u(a_{i0}) + \frac{1}{1+r}u(a_{i1})$$

subject to

$$\hat{\phi}_{i2} = \left(1-\epsilon^{\mathbb{1}[a_{i0}=S]}\right)\gamma\left(1-\delta^{\mathbb{1}[a_{i0}=S]}\right)\hat{v}(a_{i0}) + \left(1-\delta^{\mathbb{1}[a_{i1}=S]}\right)\hat{v}(a_{i1}) \geq \bar{\phi}_i.$$

If individuals' self-image constraint is too high to be satisfied by acting selfishly in both periods, they prefer to act selfishly in $t = 0$ rather than in $t = 1$, because this yields a greater material utility, as future material utility is discounted. Further, a higher self-image can be achieved in $t = 2$ by first choosing the selfish action and then the prosocial action since the negative signal does not fully persist until $t = 2$ when the selfish action is chosen in $t = 0$.

Proof. *$\hat{\phi}_{i2}(S,P) > \hat{\phi}_{i2}(P,S)$ is equivalent to*

$$\gamma' v_L' + v_H > \gamma v_H + v_L' \iff (1-\gamma)v_H > (1-\gamma')v_L' \iff \frac{v_H}{v_L'} < \frac{1-\gamma'}{1-\gamma}.$$

The last inequality always holds since $v_H > 0$, $v_L' < 0$, and $\gamma, \gamma' \in (0,1)$.

□

Thus, $(a_{i0}, a_{i1}) = (P,S)$ is strictly dominated by $(a_{i0}, a_{i1}) = (S,P)$ and the only plausible choices are $(a_{i0}, a_{i1}) \in \left\{(S,S),(S,P),(P,P)\right\}$.[3]

[2] Appendix B proves that Proposition 1 also holds for the extreme cases $\delta, \epsilon \to 0$ and $\delta, \epsilon \to 1$.

[3] Appendix C shows that none of the other choices is strictly dominated.

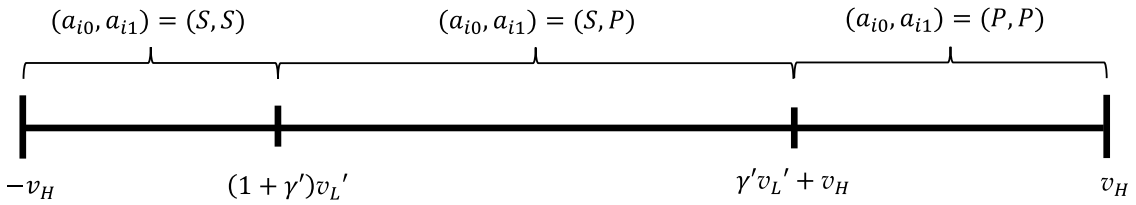


Fig. 3. Mapping of self-image constraints and choices with active self-image constraint in $t = 2$.

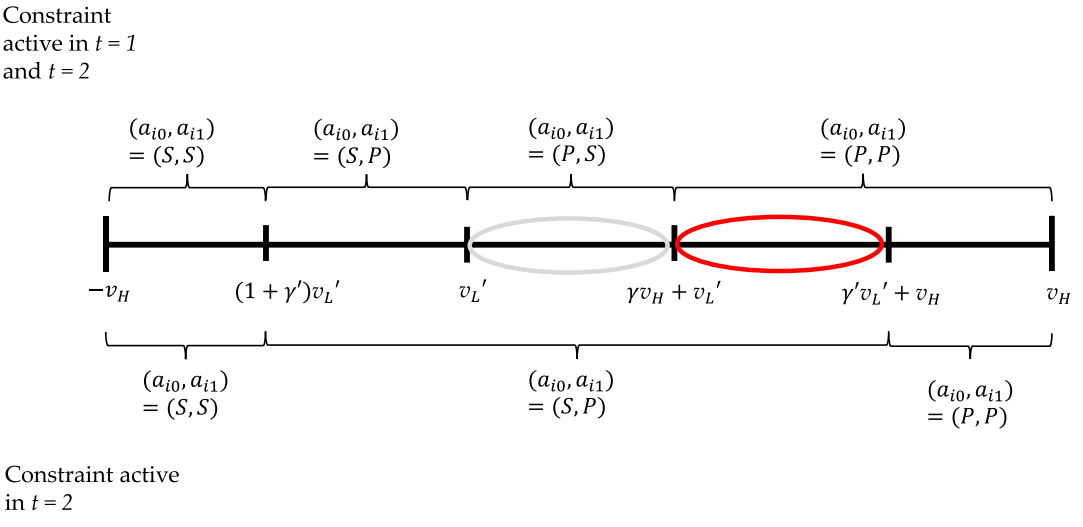


Fig. 4. Comparison between active constraint in $t = 1, 2$ and active constraint only in $t = 2$.

Proposition 2. *When the self-image constraint is only active in $t = 2$, there are three types of individuals that are characterized by their self-image constraint $\bar{\phi}_i$, which in turn determines their choice of (a_{i0}, a_{i1}). Fig. 3 maps self-image constraints and the corresponding actions.*[4]

a) Individuals with ***low self-image constraints*** $\bar{\phi}_i \in \left[-v_H, (1+\gamma')v_L'\right]$ *choose* $(a_{i0}, a_{i1}) = (S, S)$.

b) Individuals with ***intermediate self-image constraints*** $\bar{\phi}_i \in \left((1+\gamma')v_L', \gamma' v_L' + v_H\right]$ *choose* $(a_{i0}, a_{i1}) = (S, P)$.

c) Individuals with ***high self-image constraints*** $\bar{\phi}_i \in \left(\gamma' v_L' + v_H, v_H\right]$ *choose* $(a_{i0}, a_{i1}) = (P, P)$.

Proof. *Because $u(S) > u(P)$, individuals who have chosen $a_{i0} = S$ choose $a_{i1} = S$ if this choice satisfies their self-image constraint, i. e. if $(1+\gamma')v_L' \geq \bar{\phi}_i$. Otherwise, they choose $a_{i1} = P$. By choosing $(a_{i0}, a_{i1}) = (S, P)$ individuals achieve long-run self-image $\hat{\phi}_{i2} = \gamma' v_L' + v_H$. Thus, individuals with $\bar{\phi}_i > \gamma' v_L' + v_H$ would never choose $a_{i0} = S$. Instead, they choose $(a_{i0}, a_{i1}) = (P, P)$.*

□

Propositions 1 and 2 can explain the observation that some individuals consistently display either prosocial or selfish behavior, while others fluctuate between prosocial and selfish actions.

3. Comparison of behavior in model version with active constraint in $t = 1, 2$ and model version with active constraint only in $t = 2$

This section compares the two versions of the model described in the previous section. Transitioning from the version with an active self-image constraint in both the short and long run to the version with an active constraint only in the long run can be viewed as lifting the constraint in $t = 1$, such that self-image in $t = 1$ becomes irrelevant.

In both versions of the model, individuals' choices of (a_{i0}, a_{i1}) depend on whether their self-image constraints are smaller or larger than a given threshold. Fig. 4 maps thresholds of $\bar{\phi}_i$ and choices of (a_{i0}, a_{i1}) for both scenarios. The scenario with an active constraint in $t = 1, 2$ is displayed in the upper part, the scenario with an active constraint only in $t = 2$ is displayed in the lower part.

Fig. 4 shows that when the constraint in $t = 1$ is removed, two types of individuals change their behavior. First, individuals with $\bar{\phi}_i \in \left(v_L', \gamma v_H + v_L'\right]$ switch from $(a_{i0}, a_{i1}) = (P, S)$ to $(a_{i0}, a_{i1}) = (S, P)$. In this range, self-image constraints are low enough such that

[4] Appendix D proves that Proposition 2 also holds for the extreme cases $\delta, \epsilon \to 0$ and $\delta, \epsilon \to 1$.

they can still be met in the long run if the selfish action, S, is chosen once. Because of intertemporal discounting, material utility is larger if individuals choose the selfish action earlier. However, since $\bar{\phi}_i > v'_L$, the self-image constraint in $t = 1$ cannot be fulfilled by choosing $a_{i0} = S$. Thus, when the constraint in $t = 1$ is removed, individuals with $\bar{\phi}_i \in \left(v'_L, \gamma v_H + v'_L\right]$ can increase their material utility by postponing the prosocial action and choosing $(a_{i0}, a_{i1}) = (S, P)$ instead of $(a_{i0}, a_{i1}) = (P, S)$.

Second, individuals with $\bar{\phi}_i \in \left(\gamma v_H + v'_L, \gamma' v'_L + v_H\right]$ switch from $(a_{i0}, a_{i1}) = (P, P)$ to $(a_{i0}, a_{i1}) = (S, P)$. For individuals with self-image constraints in this range, the loss in self-image from choosing $a_{i0} = S$ is sufficiently compensated by the subsequent gain from choosing $a_{i1} = P$, such that the constraint in $t = 2$ can still be satisfied. Since $\bar{\phi}_i > v'_L$, the constraint in $t = 1$ cannot be satisfied by choosing $a_{i0} = S$. Therefore, individuals with $\bar{\phi}_i \in \left(\gamma v_H + v'_L, \gamma' v'_L + v_H\right]$ can increase their material utility by choosing $a_{i0} = S$ instead of $a_{i0} = P$ if the constraint in $t = 1$ is removed. Thus, overall selfish behavior increases and selfish behavior is more concentrated in $t = 0$ when the self-image constraint in $t = 1$ is lifted. Proposition 3 summarizes these findings.[5]

Proposition 3. *When the self-image constraint in $t = 1$ is lifted, such that there is an active constraint only in $t = 2$, individuals with $\bar{\phi}_i \in \left(v'_L, \gamma v_H + v'_L\right]$, i. e. all individuals who choose $(a_{i0}, a_{i1}) = (P, S)$ under an active constraint in $t = 1$ (see Proposition 1), switch from $(a_{i0}, a_{i1}) = (P, S)$ to $(a_{i0}, a_{i1}) = (S, P)$ and individuals with $\bar{\phi}_i \in \left(\gamma v_H + v'_L, \gamma' v'_L + v_H\right]$ switch from $(a_{i0}, a_{i1}) = (P, P)$ to $(a_{i0}, a_{i1}) = (S, P)$. Overall selfish behavior increases and it is more concentrated in $t = 0$. Further, prosocial behavior in $t = 1$ increases when the self-image constraint in $t = 1$ is lifted.*

Proposition 3 is consistent with empirical results from Gneezy et al. (2014) and Khan and Dhar (2007), who find that individuals are less likely to behave ethically in the first round of the experiment when they know that there will be a second round with another self-image relevant choice compared to when they are not aware. It further implies that imposing higher self-image constraints in one period leads to lower prosocial behavior in later periods.

4. Comparative statics: Different degrees of motivated reasoning

This section examines how the choice of (a_{i0}, a_{i1}) varies with the tendency toward motivated reasoning. As outlined in Section 2, motivated reasoning is determined by two parameters: the extent δ to which negative signals receive less weight in the formation of self-image and their reduced persistence, ϵ.

Motivated reasoning implicitly relaxes the self-image constraint by reducing the impact on self-image of negative signals extracted from selfish behavior. Thus, selfish behavior should increase with motivated reasoning. As the analyses below show, the exact impact depends on which component of motivated reasoning increases and on the time horizon for which the self-image constraint is active.

In what follows, $Z^s_{(a_{i0},a_{i1})}$ denotes the share of individuals who choose a specific behavior pattern (a_{i0}, a_{i1}) given that the self-image constraint is active in period s. Θ^s_t stands for the amount of selfish behavior in period t given that the self-image constraint is active in period s.

4.1. Self-image constraint active in the short and long run ($t = 1, 2$)

From Proposition 1 and because individuals' self-image constraints, $\bar{\phi}_i$, are distributed over $\left[-v_H, v_H\right]$ with density function $f(\bar{\phi})$, it follows that when the self-image constraint is active in $t = 1$ and $t = 2$, the share of individuals who choose $(a_{i0}, a_{i1}) = (S, S)$ is

$$Z^{1,2}_{(S,S)} = \int_{-v_H}^{(1+\gamma')v'_L} f(\bar{\phi})d\bar{\phi} = F\left((1+\gamma')v'_L\right), \tag{9}$$

the share of individuals who choose $(a_{i0}, a_{i1}) = (S, P)$ is

$$Z^{1,2}_{(S,P)} = \int_{(1+\gamma')v'_L}^{v'_L} f(\bar{\phi})d\bar{\phi} = F\left(v'_L\right) - F\left((1+\gamma')v'_L\right), \tag{10}$$

and the share of individuals who choose $(a_{i0}, a_{i1}) = (P, S)$ is

$$Z^{1,2}_{(P,S)} = \int_{v'_L}^{\gamma v_H + v'_L} f(\bar{\phi})d\bar{\phi} = F\left(\gamma v_H + v'_L\right) - F\left(v'_L\right). \tag{11}$$

Thus, the amount of selfish behavior in $t = 0$ is

$$\Theta^{1,2}_0 = Z^{1,2}_{(S,S)} + Z^{1,2}_{(S,P)} = F\left(v'_L\right), \tag{12}$$

and the amount of selfish behavior in $t = 1$ is

$$\Theta^{1,2}_1 = Z^{1,2}_{(S,S)} + Z^{1,2}_{(P,S)} = F\left((1+\gamma')v'_L\right) + F\left(\gamma v_H + v'_L\right) - F\left(v'_L\right). \tag{13}$$

[5] Since Proposition 3 builds on the thresholds of $\bar{\phi}_i$ established in Propositions 1 and 2, and Propositions 1 and 2 continue to hold for the extreme cases of $\delta, \epsilon \to 0$ and $\delta, \epsilon \to 1$ (see Appendices B and D), Proposition 3 holds for these extreme cases as well.

Recall that $v_L < 0$, $v_H > 0$, $v'_L = (1-\delta)v_L$, and $\gamma' = (1-\epsilon)\gamma$ (see Expressions (3) and (4)). Thus, when negative signals receive less weight, i. e. when δ increases, more individuals can fulfill their self-image constraints by acting selfishly and there is more selfish behavior compared to a situation with lower δ. Selfish behavior in $t = 0$ increases unambiguously, whereas increases in $t = 1$ are contingent on the distribution of self-image constraints. The increase of selfish behavior in $t = 0$ is always larger than a potential decrease in $t = 1$, such that overall selfish behavior increases. Regarding the persistence of negative signals, the amount of selfish behavior in $t = 0$ is independent from ϵ, that is, from how quickly negative signals fade. If the self-image constraint is active in $t = 1$, individuals cannot "borrow" from the second period to increase their short-run material utility. Thus, it does not matter how large their debt would be, i. e. which proportion of the violation of their self-image constraint would be carried over to the next period. Selfish behavior in $t = 1$, by contrast, is increasing in ϵ. Whether increases in δ and ϵ attenuate or reinforce each other is contingent on the distribution of self-image constraints. Proposition 4 summarizes these findings.

Proposition 4. *In the version of the model with an active self-image constraint in $t = 1$ and $t = 2$,*

a) selfish behavior in $t = 0$ increases with δ. Selfish behavior in $t = 1$ can either increase or decrease with δ, depending on the distribution of self-image constraints. It increases if $f\left(v'_L\right) < f\left((1+\gamma')v'_L\right)(1+\gamma') + f\left(\gamma v_H + v'_L\right)$. Overall selfish behavior is unambiguously increasing in δ.
b) the amount of selfish behavior in $t = 0$ is independent from ϵ and the amount of selfish behavior in $t = 1$ is increasing in ϵ.
c) whether the effects of increases in δ and ϵ on selfish behavior in $t = 1$ attenuate or reinforce each other, depends on the distribution of $\bar{\phi}$. They attenuate each other if $f'\left((1+\gamma')v'_L\right)(1+\gamma')v'_L > -f\left((1+\gamma')v'_L\right)$.

Proof. *From Expressions (12) and (13) follows that*

a) (i) $\frac{\partial \Theta_0^{1,2}}{\partial \delta} = -f\left(v'_L\right)v_L > 0$;
(ii) $\frac{\partial \Theta_1^{1,2}}{\partial \delta} = v_L\left[-f\left((1+\gamma')v'_L\right)(1+\gamma') - f\left(\gamma v_H + v'_L\right) + f\left(v'_L\right)\right]$,
$\frac{\partial \Theta_1^{1,2}}{\partial \delta} > 0 \iff f\left(v'_L\right) < f\left((1+\gamma')v'_L\right)(1+\gamma') + f\left(\gamma v_H + v'_L\right)$;
(iii) $\frac{\partial \Theta_0^{1,2}}{\partial \delta} + \frac{\partial \Theta_1^{1,2}}{\partial \delta} = v_L\left[-f\left((1+\gamma')v'_L\right)(1+\gamma') - f\left(\gamma v_H + v'_L\right)\right] > 0$.
b) (i) $\frac{\partial \Theta_0^{1,2}}{\partial \epsilon} = 0$,
(ii) $\frac{\partial \Theta_1^{1,2}}{\partial \epsilon} = -f\left((1+\gamma')v'_L\right)\gamma v'_L > 0$.
c) *Further, the cross-derivatives are given by*
(i) $\frac{\partial^2 \Theta_0^{1,2}}{\partial \delta \partial \epsilon} = 0$ *and*
(ii) $\frac{\partial^2 \Theta_1^{1,2}}{\partial \delta \partial \epsilon} = \gamma v_L\left[f'\left((1+\gamma')v'_L\right)(1+\gamma')v'_L + f\left((1+\gamma')v'_L\right)\right]$, $\frac{\partial^2 \Theta_1^{1,2}}{\partial \delta \partial \epsilon} < 0 \iff f'\left((1+\gamma')v'_L\right)(1+\gamma')v'_L > -f\left((1+\gamma')v'_L\right)$.

□

4.2. Self-image constraint active only in the long run ($t = 2$)

From Proposition 2 and because individuals' self-image constraints, $\bar{\phi}_i$, are distributed over $\left[-v_H, v_H\right]$ with density function $f(\bar{\phi})$, it follows that when the self-image constraint is active only in $t = 2$, the share of individuals who choose $(a_{i0}, a_{i1}) = (S, S)$ is

$$Z^2_{(S,S)} = \int_{-v_H}^{(1+\gamma')v'_L} f(\bar{\phi})d\bar{\phi} = F\left((1+\gamma')v'_L\right) \tag{14}$$

and the share of individuals who choose $(a_{i0}, a_{i1}) = (S, P)$ is

$$Z^2_{(S,P)} = \int_{(1+\gamma')v'_L}^{\gamma' v'_L + v_H} f(\bar{\phi})d\bar{\phi} = F\left(\gamma' v'_L + v_H\right) - F\left((1+\gamma')v'_L\right). \tag{15}$$

Thus, the amount of selfish behavior in $t = 0$ is

$$\Theta^2_0 = Z^2_{(S,S)} + Z^2_{(S,P)} = F\left(\gamma' v'_L + v_H\right), \tag{16}$$

and the amount of selfish behavior in $t = 1$ is

$$\Theta^2_1 = Z^2_{(S,S)} = F\left((1+\gamma')v'_L\right). \tag{17}$$

When negative signals are less important, i. e. when δ is larger, selfish behavior in $t = 0, 1$ increases. Selfish behavior also increases when signals extracted from selfish behavior are less persistent, i. e. when ϵ is higher. Whether the effects of increases in δ and ϵ attenuate or reinforce each other, depends on the distribution of $\bar{\phi}$. Proposition 5 summarizes these findings.

Proposition 5. *In the version of the model with an active self-image constraint only in $t = 2$,*

a) selfish behavior in $t = 0, 1$ is increasing in δ.

b) selfish behavior in $t = 0, 1$ is increasing in ϵ.

c) whether the effects of increases in δ and ϵ on selfish behavior in $t = 0, 1$ attenuate or reinforce each other, depends on the distribution of $\bar{\phi}$. In $t = 0$, they attenuate each other if $f'\left(\gamma' v'_L + v_H\right)\gamma' v'_L > -f\left(\gamma' v'_L + v_H\right)$ and in $t = 1$ they attenuate each other if $f'\left((1+\gamma')v'_L\right)(1+\gamma')v'_L > -f\left((1+\gamma')v'_L\right)$.

Proof. *From Expressions (16) and (17) it follows that*

a) (i) $\frac{\partial \Theta_0^2}{\partial \delta} = -f\left(\gamma' v'_L + v_H\right)\gamma' v_L > 0;$

(ii) $\frac{\partial \Theta_1^2}{\partial \delta} = -f\left((1+\gamma')v'_L\right)(1+\gamma')v_L > 0.$

b) (i) $\frac{\partial \Theta_0^2}{\partial \epsilon} = -f\left(\gamma' v'_L + v_H\right)\gamma v'_L > 0$

(ii) $\frac{\partial \Theta_1^2}{\partial \epsilon} = -f\left((1+\gamma')v'_L\right)\gamma v'_L > 0.$

c) *Further, the cross-derivatives of (16) and (17) are given by*

(i) $\frac{\partial^2 \Theta_0^2}{\partial \delta \partial \epsilon} = \gamma v_L\left[f'\left(\gamma' v'_L + v_H\right)\gamma' v'_L + f\left(\gamma' v'_L + v_H\right)\right];$

$\frac{\partial^2 \Theta_0^2}{\partial \delta \partial \epsilon} < 0 \iff f'\left(\gamma' v'_L + v_H\right)\gamma' v'_L > -f\left(\gamma' v'_L + v_H\right).$

(ii) $\frac{\partial^2 \Theta_1^2}{\partial \delta \partial \epsilon} = \gamma v_L\left[f'\left((1+\gamma')v'_L\right)(1+\gamma')v'_L + f\left((1+\gamma')v'_L\right)\right];$

$\frac{\partial^2 \Theta_1^2}{\partial \delta \partial \epsilon} < 0 \iff f'\left((1+\gamma')v'_L\right)(1+\gamma')v'_L > -f\left((1+\gamma')v'_L\right).$

□

5. Calibration

In the following, I calibrate the model to an empirical dataset on travel mode choice. Participants in an incentivized online experiment had to make two self-image-relevant decisions: First, they had to decide whether they would prefer to win a voucher for a flight or a train ticket. Second, those who chose the flight voucher had the option of making a donation to offset their emissions from flying. This design can be used to test Proposition 3 because for one half of the sample, the offset decision came as a surprise, i. e. their self-image constraint was active in both the short and long run. For the other half, the self-image constraint was active only in the long run, because they were informed about the offset possibility before their voucher choice. Another treatment dimension varies the scope δ to give less weight to the signals extracted from the flight voucher choice, which facilitates the testing of Propositions 4 a) and 5 a).

Section 5.1 describes the experimental design and additional assumptions made for the model calibration. Section 5.2 calibrates the model version with an active self-image constraint in the short and long run and tests Proposition 4 a). Section 5.3 calibrates the model version with an active self-image constraint only in the long run and tests Propositions 3 and 5 a).

5.1. Experimental design and assumptions

I calibrate the model to data from an experiment that was part of a survey conducted in late 2019, in collaboration with *forsa* (Matejko and Sommer, 2022). The relevant survey questions can be found in Appendix F. *forsa* maintains a panel of about 100,000 members, being representative of the German-speaking population in Germany. The survey was addressed to household heads, defined as those individuals who are responsible for the household's financial decisions. Data collection was based on a state-of-the-art tool that allows panelists to retrieve and return the questionnaire from home or from mobile devices connected to the internet. The survey could be interrupted and resumed at any time.

At the heart of the experiment lies the decision between the possibility of winning a voucher worth 40 euros for either a flight or a train ticket.[6] Subsequently, respondents who had chosen the flight voucher had to decide whether to offset greenhouse gas emissions from flying through a donation to atmosfair, an NGO dedicated to offsetting emissions from traveling. In the last years, because of its contribution to climate change, flying has been discredited by an anti-flying movement that is spearheaded by *Fridays for Future* (CNN, 2019).[7] Accordingly, traveling by train is the option that is conceived in the experiment to correspond to the prosocial action. To obtain incentive-compatible responses, one out of 50 respondents was randomly selected to have their choice implemented.

The experiment followed a 2×2 design. One half of the respondents were randomly assigned to group $Informed$ or group $NotInformed$. Respondents in group $Informed$ were informed about the possibility to offset emissions before their voucher choice. For group $NotInformed$, the offsetting decision came as a surprise after having selected the flight voucher. The second treatment

[6] The voucher was valid either for the German Railways (*Deutsche Bahn*) or the website of Flightgift, where flights from a large variety of airlines can be booked.

[7] The origin of the *Fridays for Future* movement dates back to August 2018, when Greta Thunberg decided to strike for climate protection outside the Swedish Parliament on each school day. She was soon joined by youth from all around the world, leading to a global strike movement (Fridays for Future, 2022). In Germany, the first protest was organized in September 2018. Thereafter, the movement grew substantially, comprising a total of 1.4 million protesters across more than 550 cities in September 2019 (Neuber and Gardner, 2020).

dimension consisted in whether, immediately before their voucher choice, respondents were reminded of their attitude toward the *Fridays for Future* movement, which they had stated a few questions before. To summarize, respondents in group $Informed \times Reminder$ (abbreviated by $I \times R$) knew that it would be possible to offset emissions from flying and they were reminded of their attitude toward *Fridays for Future*. Respondents in group $Informed \times NoReminder$ ($I \times NR$) also knew about the possibility to offset and were not reminded of their attitude. Respondents in group $NotInformed \times Reminder$ ($NI \times R$), when choosing their voucher, were not aware of the possibility to offset emissions and were reminded of their attitude toward *Fridays for Future*. Finally, respondents in group $NotInformed \times NoReminder$ ($NI \times NR$) were not aware of the offsetting opportunity and received no reminder.

The experimental design is closely aligned with the model presented in this paper. First, when respondents are not aware of the second self-image-relevant decision when choosing their voucher (i. e. when the offsetting decision comes as a surprise, as in group $NotInformed$), the self-image constraint is active in both the short and the long run. By contrast, when respondents are aware of the offsetting opportunity before the voucher choice (as in group $Informed$), at least some will apply broad choice bracketing, i. e. their self-image constraint is only active in the long run. Second, due to the extensive media coverage of *Fridays for Future's* criticism of flying (see, for example, CNN, 2019), the reminder of their attitude toward *Fridays for Future* increases individuals' awareness of the negative externalities of flying and thus leaves less scope for reducing the importance of signals extracted from the choice of the flight voucher, i. e. the reminder reduces δ.

One drawback of the experimental design is that there is no second choice for respondents who opted for the train voucher, i. e. a_{i1} is only observed if $a_{i0} = S$. Accordingly, in all four treatment conditions empirical values are available for the amount of selfish behavior in $t = 0$ (i. e. the share of individuals who choose the flight), the share of individuals who choose $(a_{i0}, a_{i1}) = (S, S)$ (i. e. those who choose the flight without offsetting), and the share of those who choose $(a_{i0}, a_{i1}) = (S, P)$ (i. e. those who choose the flight and subsequently offset) – although this last category is not needed for the model calibration. Adopting the notation from Section 4, the values required for the model calibration are denoted by $\Theta_0^{1,2(NR)}$ and $Z_{S,S}^{1,2(NR)}$ for condition $NotInformed \times NoReminder$, $\Theta_0^{1,2(R)}$ and $Z_{S,S}^{1,2(R)}$ for condition $NotInformed \times Reminder$, $\Theta_0^{2(NR)}$ and $Z_{S,S}^{2(NR)}$ for condition $Informed \times NoReminder$, and $\Theta_0^{2(R)}$ and $Z_{S,S}^{2(R)}$ for condition $Informed \times Reminder$. After choosing values for the negative signal, v_L, the positive signal, v_H, and the persistence of signals, γ, as well as making an assumption about the distribution of moral constraints, $f(\bar{\phi})$, one can calculate the parameters for motivated reasoning for each condition, $\delta^{G1\times G2}$ and $\epsilon^{G1\times G2}$, with $G1 \in \{NI, I\}$ and $G2 \in \{NR, R\}$.

To fulfill the conditions that $\gamma \in (0,1)$ and $-v_H < (1+\gamma')v_L'$ (see Section 2), I set $v_L = -120$, $v_H = 200$, and $\gamma = 0.6$.[8] For simplicity, I assume that $\bar{\phi}$ is uniformly distributed over $[-v_H, v_H]$ with density function $f(\bar{\phi}) = \frac{1}{2v_H}$.

5.2. Self-image constraint active in the short and long run ($t = 1, 2$)

I start the calibration of the model with the scenario in which the self-image constraint is active in both the short and the long run, i. e. the offsetting decision comes as a surprise. In condition $NotInformed \times NoReminder$, 24.09 % of respondents choose the flight voucher and 10.32 % choose the flight voucher without offsetting, i. e. $\Theta_0^{1,2(NR)} = 0.2409$ and $Z_{(S,S)}^{1,2(NR)} = 0.1032$. Plugging $f(\bar{\phi}) = \frac{1}{2v_H}$ and $v_L' = (1-\delta)v_L$ (see (3)) into Expression (12) and solving for δ, one obtains

$$\delta^{NI\times G2} = \frac{v_L + v_H - 2v_H\Theta_0^{1,2}}{v_L}. \tag{18}$$

Similarly, plugging $f(\bar{\phi}) = \frac{1}{2v_H}$, (3), and $\gamma' = (1-\epsilon)\gamma$ (see (4)) into Expression (9) and solving for ϵ yields

$$\epsilon^{NI\times G2} = \frac{(1+\gamma)(1-\delta)v_L + v_H - 2v_H Z_{(S,S)}^{1,2}}{\gamma(1-\delta)v_L}. \tag{19}$$

Finally, plugging in $\Theta_0^{1,2(NR)}$ and $Z_{(S,S)}^{1,2(NR)}$ gives empirical values for the parameters of motivated reasoning when the offsetting decision comes as a surprise and respondents are not reminded of their attitude toward *Fridays for Future*: $\delta^{NI\times NR} \approx 0.1363$ and $\epsilon^{NI\times NR} \approx 0.1142$. This is consistent with the model assumption that $\delta, \epsilon \in (0,1)$. It implies that the impact of negative signals on self-image is 86.37 %(= $100(1-\delta^{NI\times NR})$) of the impact of a positive signal. Further, the persistence of a negative signal is 88.58 %(= $100(1-\epsilon^{NI\times NR})$) of the persistence of positive signals. Compared to the results of Zimmermann (2020), who found limited asymmetry in updating to positive and negative signals in the short run, $\delta^{NI\times NR} = 0.1363$ seems reasonable, perhaps even reaching the upper bound. By contrast, in the long run, Zimmermann (2020) finds barely any response to negative signals, suggesting that the value of the parameter $\epsilon^{NI\times NR} = 0.1142$ is rather low. While a comparison with the results of Zimmermann (2020) can provide an indication of how realistic the motivated reasoning parameters are, it is important to bear in mind that Zimmermann (2020) investigates beliefs about performance in an intelligence test, whereas this paper addresses prosocial behavior.

Moving on to the next treatment condition, reminding respondents of their attitude toward *Fridays for Future* prior to their voucher choice should increase their awareness of the negative externalities of flying and, thereby, reduce the scope δ to place less weight on the signals extracted from the flight voucher choice. According to Proposition 4 a), the decrease in δ should lead to a decrease in selfish behavior in both periods. Empirically, in condition $NotInformed \times Reminder$, 21.28 % of respondents choose the flight voucher

[8] $-v_H < (1+\gamma')v_L'$ is fulfilled with these values because (i) $-v_H < (1+\gamma)v_L$, (ii) $(1+\gamma')v_L' = (1+\gamma)v_L$ iff $\delta, \epsilon = 0$, (iii) $\frac{\partial(1+\gamma')v_L'}{\partial\delta} = -(1+\gamma')v_L > 0$, $\frac{\partial(1+\gamma')v_L'}{\partial\epsilon} = -v_L' > 0$, and (iv) $\delta, \epsilon \in (0,1)$ (see Section 2).

and 8.74 % choose the flight voucher without offsetting, i. e. selfish behavior indeed decreases when respondents are reminded of their attitude toward *Fridays for Future* since $\Theta_0^{1,2(R)} = 0.2128 < \Theta_0^{1,2(NR)}$ and $Z_{(S,S)}^{1,2(R)} = 0.0874 < Z_{(S,S)}^{1,2(NR)}$. To check whether this goes along with a decrease in δ, I plug $\Theta_0^{1,2(R)}$ into Expression (18). This gives $\delta^{NI\times R} = 0.0427 < \delta^{NI\times NR}$ and is, thus, consistent with Proposition 4 a). Calculating the empirical value for the second parameter of motivated reasoning, the reduced persistence ϵ of negative signals, by plugging $\delta^{NI\times R}$ and $Z_{(S,S)}^{1,2(R)}$ into Expression (19) yields $\epsilon^{NI\times R} = 0.2723 > \epsilon^{NI\times NR}$. This implies that, along with the decrease in δ, a simultaneous increase in ϵ is required to produce the behavioral patterns observed in the data. According to Proposition 4 c), a simultaneous decrease in δ and an increase in ϵ entail a greater reduction in selfish behavior at $t = 1$ than a decrease in δ alone. Thus, the data are consistent with the model.

5.3. *Self-image constraint active only in the long run ($t = 2$)*

I now move to the scenario where the self-image constraint is only active in $t = 2$, i. e. respondents are informed about the possibility to offset emissions before they choose their voucher. Assuming that δ and ϵ remain the same as in the scenario with an active constraint in $t = 1, 2$ and thus plugging $\delta^{NI\times NR}$ and $\epsilon^{NI\times NR}$ into Expression (16), the model predicts that if respondents are not reminded of their attitude toward *Fridays for Future*, selfish behavior in $t = 0$ should increase to $\hat{\Theta}_0^{2(NR)} = 0.8623$ when the self-image constraint in $t = 1$ is lifted.

The data show a different picture: In condition $Informed \times NoReminder$, 25.43 % of respondents choose the flight voucher and 11.10 % of respondents do so without offsetting emissions. Thus, in line with Proposition 3, selfish behavior in $t = 0$ is more pronounced when there is no self-image constraint in $t = 1$ as $\Theta_0^{2(NR)} = 0.2543 > \Theta_0^{1,2(NR)}$. However, the amount of selfish behavior in $t = 0$ is much smaller than predicted by the model through $\hat{\Theta}_0^{2(NR)} = 0.8623$ and selfish behavior in $t = 1$ increases as well – which is in contrast to Proposition 3. This means that the data cannot be explained by the removal of the self-image constraint in $t = 1$ alone. This also becomes evident when calculating empirical values for δ and ϵ in the scenario with an active self-image constraint only in $t = 2$. First, I plug $f(\bar{\phi}) = \frac{1}{2v_H}$, (3), and (4) into Expression (16) and Expression (14) and solve both (16) and (14) for ϵ. Then, I set $\epsilon = \epsilon$ and solve for δ. This yields

$$\delta^{I\times G2} = \frac{v_L + 2v_H(\Theta_0^2 - Z_{(S,S)}^2) - v_H}{v_L} \tag{20}$$

and

$$\epsilon^{I\times G2} = \frac{(1+\gamma)(1-\delta)v_L - 2v_H Z_{(S,S)}^2 + v_H}{\gamma(1-\delta)v_L}. \tag{21}$$

When respondents are not reminded of their attitude toward *Fridays for Future*, δ and ϵ can be calculated by plugging $\Theta_0^{2(NR)}$ and $Z_{(S,S)}^{2(NR)}$ into (20) and (21). This gives $\delta^{I\times NR} \approx 2.1890$ and $\epsilon^{I\times NR} \approx 4.4843$ and is in contrast to the model, which assumes that $\delta, \epsilon \in (0, 1)$.

$\delta > 2$ implies that motivated reasoning is so pronounced that a negative signal is perceived as a positive signal and, additionally, it receives more weight than a genuinely positive signal. Similarly, $\epsilon > 2$ means that in period t, a negative signal from period $t - 2$ is perceived as a positive signal and it is more persistent than a genuinely positive signal. This is very unlikely to hold in reality.

Moving to condition $Informed \times Reminder$, the data show that 20.80 % of respondents choose the flight voucher and 8.10 % do so without offsetting, i. e. $\Theta_0^{2(R)} = 0.2080$ and $Z_{(S,S)}^{2(R)} = 0.0810$. This is in line with Proposition 5 a), as $\Theta_0^{2(R)} = 0.2080 < \Theta_0^{2(NR)}$ and $Z_{(S,S)}^{2(R)} = 0.0810 < Z_{(S,S)}^{2(NR)}$. Plugging these values into Expressions (20) and (21) yields $\delta^{I\times R} \approx 2.2433$ and $\epsilon^{I\times R} \approx 4.5389$, which is equally unrealistic as the values obtained for condition $Informed \times NoReminder$. Further, Proposition 3 predicts that there should be more selfish behavior in $t = 0$ when the self-image constraint is only active in the long run compared to when it is active in both the short and the long run. However, when respondents are reminded of their attitude towards *Fridays for Future*, the data show the opposite as $\Theta_0^{2(R)} = 0.2080 < \Theta_0^{1,2(R)}$.

While Appendix E provides a detailed description of why the data may not fully align with the model, this section gives some intuition. Firstly, some respondents may be intrinsically motivated to always maintain a satisfactory level of self-image, i. e. their self-image constraint is active even when they are aware of the second self-image-relevant choice. Secondly, informing respondents about the possibility of offsetting emissions from flying before they choose a voucher influences voucher choice not only by removing the self-image constraint in $t = 1$, but also by highlighting the negative environmental impact of flying. Thus, the treatment has two opposite effects: On the one hand, it leads to an increase in flying by providing an opportunity to offset negative externalities, and on the other hand, it leads to a decrease in flying by highlighting negative externalities. Thirdly, respondents are likely to travel by train more often than by air, so, irrespective of the treatment condition, the train voucher is more useful to them than the flight voucher. Finally, some assumptions may be overly simplistic. Taken together, these facts may explain why selfish behavior in the first period is empirically lower than predicted by the model when respondents are aware of the offsetting opportunity.

6. Summary and conclusion

In a three-period model with discrete choices between selfish and prosocial actions, this paper provides the first joint analysis of self-signaling, moral balancing, and motivated reasoning. Individuals are constrained in their pursuit of self-interest by their own moral standards in the sense that they need to achieve a certain level of self-image that satisfies their standards, but they cannot gain additional utility by overfulfilling those standards. Since the utility of prosocial actions is lower than the utility of selfish actions, e. g.

because sustainable products are more expensive or less convenient, individuals try to avoid prosocial actions as much as possible. The avoidance of prosocial actions is facilitated by biased information processing which gives less weight to signals that contradict the individuals' desired self-image, i. e. motivated reasoning. This is in line with empirical observations that individuals actively seek excuses to avoid prosocial behavior (see, for example, Dana et al., 2006; Exley, 2016; Andreoni et al., 2017; Momsen and Ohndorf, 2022).

The model shows that individuals behave differently depending on whether they strive to fulfill their self-image constraint only in the long run or in both the short and the long run. Selfish behavior is more prevalent and more concentrated in the first period when the satisfactory self-image has to be achieved only in the long run, as prosocial behavior in subsequent periods can repair the damage inflicted to self-image from selfish conduct in earlier periods. Motivated reasoning further impedes prosocial behavior.

Calibrating the model to data from an experiment on travel mode choice shows that empirical observations are partially in line with the model predictions. These discrepancies may be due to the experimental setup not being entirely consistent with the model assumptions. Therefore, I encourage future research to design alternative experiments to test the model.

The model results imply that voluntary prosocial behavior is unlikely to be sustained. Thus, if prosocial behavior is socially desirable, i. e. if societal gains from prosocial behavior outweigh individual utility losses, policymakers should foster prosocial behavior through adequate policy measures, such as taxing selfish behavior or implementing mandatory behavioral standards. In addition, real-time feedback (see Andor et al., 2023; Byrne et al., 2023; Tiefenbeck et al., 2018, 2019; Fang et al., 2023) can foster regular reflection on self-image and thereby lead to an additional increase in prosocial behavior.

Data availability

The authors do not have permission to share data.

Funding

This work was supported by the Federal Ministry of Research, Technology and Space (BMFTR) in the framework of the Kopernikus projects ENavi (grant 03SFK4B0), Ariadne (grant 03SFK5C0), and Ariadne 2 (grant 03SFK5C0-2). Funding sources had no involvement with the study design, collection, analysis and interpretation of data, writing of the report and decision to submit the article for publication.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

I am highly grateful for valuable comments and suggestions by two anonymous reviewers as well as Fabian Castelli, Jana Eßer, Manuel Frondel, Philipp Hiemann, Eva Hümmecke, Kathrin Kaestner, David Schulze, Lukas Tomberg, Colin Vance, Madeline Werthschulte, and participants of the EAERE and VfS 2025 annual conferences and the RWI therapy research seminar. Furthermore, I thank Lisa Kruse, Nico Reichert, Ramona Wenzinger, and Julia Wittmann for valuable research assistance.

Appendix A. Behavior of forward-looking agents when the self-image constraint is active in $t = 1, 2$

When the self-image constraint is active in $t = 1, 2$, a forward-looking agent maximizes

$$\max_{(a_{i0},a_{i1})\in\{S,P\}^2} U = u(a_{i0}) + \frac{1}{1+r}u(a_{i1}) \text{ subject to}$$

$$\hat{\phi}_{i1} = \left(1-\delta^{\mathbb{1}[a_{i0}=S]}\right)\hat{v}(a_{i0}) \geq \bar{\phi}_i \text{ and}$$

$$\hat{\phi}_{i2} = \left(1-\epsilon^{\mathbb{1}[a_{i0}=S]}\right)\gamma\left(1-\delta^{\mathbb{1}[a_{i0}=S]}\right)\hat{v}(a_{i0}) + \left(1-\delta^{\mathbb{1}[a_{i1}=S]}\right)\hat{v}(a_{i1}) \geq \bar{\phi}_i.$$

Agents choose $(a_{i0}, a_{i1}) = (S, S)$ whenever this choice fulfills both of their self-image constraints. Since $\hat{\phi}_{i1}(S) = v'_L > \hat{\phi}_{i2}(S, S) = (1+\gamma')v'_L$, this is the case for agents with a self-image constraint $\bar{\phi}_i \leq (1+\gamma')v'_L$.

Agents with moral standards higher than $(1+\gamma')v'_L$ have to choose the prosocial action at least once. Because utility is higher when the selfish action is chosen in $t = 0$ rather than $t = 1$, individuals with $\bar{\phi}_i > (1+\gamma')v'_L$ choose $(a_{i0}, a_{i1}) = (S, P)$ whenever this fulfills both of their self-image constraints. Since $\hat{\phi}_{i1}(S) = v'_L < \hat{\phi}_{i2}(S, P) = \gamma' v'_L + v_H$, this is the case for agents with a self-image constraint $(1+\gamma')v'_L < \bar{\phi}_i \leq v'_L$.

$(a_{i0}, a_{i1}) = (P, S)$ yields a smaller overall utility and a smaller self-image in $t = 2$ than $(a_{i0}, a_{i1}) = (S, P)$, but a larger self-image in $t = 1$: $\hat{\phi}_{i1}(S) = v'_L < \hat{\phi}_{i1}(P) = v_H$ and $\hat{\phi}_{i2}(S, P) = \gamma' v'_L + v_H > \gamma v_H + v'_L = \hat{\phi}_{i2}(P, S)$. Therefore, and because $\gamma v_H + v'_L < v_H$, agents with moral standards $v'_L < \bar{\phi}_i \leq \gamma v_H + v'_L$ choose $(a_{i0}, a_{i1}) = (P, S)$. Finally, agents with moral standards above $\gamma v_H + v'_L$ can only fulfill their self-image constraints if they choose $(a_{i0}, a_{i1}) = (P, P)$.

Thus, when the self-image constraint is active in $t = 1, 2$, the threshold values for the self-image constraint, which determine agents' choice of (a_{i0}, a_{i1}) are exactly the same for forward-looking agents as for myopic agents (see Proposition 1).

Appendix B. Proof of Proposition 1 for $\delta, \epsilon \to 0$ and $\delta, \epsilon \to 1$

From Expressions (3) and (4), $\lim_{\delta \to 0} v'_L = v_L$, and $\lim_{\epsilon \to 0} \gamma'(\epsilon) = \gamma$, whereas $\lim_{\delta \to 1} v'_L = 0$, and $\lim_{\epsilon \to 1} \gamma'(\epsilon) = 0$. Thus, for extreme values of δ and ϵ Proposition 1 holds when substituting v'_L and γ' with their respective limits.

$\delta \to 1$ implies that $(1+\gamma')v'_L = v'_L = 0$. This means that no individual chooses $(a_{i0}, a_{i1}) = (S, P)$. This combination of actions cannot satisfy a self-image constraint $\bar{\phi}_i > 0$ in $t = 1$ and for $\bar{\phi}_i \leq 0$ it is strictly dominated by $(a_{i0}, a_{i1}) = (S, S)$ because $U(S, P) < U(S, S)$. □

Appendix C. Choice set for model version with active self-image constraint in the long run

In general, there are four possible combinations of selfish and prosocial actions:

$$(a_{i0}, a_{i1}) \in \Big\{(S, S), (S, P), (P, S), (P, P)\Big\}.$$

When deciding on their actions, in addition to material utility, individuals take into account the actions' implications for their self-image in $t = 2$.

For $(a_{i0}, a_{i1}) = (S, S)$ material utility is $U(S, S) = \frac{2+r}{1+r} u(S)$, and self-image in $t = 2$ is $\hat{\phi}_{i2}(S, S) = (1+\gamma')v'_L$.

For $(a_{i0}, a_{i1}) = (S, P)$ material utility is $U(S, P) = u(S) + \frac{1}{1+r} u(P)$, and self-image in $t = 2$ is $\hat{\phi}_{i2}(S, P) = \gamma' v'_L + v_H$.

For $(a_{i0}, a_{i1}) = (P, S)$ material utility is $U(P, S) = u(P) + \frac{1}{1+r} u(S)$, and self-image in $t = 2$ is $\hat{\phi}_{i2}(P, S) = \gamma v_H + v'_L$.

For $(a_{i0}, a_{i1}) = (P, P)$ material utility is $U(P, P) = \frac{2+r}{1+r} u(P)$, and self-image in $t = 2$ is $\hat{\phi}_{i2}(P, P) = (1+\gamma)v_H$.

$(a_{i0}, a_{i1}) = (S, S)$ is not strictly dominated by any other option, because

$$U(S, S) > U(S, P) \text{ and } \hat{\phi}_{i2}(S, S) < \hat{\phi}_{i2}(S, P),$$

$$U(S, S) > U(P, S) \text{ and } \hat{\phi}_{i2}(S, S) < \hat{\phi}_{i2}(P, S), \text{ and}$$

$$U(S, S) > U(P, P) \text{ and } \hat{\phi}_{i2}(S, S) < \hat{\phi}_{i2}(P, P).$$

$(a_{i0}, a_{i1}) = (S, P)$ is not strictly dominated by any other option, because

$$U(S, P) < U(S, S) \text{ and } \hat{\phi}_{i2}(S, P) > \hat{\phi}_{i2}(S, S),$$

$$U(S, P) > U(P, S) \text{ and } \hat{\phi}_{i2}(S, P) > \hat{\phi}_{i2}(P, S), \text{ and}$$

$$U(S, P) > U(P, P) \text{ and } \hat{\phi}_{i2}(S, P) < \hat{\phi}_{i2}(P, P).$$

$(a_{i0}, a_{i1}) = (P, S)$ is strictly dominated by $(a_{i0}, a_{i1}) = (S, P)$, because

$$U(P, S) < U(S, P) \text{ and } \hat{\phi}_{i2}(P, S) < \hat{\phi}_{i2}(S, P).$$

$(a_{i0}, a_{i1}) = (P, P)$ is not strictly dominated by any other option, because

$$U(P, P) < U(S, S) \text{ and } \hat{\phi}_{i2}(P, P) > \hat{\phi}_{i2}(S, S),$$

$$U(P, P) < U(S, P) \text{ and } \hat{\phi}_{i2}(P, P) > \hat{\phi}_{i2}(S, P), \text{ and}$$

$$U(P, P) < U(P, S) \text{ and } \hat{\phi}_{i2}(P, P) > \hat{\phi}_{i2}(P, S).$$

Thus, plausible combinations of actions with an active self-image constraint in $t = 2$ are

$$(a_{i0}, a_{i1}) \in \Big\{(S, S), (S, P), (P, P)\Big\}.$$

Appendix D. Proof of Proposition 2 for $\delta, \epsilon \to 0$ and $\delta, \epsilon \to 1$

From Expressions (3) and (4), $\lim_{\delta \to 0} v'_L = v_L$, and $\lim_{\epsilon \to 0} \gamma'(\epsilon) = \gamma$, whereas $\lim_{\delta \to 1} v'_L = 0$, and $\lim_{\epsilon \to 1} \gamma'(\epsilon) = 0$. Thus, for extreme values of δ and ϵ Proposition 2 holds when substituting v'_L and γ' with their respective limits.

Appendix E. Post-mortem analysis of model calibration

This section provides an analysis of the potential causes for the failure of some model predictions in the calibration. While the data are consistent with the model for the scenario with an active self-image constraint in $t = 0, 1$, this is not the case for the scenario where the self-image constraint is only active in $t = 2$.

According to Proposition 2, when the self-image constraint is active only in the long run, only individuals with very high self-image constraints should act prosocially in $t = 0$, i. e. choose the train voucher, and individuals with intermediate self-image constraints should offset their emissions after choosing the flight voucher. Assuming that informing respondents about the offsetting possibility before voucher choice only removes the short-run self-image constraint without changing anything else and setting the model parameters as specified in Section 5.1, a large majority of 86.23 % should choose the flight voucher.

There are several likely explanations for why this may not hold empirically. Firstly, some respondents may be intrinsically motivated to always maintain a satisfactory level of self-image, i. e. their self-image constraint might be active even when they are aware of the second self-image relevant choice. Thus, in treatment condition $Informed$, the self-image constraint in $t = 1$ might be lifted only for a fraction α of respondents, while it is still active for the remaining fraction $1 - \alpha$.

Secondly, the information about the offsetting possibility may have additional effects other than just removing the short-run self-image constraint. Most likely, it increases awareness of the negative environmental impacts of flying, leading to a reduction in the scope for placing less weight on negative signals (i. e. δ). This means that the treatment would have two opposite effects: On the one hand, it leads to an increase in flying by providing an opportunity to offset negative externalities (see Proposition 3), and on the other hand, it leads to a decrease in flying by highlighting its negative externalities (see Proposition 5 a)).

Thirdly, for simplicity, I have assumed in the calibration that the self-image constraints are uniformly distributed. In reality, however, the distribution can take any form. If the density is higher for high self-image constraints, a larger share of respondents will opt for the train voucher even when the self-image constraint in $t = 1$ is lifted. In a similar vein, the empirical values for v_L, v_H, and γ may be different than assumed.

Fourthly, individuals are likely to be heterogeneous in reality. Although the model assumes that v_H, v_L, γ, δ and ϵ are the same for all individuals, individuals may actually differ substantially in the strength of their signals, their discounting and their tendency towards motivated reasoning (see, for example, Eßer et al., 2025). In addition, the model assumes that $u(S) > u(P)$. In the empirical study, some respondents may not have planned any trips for which flying would be a sensible option. Others may prefer to take the train, even when flying is an option – for example, because traveling by train is more convenient for certain routes, or because they are afraid of flying. All of this implies that a fraction β of respondents for whom $u(P) > u(S)$ would always opt for the train voucher, regardless of whether their self-image constraint is active in $t = 1, 2$ or only in $t = 2$. Finally, all the possible explanations discussed above may interact with each other.

Appendix F. Relevant survey questions

Question A13: Currently, students in many cities are protesting for more climate protection every Friday during and outside school hours. What is your attitude toward these so-called ”Fridays for Future” protests?

- (1) I don’t like them at all
- (2) I don’t like them
- (3) neither in favor nor don’t like them
- (4) I’m in favor
- (5) I’m totally in favor
- Don’t know/No answer

Question KOMP1: Do you plan to take a private trip within the next year using either the train or the plane? (Note: Please do not consider a trip where you would take a car (either as a driver or passenger) or a bus).

- Yes
- No
- Don’t know/ No answer

If respondent is in the treatment group FfF:
Question KOMP2: *Filtering into different groups depending on answer to Question A13.*
For participants who answered Question A13 with 1 or 2:
Would you agree that, overall, you do not support the "Fridays for Future" movement?

- Yes
- No
- Don’t know/ No answer

For participants who answered Question A13 with 3:
Would you agree that, overall, you have a neutral attitude toward the "Fridays for Future" movement?

- Yes
- No
- Don't know/ No answer

For participants who answered Question A13 with 4 or 5:
Would you agree that, overall, you support the "Fridays for Future" movement?

- Yes
- No
- Don't know/ No answer

If Group Info & KOMP1 = Yes:
In the following, you can win a travel voucher worth 40 euros for your upcoming trip. The winners will be selected randomly. One out of 50 respondents will receive a voucher. You can choose whether you would prefer to receive a flight voucher or a rail voucher for your trip, should you be one of the winners. Flight vouchers can be redeemed via the Flightgift website with more than 300 different airlines, while rail vouchers can be redeemed with *Deutsche Bahn AG*.
If you travel by air, you have the option of offsetting the CO2 emissions generated for a specified fee. If you decide to offset your CO2 emissions, this fee will be donated to the non-profit organization atmosfair. As a result, your flight voucher will be reduced by the amount of the CO2 offset.
If Group No Info & KOMP1 = Yes:
In the following, you can win a travel voucher worth 40 euros for your upcoming trip. The winners will be selected randomly. One out of 50 respondents will receive a voucher. You can choose whether you would prefer to receive a flight voucher or a rail voucher for your trip, should you be one of the winners. Flight vouchers can be redeemed via the Flightgift website with more than 300 different airlines, while rail vouchers can be redeemed with *Deutsche Bahn AG*.
If Group Info & KOMP1 = No:
Imagine that you are planning to take a private trip using either the train or the plane. In the following, you can win a travel voucher worth 40 euros for your upcoming trip. The winners will be selected randomly. One out of 50 respondents will receive a voucher. You can choose whether you would prefer to receive a flight voucher or a rail voucher for your trip, should you be one of the winners. Flight vouchers can be redeemed via the Flightgift website with more than 300 different airlines, while rail vouchers can be redeemed with *Deutsche Bahn AG*.
If you travel by air, you have the option of offsetting the CO2 emissions generated for a specified fee. If you decide to offset your CO2 emissions, this fee will be donated to the non-profit organization atmosfair. As a result, your flight voucher will be reduced by the amount of CO2 offset.
If Group No Info & KOMP1 = No:
Imagine that you are planning to take a private trip using either the train or the plane. In the following, you can win a travel voucher worth 40 euros for your upcoming trip. The winners will be selected randomly. One out of 50 respondents will receive a voucher. You can choose whether you would prefer to receive a flight voucher or a rail voucher for your trip, should you be one of the winners. Flight vouchers can be redeemed via the Flightgift website with more than 300 different airlines, while rail vouchers can be redeemed with *Deutsche Bahn AG*.
For everyone:
Question KOMP3: If you are among the winners, which voucher do you choose?

- Flight voucher worth 40 euros
- Train voucher worth 40 euros
- don't know / no answer

If Group No Info & Komp3 = Flight:
When traveling by air, you have the option of offsetting the CO2 emissions generated. If you decide to offset, the amount you choose will be donated to the non-profit organization atmosfair. Your flight voucher will be reduced by this amount.
If Komp3 = Flight:
Question KOMP3b: If you win a flight voucher: Would you like to pay a fee of 5/10/15 euros (randomized) to offset the emissions generated by your flight? Reminder: This amount will reduce the value of your voucher if you are one of the lucky winners of a voucher.

- yes
- no
- don't know / no answer

References

Andor, M.A., Goette, L., Price, M.K., Schulze Tilling, A., Tomberg, L., 2023. Differences in how and why social comparison and real-time feedback impact resource use: evidence from a field experiment. NBER Working Papers (31845). https://doi.org/10.3386/w31845
Andreoni, J., Rao, J.M., Trachtman, H., 2017. Avoiding the ask: a field experiment on altruism, empathy, and charitable giving. J. Pol. Econ. 125 (3), 625–653. https://doi.org/10.1086/691703

Bénabou, R., 2015. The economics of motivated beliefs. Revue d'économie politique (5), 665–685. https://doi.org/10.3917/redp.255.0665
Bénabou, R., Tirole, J., 2002. Self-confidence and personal motivation. Quart. J. Econ. 117, 817–915. https://doi.org/10.1162/003355302760193913
Bénabou, R., Tirole, J., 2004. Willpower and personal rules. J. Pol. Econ. 112 (4), 848–886. https://doi.org/10.1086/421167
Bénabou, R., Tirole, J., 2006. Incentives and prosocial behavior. Am. Econ. Rev. 96 (5), 1652–1678. https://doi.org/10.1257/aer.96.5.1652
Bénabou, R., Tirole, J., 2011. Identity, morals, and taboos: beliefs as assets. Q. J. Econ. 126 (2), 805–855. https://doi.org/10.1093/qje/qjr002
Bodner, R., Prelec, D., 2003. Self-signaling and diagnostic utility in everyday decision making. Psychol. Econ. Decis. 1 (105), 26. https://doi.org/10.1093/oso/9780199251063.003.0006
Brañas-Garza, P., Bucheli, M., Espinosa, M.P., García-Muñoz, T., 2013. Moral cleansing and moral licenses: experimental evidence. Econ. Phil. 29 (2), 199–212. https://doi.org/10.1017/S0266267113000199
Byrne, D.P., Goette, L., Martin, L.A., Miles, A., Jones, A., Schob, S., Staake, T., Tiefenbeck, V., 2023. How nudges create habits: theory and evidence from a field experiment. Available at SSRN 3974371. https://doi.org/10.2139/ssrn.3974371
Chew, S.H., Huang, W., Zhao, X., 2020. Motivated false memory. J. Pol. Econ. 128 (10), 3913–3939. https://doi.org/10.1086/709971
CNN, 2019. Teen activist tells Davos elite they're to blame for climate crisis. https://edition.cnn.com/2019/01/25/europe/greta-thunberg-davos-world-economic-forum-intl/index.html. Last access on July 15, 2021.
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. Organ. Behav. Hum. Decis. Process. 100 (2), 193–201. https://doi.org/10.1016/j.obhdp.2005.10.001
Druckman, J.N., McGrath, M.C., 2019. The evidence for motivated reasoning in climate change preference formation. Nat. Clim. Change 9 (2), 111–119. https://doi.org/10.1038/s41558-018-0360-1
Dubé, J.-P., Luo, X., Fang, Z., 2017. Self-signaling and prosocial behavior: a cause marketing experiment. Market. Sci. 36 (2), 161–186. https://doi.org/10.1287/mksc.2016.1012
Eßer, J., Flörchinger, D., Frondel, M., Sommer, S., 2025. Climate change and the covid-19 pandemic: testing for motivated reasoning. USAEE Working Paper No. 25-648. https://doi.org/10.2139/ssrn.5235597
Exley, C.L., 2016. Excusing selfishness in charitable giving: the role of risk. Rev. Econ. Stud. 83 (2), 587–628. https://doi.org/10.1093/restud/rdv051
Falk, A., 2021. Facing yourself–a note on self-image. J. Econ. Beh. Organ. 186, 724–734. https://doi.org/10.1016/j.jebo.2020.11.003
Fang, X., Goette, L., Rockenbach, B., Sutter, M., Tiefenbeck, V., Schoeb, S., Staake, T., 2023. Complementarities in behavioral interventions: evidence from a field experiment on resource conservation. J. Public Econ. 228, 105028. https://doi.org/10.1016/j.jpubeco.2023.105028
Flörchinger, D., Frondel, M., Sommer, S., Andor, M., 2025. Pro-environmental behavior and environmentalist movements: evidence from the identification with Fridays for Future. Ecol. Econ. 235, 108609. https://doi.org/10.1016/j.ecolecon.2025.108609
Fridays for Future, 2022. Who we are. https://fridaysforfuture.org/what-we-do/who-we-are/. Last access on January 26, 2022.
Gneezy, A., Gneezy, U., Riener, G., Nelson, L.D., 2012. Pay-what-you-want, identity, and self-signaling in markets. Proc. Nat. Acad. Sci. 109 (19), 7236–7240. https://doi.org/10.1073/pnas.1120893109
Gneezy, U., Imas, A., Madarász, K., 2014. Conscience accounting: emotion dynamics and social behavior. Manage. Sci. 60 (11), 2645–2658. https://doi.org/10.1287/mnsc.2014.1942
Grossman, Z., 2014. Strategic ignorance and the robustness of social preferences. Manage. Sci. 60 (11), 2659–2665. https://doi.org/10.1287/mnsc.2014.1989
Grossman, Z., van der Weele, J.J., 2017. Self-image and willful ignorance in social decisions. J. Eur. Econ. Assoc. 15 (1), 173–217. https://doi.org/10.1093/jeea/jvw001
Ho, B., Taber, J., Poe, G., Bento, A., 2016. The effects of moral licensing and moral cleansing in contingent valuation and laboratory experiments on the demand to reduce externalities. Environ. Resour. Econ. 64 (2), 317–340. https://doi.org/10.1007/s10640-014-9872-y
Jordan, J., Mullen, E., Murnighan, J.K., 2011. Striving for the moral self: the effects of recalling past moral actions on future moral behavior. Pers. Soc. Psychol. Bull. 37 (5), 701–713. https://doi.org/10.1177/0146167211400208
Kahan, D.M., 2013. Ideology, motivated reasoning, and cognitive reflection. Judgement Decis. Mak. 8 (4), 407–424. https://doi.org/10.1017/S1930297500005271
Khan, U., Dhar, R., 2007. Where there is a way, is there a will? The effect of future choices on self-control. J. Experim. Psychol. Gen. 136 (2), 277. https://doi.org/10.1037/0096-3445.136.2.277
Kunda, Z., 1990. The case for motivated reasoning. Psychol. Bull. 108 (3), 480. https://doi.org/10.1037/0033-2909.108.3.480
List, J.A., Momeni, F., 2017. When corporate social responsibility backfires: theory and evidence from a natural field experiment. NBER Working Paper (w24169). https://doi.org/10.3386/w24169
List, J.A., Momeni, F., 2021. When corporate social responsibility backfires: evidence from a natural field experiment. Manage. Sci. 67 (1), 8–21. https://doi.org/10.1287/mnsc.2019.3540
Matejko, L., Sommer, S., 2022. The social sustainability barometer 2017–2019: survey data on the Energiewende. Jahrbücher für Nationalökonomie und Statistik 242 (4), 529–545. https://doi.org/10.1515/jbnst-2021-0049
Mazar, N., Zhong, C.-B., 2010. Do green products make us better people? Psychol. Sci. 21 (4), 494–498. https://doi.org/10.1177/0956797610363538
Merritt, A.C., Effron, D.A., Fein, S., Savitsky, K.K., Tuller, D.M., Monin, B., 2012. The strategic pursuit of moral credentials. J. Exp. Soc. Psychol. 48 (3), 774–777. https://doi.org/10.1016/j.jesp.2011.12.017
Merritt, A.C., Effron, D.A., Monin, B., 2010. Moral self-licensing: when being good frees us to be bad. Soc. Personal Psychol. Compass 4 (5), 344–357. https://doi.org/10.1111/j.1751-9004.2010.00263.x
Möbius, M.M., Niederle, M., Niehaus, P., Rosenblat, T.S., 2022. Managing self-confidence: theory and experimental evidence. Manage. Sci. 68 (11), 7793–7817. https://doi.org/10.1287/mnsc.2021.4294
Momsen, K., Ohndorf, M., 2022. Information avoidance, selective exposure, and fake (?) news: theory and experimental evidence on green consumption. J. Econ. Psychol. 88, 102457. https://doi.org/https://doi.org/10.1016/j.joep.2021.102457
Monin, B., Miller, D.T., 2001. Moral credentials and the expression of prejudice. J. Pers. Soc. Psychol. 81 (1), 33. https://doi.org/10.1037/0022-3514.81.1.33
Munro, G.D., Stansbury, J.A., 2009. The dark side of self-affirmation: confirmation bias and illusory correlation in response to threatening information. Pers. Soc. Psychol. Bull. 35 (9), 1143–1153. https://doi.org/10.1177/0146167209337163
Neuber, M., Gardner, B.G., 2020. Germany. In: de Moor, J., Uba, K., Wahlström, M., Wennerhag, M., de Vydt, M. (Eds.), Protest for a Future II: Composition, Mobilization and Motives of the Participants in Fridays for Future Climate Protests on 20–27 September, 2019, in 19 Cities around the World. Retrieved from: https://www.diva-portal.org/smash/get/diva2:1397070/FULLTEXT01.pdf. Last access on November 21, 2024.
Nisan, M., Horenczyk, G., 1990. Moral balance: the effect of prior behaviour on decision in moral conflict. Brit. J. Soc. Psychol. 29 (1), 29–42. https://doi.org/10.1111/j.2044-8309.1990.tb00884.x
Ploner, M., Regner, T., 2013. Self-image and moral balancing: an experimental analysis. J. Econ. Behav. Organ. 93, 374–383. https://doi.org/10.1016/j.jebo.2013.03.030
Rabin, M., 1995. Moral preferences, moral constraints, and self-serving biases Available at https://scholar.harvard.edu/files/rabin/files/moral_preferences.pdf.
Read, D., Loewenstein, G., Rabin, M., 1999. Choice bracketing. J. Risk Uncertain. 19 (1-3), 171–197. https://doi.org/10.1023/A:1007879411489
Sachdeva, S., Iliev, R., Medin, D.L., 2009. Sinning saints and saintly sinners: the paradox of moral self-regulation. Psychol. Sci. 20 (4), 523–528. https://doi.org/10.1111/j.1467-9280.2009.02326.x
Tiefenbeck, V., Goette, L., Degen, K., Tasic, V., Fleisch, E., Lalive, R., Staake, T., 2018. Overcoming salience bias: how real-time feedback fosters resource conservation. Manage. Sci. 64 (3), 1458–1476. https://doi.org/10.1287/mnsc.2016.2646
Tiefenbeck, V., Wörner, A., Schöb, S., Fleisch, E., Staake, T., 2019. Real-time feedback reduces energy consumption among the broader public without financial incentives. Nat. Energy 4 (10), 831–832. https://doi.org/10.1038/s41560-019-0480-5
Zimmermann, F., 2020. The dynamics of motivated beliefs. Am. Econ. Rev. 110 (2), 337–363. https://doi.org/10.1257/aer.20180728