Identity as Self-Image

Roland Bénabou Luca Henkel

Preliminary version

August 19, 2025

Abstract

We review the economic literature on self-image, which conceptualizes identity as a set of beliefs about one's core traits, values, goals, and social ties. Self-image concerns lead individuals to process information and make choices in non-standard ways that help affirm and protect certain valued identities. We first present the main cognitive mechanisms involved within a simple unifying framework. We then survey the extensive laboratory, online, and field experimental literature on the nature and behavioral implications of self-image concerns. We discuss in particular how they give rise to information and decision avoidance, motivated memory and beliefs, excuse-driven behavior, preferences for truth-telling, hypothetical bias, moral cleansing and moral licensing, collective identities, political preferences, and other forms of self-signaling or self-deception. We subsequently discuss common empirical strategies used to identify self-image concerns, as well as the threats to their validity and how to alleviate them. We conclude by outlining open questions and directions for future research on the belief-based approach to identity.

Contact: Roland Bénabou; Princeton University, NBER, CEPR, IZA, BREAD, and THRED, rbenabou@princeton.edu. Luca Henkel; Erasmus University Rotterdam, ERIM, CESifo, JILAEE, IZA, Tinbergen Institute; henkel@ese.eur.nl.
Acknowledgements: Adam Rangel provided excellent research assistance.

"An identity is a definition, an interpretation, of the self (...) People who have problems with identity are generally struggling with the difficult aspects of defining the self, such as the establishing of long-term goals, major affiliations, and basic values." (Baumeister, 2022).

1 Introduction

One can broadly distinguish two complementary approaches to the study of identity in economics. The first one is based on preferences and emphasizes the social aspects of identity. The second one is based on beliefs about the self and emphasizes the cognitive aspects of identity.

In the line of work pioneered by Akerlof and Kranton (2000, 2005, 2010), an identity is modeled as a set of payoffs and internalized social norms attached to membership in a group. Individuals gain or lose material and symbolic payoffs, such as social status or self-esteem, when identified with some group, and incur psychic losses when their actions or personal characteristics deviate from those prescribed by the group's norm.[1] Identities can be exogenously assigned by nature or society (gender, race), manipulated by some principal (firms, the military, politicians, experimenters), or determined by previous actions of the individual (membership in a political party or religion). Alternatively, an individual may themselves choose to identify with one group or the other, based on the associated material and psychological costs (Shayo, 2009, 2020). In that case, they are essentially choosing their preferences from a socially determined set of possibilities.[2] The psychological foundations of this approach draw on the vast classical literature on conformity (Asch, 1955; Cialdini and Goldstein, 2004), norm formation (Sherif, 1935), self-categorization and ingroup-outgroup rivalry (Tajfel, 1978; Tajfel and Turner, 1979).

This preference-based modeling framework is highly versatile and has been applied to

[1] Relatedly, in Brekke, Kverndokk, and Nyborg (2003)'s model of moral motivation, an individual experiences a psychic cost (interpreted as a loss in self-esteem) to the extent that their action differs from the "universalizing" Kantian norm or rule, namely the counterfactual action which, if it were taken by everyone, would maximize social welfare.

[2] A related approach is that of Bisin and Verdier (2001), in which it is parents who choose the preferences of their children, trading off a desire for similarity to their own tastes against how adapted the inculcated preferences will be to the society in which their offspring will live.

a wide range of domains, many of which are reviewed in other chapters of this volume: gender roles (Bertrand, Kamenica, and Pan, 2015), workplace relations (Akerlof and Kranton, 2005), redistributive politics (Shayo, 2009; Bonomi, Gennaioli, and Tabellini, 2021), attitudes toward immigration (Grossman and Helpman, 2020), labor supply decisions (Oh, 2023) and ethnic conflict (Shayo, 2020), among others.

In the second approach, introduced by Bénabou and Tirole (2002, 2004, 2011), an identity is modeled as a self-image: a set of beliefs about one's core traits, values, long-term goals, and social ties. Psychologically, this view builds on the likewise vast literature on "the self" (Baumeister, 2022), which encompasses work on self-perception and self-awareness, self-esteem, cognitive dissonance, and self-deception. Self-perception theory, in particular, emphasizes that people's true identities (attitudes, preferences, emotions) are not directly accessible or transparent to them, but instead inferred by observing their own behavior and its context (Bem, 1972). In self-awareness theory (Duval and Wicklund, 1972), increased attention to the self raises the motivation to behave consistently with one's standards of behavior; if this is too costly, attention will instead be shifted away (see Morin (2011) and Smári, Ólason, and Ólafsson (2008) for reviews). Relatedly, when an inconsistency arises between a deeply held self-concept and what the evidence points to, cognitive dissonance results (e.g., Aronson, 1969); people then typically resolve it in favor of their existing self-image, through various forms of motivated reasoning and self-deception (Kunda, 1990).[3]

Indeed, certain beliefs about oneself and one's place in the world can be more desirable to hold than fully objective ones, a point first made in economics by Akerlof and Dickens (1982) and Akerlof (1989) and reviewed extensively in Bénabou and Tirole (2016). Positive views of one's abilities, social worth, or endowments (professional, social, cultural) enhance self-esteem and hopefulness, foster a sense of belonging, and lower anxiety, thus providing hedonic benefits (Loewenstein, 1987; Caplin and Leahy, 2001; Köszegi, 2006; Bénabou and Tirole, 2011). Optimistic beliefs can also be functionally adaptive: they serve as mental commitment devices encouraging effort, resilience in the face of adversity, or stronger dedication to a cause (Bénabou and Tirole, 2002, 2004; Borisova and Vellodi, 2024). Another

[3]In recent years, work in neuroscience on self-referential processing and active inference has begun to provide new foundations for these theories (Friston, 2010; Sharot and Garrett, 2016).

instrumental benefit is that being convinced that one has such qualities makes it easier to convince others that one possesses them (see Hippel and Trivers (2011) for an evolutionary explanation).

These non-standard roles of beliefs have important informational and behavioral consequences bearing on identity. External feedback provides the individual with information about their type (ability, resilience, generosity, etc.), as does reflecting upon their own actions – a self-inference process akin to that of social reputation. Consequently, people will process information and make choices in ways that help affirm and defend their valued identities. In simplified terms, in the first type of model, agents choose their preferences, whereas in the second, they choose their beliefs, subject in both cases to reality constraints, cost-benefit tradeoffs, and potential manipulation by others.

Because it is explicitly cognitive, the identity-as-self-image modeling framework can account for many phenomena, reviewed in this chapter, that cannot arise when individuals have full knowledge of their preferences – whether standard or "behavioral." Among those that have been formally modeled are information avoidance and ego-protecting task choices (Carrillo and Mariotti, 2000; Köszegi, 2006); selective memory or awareness and asymmetric updating of beliefs to good and bad news (Bénabou and Tirole, 2002; Chew, Huang, and Zhao, 2020); and self-signaling behaviors –taking actions intended to convince or reassure oneself about one's type– that can lead to rigid personal rules, endowment effects, and escalating commitments (Bodner and Prelec, 2003; Bénabou and Tirole, 2004, 2011; Mijović-Prelec and Prelec, 2010). Closely related behaviors flowing from this framework include directionally motivated "errors" in simple computations or inferences (González-Jiménez, 2022; Bolte and Fan, 2023; Exley and Kessler, 2024) and self-serving failures of Bayesian skepticism, achieved through selective depth of reasoning about good and bad news (Taber and Lodge, 2006; Hagenbach and Saucet, 2025). Further distinctive implications include the crowding of intrinsic motivation by extrinsic incentives (Bénabou and Tirole, 2006), moral cleansing or licensing, and mental taboos (Bénabou and Tirole, 2011; Fershtman, Gneezy, and Hoffman, 2011; Hong, Tirole, and Zhang, 2025).

Imperfect self-knowledge also combines with selective recall in Kőszegi, Loewenstein,

and Murooka (2022) to generate fragile, history-dependent self-esteem. In Heidhues, Kőszegi, and Strack (2023) and Hestermann and Le Yaouanq (2021), dogmatic priors about one's abilities or preferences (possibly a strong form of motivated beliefs), or even just incorrect initial priors, induce persistent attribution errors: taking credit for successes while blaming external factors for failures,[4] persistent misunderstanding of one's own motives (true identity), and decision mistakes that can worsen over time or trap the individual in situations that it would be optimal to quit.[5]

Another important feature of self-image concerns is that, since they are inherently inward-looking, they can shape behavior even in private decisions such as charitable giving (e.g., Exley, 2016), investment choices (Henkel and Zimpelmann, 2025), or modify responses in anonymous surveys (Bursztyn et al., 2025). They also need not always be anchored in group identification or social categorizations, but can also pertain to personal dimensions of identity, such as being honest, competent, or disciplined.

While the two modeling approaches highlight different facets of identity, they have strong connections to each other, as should be expected from the significant overlap in the psychology literatures on which they draw. First, what beliefs are sources of self-worth is in part socially determined, like preferences in the first type of model. Strength and courage, intelligence, or patience are valued differently across settings: army versus academia, blue versus white collar jobs, etc. Even notions of morality vary significantly across societies and periods. Examples include differing "moral foundations" (Graham et al., 2011, 2013), universalism versus parochialism (Enke, Rodríguez-Padilla, and Zimmermann, 2023; Cappelen, Enke, and Tungodden, 2025), and consequentialist versus deontological ethics (Alger and Weibull, 2013; Bénabou, Falk, and Henkel, 2024). Second, self-image is shaped both by individual experiences and by the actual or imagined judgment of others. This is what Adam Smith (1759) termed "the ideal man within the breast, (...) the impartial spectator", Cooley (1902) the "looking glass self", and Leary and Downs (1995) the "sociometer". Third,

[4]On scapegoating, see also Bénabou and Tirole (2009) in the context of bargaining breakdowns and conflicts, such as those documented by Babcock and Loewenstein (1997).

[5]In Santos-Pinto and Sobel (2005), agents have heterogeneous production functions mapping multiple skills into general ability. They invest in skills to maximize ability, then make social comparisons by (dogmatically) projecting their own production function onto others, resulting in an inflated self-image.

people selectively interact with similar individuals (homophily) whose behaviors and judgments will likely provide identity-confirming signals, while shunning and ostracizing those whose identity-dissonant opinions or norm-breaking actions could threaten their self-view (Bénabou and Tirole, 2011).

Conversely, the strength of the internalized norms and the degree of identification with a group emphasized in the first type of model are often subject to substantial cognitive uncertainty, as is the case for many other aspects of a person's preferences (Enke and Graeber, 2023). People are commonly conflicted about how much they would really sacrifice for a group's cause or adherence to its norms and duties, how important their work is versus family identity, their attachment to ancestral traditions versus embracing modern ways, their masculine or feminine identity, or their religious identity. In fact, it is often those whose identity is most insecure (adolescents, new religious or political converts, recent immigrants) who are the most zealous in affirming it. Once imperfect self-knowledge enters the picture, the role of beliefs and potentially motivated or misspecified learning about oneself emphasized by the second type of model comes into play.

Another important feature shared by preference-based and belief-based models is the potential to generate multiple social norms. These can arise both within a group, due to conformity payoffs in the first case or pooling on certain actions in the second, and in the way a population distributes itself across multiple identities, as these choices generate externalities in either preferences (Shayo, 2009) or information (Battaglini, Bénabou, and Tirole, 2005; Bénabou and Tirole, 2011).

In the remainder of this chapter, we will focus on identity as self-image. We start in Section 2 by presenting a simple theoretical framework that synthesizes the main modes of identity-management that arise when self-knowledge is imperfect. We then survey the large experimental literature that has arisen in this area over the last two decades. In Section 3, we thus review the evidence documenting many of the behaviors that can be naturally explained by self-image motives and are not easily explainable otherwise. Many applications have been in the two domains of ability and morality, where self-serving beliefs are ubiquitous. Others include attractiveness, social status, greed, and political or religious identity.

In Section 4, we move from individual self-image to the nascent literature on the cognitive foundations of collective beliefs and identity. In Section 5, we discuss common methods the literature has used to identify self-image motives, connecting papers across domains. Lastly, in Section 6, we discuss potential directions for future work.

Before proceeding, we note other review articles related to this chapter. Bénabou (2015) and Bénabou and Tirole (2016) provide broad reviews of the economics of motivated reasoning and cognition. Golman, Hagmann, and Loewenstein (2017) provide an extensive survey of the literature on information avoidance, Amelio and Zimmermann (2023) of that on motivated memory, and Vu et al. (2023) a meta-analysis of studies on willful ignorance. Most closely related, and at times overlapping, is a concurrent survey by Lindquist, Saccardo, and Serra-Garcia (2025) that focuses on the specific implications of belief management for unethical behavior.

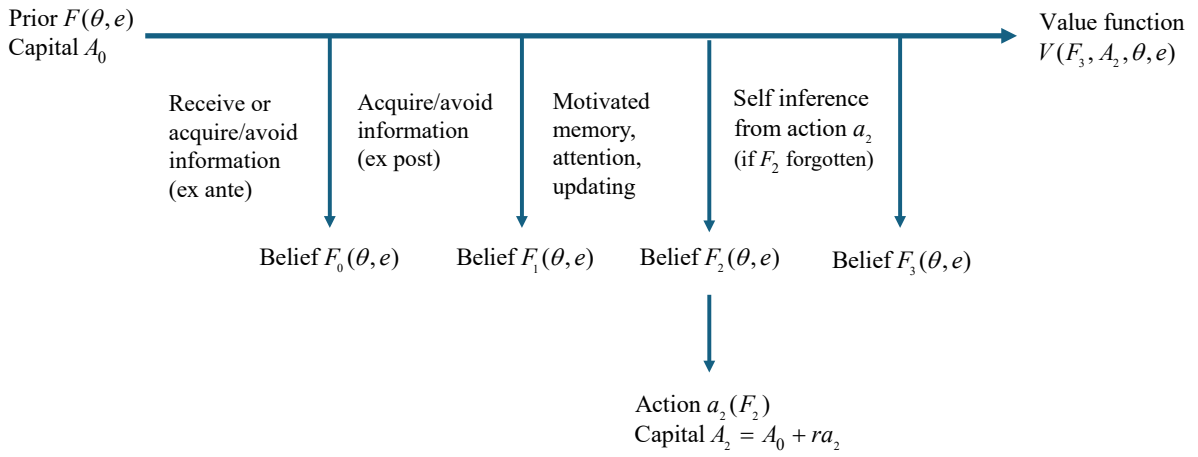
2 A simple unifying framework

In this section, we explain the main *identity-management* behaviors mentioned earlier and documented in the experimental literature we will review. Most papers focus on one or at most two, but we present them in a unified framework, illustrated in Figure 1, to make clear their similarities, differences, and complementarities. In each case, we provide examples and mention some relevant papers, revisited in more detail later on.

We consider an individual who acts and/or thinks during four periods, $t = 0, \ldots 3$. From then on, their expected present value of utility over all $t \geq 3$ is given by a value function V that is detailed below.

The individual is characterized by a personal parameter θ, such as ability, resilience, prosociality/morality, how much they value their relationship or belonging to some group, culture, or ideal (e.g., religion), or conversely how valued they are by their peers. We can think of θ as representing "who the person truly is". Self-knowledge is notoriously imperfect, so at a point in time, they may not know θ precisely, but only have a belief distribution over it. We define this belief as the agent's *self-image,* or *sense of identity.* While preferences can

Figure 1: **Modes of identity management**



Notes: Individual's full type: (θ = true preference or ability, e = externality on others). Beliefs at time t: $F_t\ (\theta, e)$, with self-image being the marginal $I_t(\theta)$. A_t: identity-congruent assets such as personal, social, or cultural capital.

of course change over time, we keep θ fixed to emphasize the dynamics of self-beliefs and contrast this with identity models based on preferences.

Depending on the setting, the individual's actions can have consequences that they may also not fully know, such as their externalities on others. We denote these situational variables as e. Initially, the agent has a prior distribution $F(\theta, e)$ over "the person and the situation" (Ross and Nisbett, 1991), which we refer to as the agent's full *type*. In later periods $t = 0, ...3$, their updated belief is denoted $F_t(\theta, e)$. This belief will often be separable in θ and e but it need not, for instance, if there is self-selection of agents into different settings. We will denote $I_t(\theta)$ the marginal distribution over θ, which is the individual's self-image at time t.

The agent starts with an initial endowment of *identity-relevant capital* A_0, which may be fixed (race, sex) or augmentable: human or professional capital, number of children or number of ties with some social group, knowledge of a culture, record of good versus bad deeds, etc. In period 2, the agent may take, at some cost, an action a_2 that potentially increases the stock to $A_2 = A_0 + ra_2$, with $r \geq 0$. In many of the papers discussed below, this will be a dictator-type allocation or other prosocial decision, but other examples include labor/family time budgeting, investing in new social ties, choice of clothing or food, conforming to group norms or religious prescriptions, etc.

During periods $0, \ldots 3$, the agent engages in various learning, remembering, and other cognitive operations that are described below, arriving at a final self-image $I_3(\theta)$, and more generally a final joint belief about themselves and their environment $F_3(\theta, e)$. From there on, the expected present value of their future utility, including from potential future decisions, is a functional $V(F_3, \theta, e, A_2)$. The dependence on the actual type (θ, e) and asset stock A_2 is standard, so we focus here on the dependence on the belief. In particular, we assume that V increases with the belief about θ (in the FOSD sense), at least over the relevant range: the agent cares about their self-image, and therefore will act and reason in ways that seek to preserve, affirm or enhance it.

As discussed earlier, the sources of such utility from self-image can be hedonic or functional, with two main variants in each case.

1. Pure, intrinsic *self-esteem*: $V = sf(I_3)$, where s is a parameter measuring salience and f is an increasing functional that can be linear, concave or convex, potentially creating a motive for information aversion or information seeking. For instance, the agent values seeing themselves as a smart, generous, honest, masculine or feminine person, *per se*.

2. *Anticipatory utility* over the future benefits to be derived from one's personal and social assets: $V = W(f(I_3)A_2, \theta, e)$. The complementarity between I and A_2 makes maintaining a positive self-image (high F_3) all the more important, the higher the level of the identity-congruent asset. For instance: (i) being more altruistic or reciprocal toward some group (national, cultural, ethnic) will be more valuable, the more social ties one has with it, or the tighter those are; (ii) being talented or intrinsically motivated for some activity (education, craft, military service) will more important, the more opportunities one has or will have to engage in it; (iii) being a family-loving person will be more valuable, the more relatives one has; (iv) having a strong faith in some religion involving afterlife rewards and punishments (heaven and hell, karma) will be more reassuring, the more one has followed its precepts, etc. Consequently, in each of these cases, seeing oneself as having more of the relevant attribute will raise anticipatory utility.

Conversely, the complementarity makes investing further in those identity-congruent assets (increasing A_2 via a_2) all the more desirable, the stronger is one's sense of that identity. Importantly, the combination of these two mechanisms can lead to "escalating commitments": over time, the individual simultaneously invests more and more in some asset and develops a stronger and stronger sense of the identity associated with it (e.g., workaholism and professional identity, compulsive rituals and religious identity, activism and political identity). Moreover, as suggested by the term, such escalations can be "identity traps" that actually lower the agent's intertemporal welfare (a form of hedonic treadmill, see Bénabou and Tirole (2011)).

3. An *instrumental value of self-confidence* or identification with one's work (research, firm's mission, serving one's country), which acts as an internal commitment device inducing effort and perseverance for an individual with imperfect willpower, $W = sV(I_3, \beta, \theta, e, A_2)$.

4. *Persuasiveness to others.* Alternatively, but also instrumental, V may represent the value of such beliefs emanating from the fact that it is easier to convince others of one's qualities or likelihood of success if one believes in them. Quite often, as Arendt (1972) observed, "The deceivers began by deceiving themselves."

With these elements in place, we now discuss the main cognitive strategies that people use (consciously or not) to maintain and invest in a valued identity, and some of their consequences.

Ex-ante information seeking or avoidance. In period 0, the agent may receive a signal σ_0 about their type (θ, e) according to some information structure, either exogenous or chosen by them ex ante, at some cost. The resulting belief, denoted F_0, is a (weakly) mean-preserving spread of the initial F. Of particular interest is the case where the agent pays or foregoes some gain in order not to learn about their true type. The purpose of such ex-ante willful ignorance, which can occur when the value function V is concave in beliefs, is to preserve a relatively good self-image rather than endanger it by learning possibly bad news about oneself, or the tradeoffs one is facing. The most often cited examples are that of a

patient who avoids taking a medical test and, in the context of self-image, the paradigmatic experiment of Dana, Weber, and Kuang (2007) in which subjects choose to remain ignorant of how their choices affect someone else's payoff. As we will argue, however, the latter more plausibly corresponds to the next category of information aversion.

Ex-post, motivated information seeking or avoidance. Most often, people already know something about their type –from external sources, previous experience, or even just a first, fleeting but "symptomatic" sense of how they would act if confronted with some situation. The agent may then have the opportunity to acquire a further signal σ_1, or once again pay to avoid information. Importantly, this decision now depends on their belief F_0, and hence on their current sense of identity. Accordingly, in Eil and Rao (2011), subjects who received bad news about their relative IQ have a negative willingness to pay for fully learning it, whereas it is positive for those who received good news. Similarly, in Müller et al. (2022), individuals who have exceeded the desired number of children they stated ten years earlier are willing to pay not to be reminded of what it was. In Dana, Weber, and Kuang (2007) there is no previous signal within the context of the experiment, but it is most likely those who know or suspect they would make a selfish choice in a dictator game who prefer not to learn whether their payoffs and those of the recipient are positively or negatively correlated. And indeed, almost all of them subsequently make the self-serving choice.[6]

Motivated memory, attention, or awareness. During period 2, which one should think of as lasting for a while, the agent may miscode, forget, or misremember the various signal(s) they received previously. In particular, since V is increasing in self-image, they will tend to selectively recall good news and forget bad ones, resulting in positively distorted memories $(\tilde{\sigma}_0, \tilde{\sigma}_1)$. Examples of such motivated recall include Zimmermann (2020) and Chew, Huang, and Zhao (2020) for subjects' IQ, Huffman, Raymond, and Shvets (2022) for managers' performance, and Müller et al. (2022) for fertility preferences. People also often have motivated memories of their past actions, since, as discussed below, those can be informative about who

[6]An important difference between ex-ante and ex-post information avoidance is that the former involves a simple optimization by an uninformed Self 0 of the agent, whereas the latter involves a game of strategic information transmission between an informed Self 0 and subsequent selves, potentially resulting in multiple personal equilibria. Thus, both a mode of "positive thinking" and one of "know thyself" can be sustainable.

they are. We thus all tend to remember better the instances when we acted morally than those where we behaved selfishly (e.g., Saucet and Villeval, 2019; Carlson et al., 2020).

Motivated belief updating. Together with the signals received, these and related motivated-reasoning processes result in an updated belief $F_2(\theta, e)$, and especially an updated self-image $I_2(\theta)$ at the end of period 1, that can depart from objectivity. If the agent fully understands their own tendencies to selectively acquire and process information (*metacognition)* and they are a sophisticated Bayesian, F_2 cannot be biased on average relative to F, but it can be made more or less informative, or more or less skewed, which again can be valuable depending on the properties of the intertemporal utility V with respect to beliefs. If, however, the agent is *naïve* and takes their recollection $(\tilde{\sigma}_0, \tilde{\sigma}_1)$ partly at face value, or similarly neglects the fact that the decision to acquire or avoid σ_1 depends on σ_0 and is thus informative about it –e.g., choosing ignorance about externalities is a bad sign concerning prosociality– then F_2 can be biased on average as well.

There is good evidence for at least partial naivete with respect to self-deception: agents who display self-serving recall do end up with higher mean beliefs about their IQ or professional ability (see earlier cites, and others in Section 3.2), and those who hide behind fairly transparent excuses do end up feeling less responsible for selfish outcomes they purposefully brought about (Hamman, Loewenstein, and Weber, 2010). Similarly, in several studies, agents effectively misapply Bayes' rule in an asymmetric, self-serving manner, resulting again in inflated self-views on average (e.g., Eil and Rao (2011) and Möbius et al. (2022)).

Identity-congruent choices: effect of identity on behavior. At the end of period 2, given their belief F_2, the individual chooses a_2: how much to give in a dictator game, how much to work, how well to conform to their group's norm, etc. If F_2 has been distorted by information avoidance or motivated recall/updating, the decision is also distorted. Depending on whether the value of the distortion is hedonic or instrumental, this can be harmful or beneficial. In the first case, an overconfident agent may undertake tasks that are too hard, one who denies their homosexual desires may choose partners of the wrong sex for them, one

who overestimates their altruism may give more than what really makes them happy, and one who inflates (or insufficiently deflates) their belief in some religion may over-comply with its precepts and rituals. In the second, an agent's greater self-confidence may increase their motivation for effort and perseverance, or their ability to favorably impress others. Note that, conversely, the choice of a_2 is informative about F_1, and in particular potentially "diagnostic" about the agent's personal type θ, a point we return to below.

Following their decision a_2, during period 3, the agent may again forget or misremember (exogenously or endogenously) their previous state of knowledge F_2 (or the signals that led to it) and the associated motivation that really determined their choice of a_2. They could also selectively remember a_2 itself, such as mostly recalling cases where one acted prosocially, as discussed earlier.

Self-signaling: effect of behavior on identity. At the start of period 3, if the individual has perfect introspective access to (e.g., recollection of) their previous belief F_2 or the signals (σ_0, σ_1) that led to it, their earlier choice of a_2 conveys no extra information, so their final belief remains $F_3 = F_2$. If, on the other hand, there is a loss of information during period 3, the action a_2 (when recalled) is informative and will affect F_3. Anticipating this in period 2, the agent has an incentive to choose a_2 in a way that not only reflects its cost and effect on A_2, but is also designed to affect the belief F_3 in a desirable direction. That is, *the action is used as a signal* (to future selves) of what kind of a person one is. For example, selfish types could pool with more generous ones by behaving as altruistically as them, or generous types could behave even more generously than they would absent identity concerns to separate from selfish ones. More generally, self-signaling creates a further force toward identity-congruent investments and escalating commitments of the types discussed above: the agent now also engages in these behaviors to affirm and reassure themselves about their identity.

Mental taboos. In some cases, acquiring information at the start of period 1 is itself bad news about the self: simply pondering the costs and benefits of disloyalty to the group, or those of certain morally "repugnant transactions", or merely entertaining questions about the tenets of one's faith, can be a signal of a low θ, degrading one's self-view or "sacred

values"(Bénabou and Tirole, 2011; Gneezy and Rustichini, 2011). The individual will then forbid themselves from certain *thoughts,* or even find those deeply aversive.

3 Behavioral manifestations of self-image

We now review the evidence on some of the main behaviors discussed above as arising from imperfect self-knowledge and self-image concerns. In each case, we describe the first influential studies and explain why their findings challenge both standard theory and common behavioral models, but can be naturally explained by self-image motivations. We then discuss how the literature on each of these behaviors has evolved since the initial studies.

3.1 Avoiding information and choice situations

In non-strategic settings, standard economic agents always weakly prefer more information to less, and strictly so if it is decision-relevant, helping them make better choices. Similarly, they prefer making decisions to avoiding them and letting some default prevail.[7] In contrast, and different also from preference-based models of identity, an individual who seeks to maintain a certain self-view may want to avoid information that could threaten it, as well as decision situations in which their choice could reveal an undesirable type.[8]

For instance, consider a person who is generally self-interested but would like to think of themselves as altruistic, and encounters an appeal by a charity to donate money for a good cause. An altruistic individual would gladly give if the charity uses the money efficiently and this leads to a significant improvement in other people's lives, but not if the organization is ineffective or corrupt. The charity, or some rating organization, makes credible information on its operations and impact easily available. A person who feels ungenerous when encountering the appeal, however, has an incentive to avoid looking at such information. Remaining

[7]Agents with self-control problems may choose commitment over discretion. They may also avoid information, but this is isomorphic to behavior driven by self-image concerns in which they prefer not to learn about their own type; see Carrillo and Mariotti (2000). Choice overload or mental decision costs may also explain some of these behaviors, but this is implausible in the very simple settings of the literature surveyed below, and cannot account for why agents would pay not to be informed even when the information is not decision-relevant.

[8]Self-image concerns are admittedly not the only reason for individuals to avoid information. For a discussion of non-standard preferences leading to such behaviors, see Golman, Hagmann, and Loewenstein (2017).

ignorant provides an excuse not to give, while maintaining a belief that they are generally an altruistic person.

Importantly, for this strategy to work, choosing ignorance and then just following self-interest must be less damning than having behaved selfishly with full knowledge of adverse consequences for others. This corresponds to a self-signaling model in which remaining ignorant is easier to rationalize away ("I did not really understand, was not really paying attention," etc.) and/or less memorable than an explicit sharing choice, or less informative about exactly how selfish the person is; see Section 3.4 for further discussion.

3.1.1 Experimental paradigm

Many papers that have studied information avoidance give subjects the choice of whether to reveal payoff-relevant information about prosocial decisions. Dana, Weber, and Kuang (2007) conduct an influential experiment creating such a situation. Its key feature is a "moral wiggle room game", which is a modified dictator game. Dictators allocate money between themselves and another subject by choosing between two options, A and B. They know that Option A yields \$6 for them, and Option B only \$5. Concerning the recipients, there are two equiprobable states of the world. In State *Unaligned*, they receive \$1 when the dictator chooses Option A, and \$5 when Option B is chosen. In State *Aligned*, the outcomes are switched: now Option A leads to \$5 for the recipient, and Option B to \$1. In standard conditions where there is full information about the state, under *Aligned* (almost) all Dictators naturally chose the Pareto-dominant Option A. Under *Unaligned*, which features a tradeoff between an extra \$1 for themselves and an extra \$4 for the recipient, 74% choose the prosocial (and total-payoff-maximizing) Option B, in line with standard results from the large literature on dictator games.[9]

The key feature of the moral wiggle room game, however, is that Dictators do not know which state they are in, but can learn it prior to choosing between A and B, by simply clicking a button labeled "reveal game". Learning is thus costless and effortless, yet Dana, Weber, and Kuang (2007) find that only 56% of dictators click on the button to reveal the

[9]See Engel (2011) for a meta-analysis, and Barmettler, Fehr, and Zehnder (2012) for a demonstration that dictator game giving remains even under double-blind conditions.

outcomes for the recipient. Moreover, 87% of those who do not click choose Option A, which maximizes their own payoff.

These results pose a significant challenge to the existence of social preferences, which is the standard explanation for why most subjects in Dictator games share fixed endowments, or maximize the sum of payoffs, as is the case here in State *Unaligned*. Indeed, the prediction for agents who really value $4 for the recipient at least as much as $1 for themselves is clear: always press the button and learn about the outcomes. As learning comes at no cost, even self-interested agents are at least indifferent between learning and not learning.

In contrast, a self-image explanation can explain the simultaneous propensity to give under full information about outcomes and the avoidance of such information. First, it could be that subjects are risk-averse with respect to self-image and do not want to endanger a relatively high prior by finding out more about their generosity (*ex-ante* information avoidance). More likely, many already have some insight as to how they would feel if confronted with the *Unaligned* condition, which they were shown in the explanation phase of the experiment, and is a fairly simple choice to project oneself into. As in our earlier charity example, in this case it is predominantly individuals with low altruism who choose to remain ignorant, so as not to be explicitly (and memorably) confronted with having to sacrifice either utility from money or utility from self-esteem. For the more altruistic individuals, on the contrary, there is only a potential gain in utility on the monetary side (they would happily choose Option B), so they would always want to know. In line with this *ex-post* ignorance interpretation, the vast majority of those who did not click choose Option A, which also means that such a choice is, statistically, a sign of selfishness.

3.1.2 Subsequent literature

The basic setup and results of Dana, Weber, and Kuang (2007) have been replicated and extended many times. This is important, as the initial experiment had a relatively low sample size – 32 dictators made the information choice and only 19 faced the choice in State *Aligned* without any uncertainty. In a meta-analysis, Vu et al. (2023) review decisions made by 6,531 participants across 22 studies. On average, they find that 40% of subjects avoid information

about the consequences of their actions for others, then predominantly pursue their self-interest. The pattern thus appears to be very robust. However, the original design left room for other explanations than "willful blindness" due to image concerns. Several subsequent papers therefore extended it to provide more stringent tests of the mechanism.

In the original experiment, dictators make their decisions anonymously but the experimenter can still link their choices back to them when handing out the payment, which may trigger social-image concerns. In an important replication, Larson and Capra (2009) therefore implement the moral wiggle room game in a double-blind setting. They find very similar results, with in total 53% of participants choosing to avoid the information.

Feiler (2014) also expand on the design by adding a within-subject approach. The author replicates the finding that a significant fraction of individuals avoid information in a self-serving way, and, importantly, this is true even for those who choose the prosocial action when it is not possible to remain uninformed. Matthey and Regner (2011) similarly investigate information aversion in repeated decisions, and find that a significant number of subjects who show other-regarding behavior when payoffs are known prefer to remain ignorant when given the opportunity.

While these results are very consistent with a strategy of self-image preservation, they do not yet fully prove that they are caused by it. For instance, dictators may be inattentive or confused by the setup, causing random choices that leads them to remain uniformed. Indeed, even though a dictator game choice is a very simple, non-strategic decision, Koppel et al. (2025) show that 22% of subjects misunderstand important aspects of the game. Therefore, Exley and Kessler (2023) design a clever experiment to directly assess the role of self-image in driving information avoidance in the moral wiggle room game. They add to the original setup a control condition that changes one part of the consequences of both Options A and B: it is now a third participant who receives the payoff that in the original setup went to the dictator. All other features of the design – the consequences for the recipient, the information choice, and the timing of information provision – are kept unchanged. Inattention or confusion would therefore lead to similar fractions of subjects not revealing payoffs in this control condition and in the original one. Self-image, in contrast, predicts

information avoidance only in the latter: absent consequences for the dictator, choices are unrelated to their selfishness or generosity. First, replicating once again the original paper, Exley and Kessler (2023) find, across multiple experiments, that between 62% and 73% of subjects avoid the information, whereas in the known State *Aligned*, between 61% and 68% choose the altruistic Option A. Importantly, they find significantly lower rates of information avoidance in the control condition, as predicted. The share does not drop to zero, however, indicating that image concerns do not fully explain the avoidance and that other motives are also at work. Dividing the fraction of subjects avoiding information in the control condition by the fraction avoiding it in the original game, the authors conclude that self-image concerns drive about 19% to 34% of aversion choices.

Another way of testing the robustness of identity-protecting willful ignorance is to give subjects the opportunity to avoid payoff information not only about the recipient, but instead about their own, or about both. This allows distinguishing self-serving information avoidance from pro-social information avoidance and general information avoidance. Self-image concerns would predict willful ignorance only in the first of the three cases. Accordingly, Momsen and Ohndorf (2023) give Dictators the option to avoid all payoff information, while Kandul and Ritov (2017) and Moradi (2018) also give them the option to avoid information about their own payoff. In all three experiments, the authors find substantially lower rates of information avoidance when the information is about own or both payoffs than about the recipient's.

Another line of work investigates in what settings information avoidance is more or less likely to emerge, and in particular, whether it is sensitive to tradeoffs, as predicted by self-image theories. Grossman (2014) thus expands the original design to three conditions. The first reproduces the original setting. In the second, subjects must actively choose (click) whether or not to learn, and in the third, they have to actively choose not to learn. He finds that, relative to the original setting, information avoidance is significantly lower in the second condition and almost entirely vanishes in the third one. The fact that defaults matter greatly for information avoidance is quite consistent with it being driven by self-image concerns: it is much more cognitively costly (greater dissonance) to uphold the image of being

altruistic when one has just actively decided against learning payoff-relevant information about another person, compared to passively letting it happen.

Relatedly, Serra-Garcia and Szech (2022) study how information avoidance responds to incentives. As paradigm, they use a game where subjects choose whether to open an envelope that, with equal probability, contains a $10 donation to a charitable cause or is empty. After deciding whether to open it, subjects choose whether to receive money for themselves (on the order of $2.50) or donate the envelope's content. Between subjects, the authors vary from negative to positive an extra payoff that subjects receive for opening the envelope. Replicating earlier work, they find substantial information avoidance at the zero-payoff point –slightly less than 50% of subjects do not open the envelope. More importantly, they find a well-behaved, downward-sloping demand curve for information. When not opening the envelope involves a cost of $0.10 the fraction not opening is (still) about 30%, and when it is opening that costs $0.10, the proportion rises to about 70%.

Self-serving information avoidance has also been studied in settings other than the dictator game. For instance, Conrads and Irlenbusch (2013) design an ultimatum game in which the proposer can choose to remain ignorant of the proposed split that will result from their two choice options. They find that a significant fraction of proposers choose to remain ignorant, and then predominantly select the self-serving option. Epperson and Gerster (2024) study avoidance of information about farming practices, and subsequent meat consumption choices. They find that when information can be acquired at no cost, a significant fraction prefers of paticipants not to receive it. Momsen and Ohndorf (2020) study this behavior in a market setting. Subjects purchase products with externalities and can learn about the nature of these negative spillovers. When information is costless, the authors do not find information avoidance, but it starts appearing as soon as revelation involves a trivial cost (€0.02), which most likely serves as an excuse.

3.1.3 Avoidance of prosocial decisions

Besides avoiding information that could guide potentially prosocial choices, people often avoid altogether situations where they might face such a decision, even under anonymity.

In Lazear, Malmendier, and Weber (2012), subjects can either play a standard dictator game where they have $\$y$ to share with another subject, or stay out, in which case they receive $\$x < \$y$ and the other subject nothing. The authors find that a significant fraction of participants prefer the latter option, even though when subjects are forced to make the dictator decision, the overwhelming majority give a non-zero amount, replicating standard results.

As with information avoidance, this behavior poses a challenge to both standard economic theory and models of social preference. A fully selfish agent would enter and allocate all of y to themselves. One with any degree of altruism (or other fairness preferences, such as inequality aversion) would similarly enter, so as to choose their preferred division of $y > x$. In contrast, agents with self-image concerns anticipate that they would or might feel compelled to give (or give more than they would really like) in order not to display obvious selfishness, and thus prefer to avoid the "test" of playing the game.[10]

Using a natural field experiment, Adena and Huck (2020) provide evidence for self-serving decision avoidance in an online fundraising campaign. While booking tickets for the opera, customers are asked whether they would like to donate to help disadvantaged schoolchildren. In one condition, they can continue booking without a donation only after ticking either a box saying "I have donated already" or a box saying "No, thank you". In the other condition, they can proceed without ticking any box. The authors find that this seemingly small change significantly decreases donations. Intuitively, lying by declaring to have donated before, or explicitly declaring one's selfishness, induces a hit to self-image and makes it hard to forget later on that one chose not to contribute. Not having to declare one's intention helps to maintain positive beliefs about oneself, by later conveniently "overlooking" or forgetting that there was a fundraising call. As a result, self-image-motivated customers feel less compelled to donate in that case. Importantly, social image cannot be at play here, since the opera house observes the actual donation decision in either case.[11]

[10]An alternative account of such "crossing the street to avoid walking by a beggar" strategies are as commitment devices for an agent who does not want to give but fears succumbing to a temptation to do so (Dillenberger and Sadowski, 2012). The natural and common explanation for the psychological origin of this temptation is the shame that one would feel from behaving selfishly –a case of negative self-image.

[11]In Dana, Cain, and Dawes (2006), both self and social image may be at play. Subjects first choose a potential dictator allocation over $10, then decide between having it implemented or receiving $9. In the

"Avoiding the ask" has also been observed in the context of many other fundraisers. DellaVigna, List, and Malmendier (2012) show that many people choose not to be at home when fundraisers have announced that they will visit them at a specific time to ask for donations. Similarly, Andreoni, Rao, and Trachtman (2017) show that people choose a different route of exit at a supermarket to avoid being asked to donate to a good cause. Knutsson, Martinsson, and Wollbrant (2013), Cain, Dana, and Newman (2014), Trachtman et al. (2015), and Damgaard and Gravert (2018) provide further evidence on people avoiding being asked or reminded of prosocial decisions. While these situations feature (sometimes minimal, online) social interactions and thus potentially social pressure, self-image may also contribute to the observed behaviors.

3.2 Motivated belief updating

Very often, there is no way to avoid identity-relevant information, or choice situations that will generate such signals. Examples include personal successes and failures, feedback from or treatment by one's ingroup or partners, performance reviews, or ethically charged product labels. In such cases, individuals who want to maintain a certain identity may distort *ex post* the signals received or their interpretation, and hence their belief updating, through various forms of *motivated reasoning* (Kunda, 1990; Epley and Gilovich, 2016).

3.2.1 Experimental paradigm

An often-used paradigm in the literature uses stylized updating tasks where subjects receive information about their IQ. Both their priors about their relative rankings and their posteriors following the signal(s) are elicited, allowing precise identification and quantification of updating behavior. Information about intelligence is particularly suitable because (i) a proxy for it can easily be elicited during the time frame of an experiment, (ii) intelligence scores

latter case, the receiver will never know that the game was played, but in the former, they will know and see the allocation. Thus, while self-image is most likely at play here again (having exited may be less memorable than having implemented a selfish split), choosing the "quiet exit" option, which about 28% of subjects do, also frees the dictator from any judgment by the receiver. The same is true for Broberg, Ellingsen, and Johannesson (2007), who use a similar design but elicit subjects' willingness to pay for "erasing" their initial allocation decision. Overall, 64% of them are willing to exit for less than their endowment as dictators.

are comparable, and (iii) it is highly self-relevant, being both socially valued and relevant to many educational and professional domains.

Two influential experiments employing this design are Eil and Rao (2011) and Möbius et al. (2022). The paradigm has four essential parts. First, subjects perform an IQ test. Afterwards, they are put in a comparison group with other subjects and asked to state their belief about how they rank within their group in terms of IQ performance (prior). Third, they receive repeated signals that are informative about their relative rank, but are not fully revealing. After each one, they are asked, with incentives for accuracy, about their posterior belief in their rank among the group. In a control condition, subjects receive signals that are computationally equivalent to the signals about subjects' IQ, but pertain to some arbitrarily assigned ranking.

Within this paradigm, the two papers and subsequent literature differ slightly in implementation details. In Eil and Rao (2011), subjects form groups of ten and receive three signals (without replacement) about whether they rank higher or lower than another group member. Prior to receiving any signal and then after each one, subjects are asked to state their full belief distribution over the ranks. Beliefs are incentivized using a quadratic scoring rule. In Möbius et al. (2022), subjects receive four signals about whether they rank above or below the median among all subjects. They know that each signal is only correct with a 75% probability, which has the attractive feature that it induces exogenous variation in the information that subjects receive.[12] Prior to and after receiving each signal, subjects state their subjective likelihood of actually having an above-median score on the IQ test. Their guess is incentivized by eliciting the value of x for which they are indifferent between receiving a prize with probability x and receiving the prize if their score was above the median. In the control condition, subjects in Eil and Rao (2011) update about their rank in a meaningless "card task", while in Möbius et al. (2022) they update about the performance of a "robot" rather than their own performance.

Both papers observe a strong asymmetry in updating to positive versus negative news about relative IQ performance. Subjects tend to update in a fairly Bayesian fashion in the

[12] In particular, the random variation is orthogonal to subjects' performance in the IQ-test, which may be correlated with updating behavior.

first case, but show significant unresponsiveness in the second. Importantly, this asymmetry is absent in the control condition. The asymmetry in updating about IQ therefore cannot be driven by differences in processing or perceiving negative relative to positive news *per se*. Möbius et al. (2022) further show that the resulting upward-distorted beliefs causally induce a higher propensity to compete in a tournament rather than work for a piece rate.

Both papers also elicit subjects' willingness to pay to perfectly reveal their rank after the updating rounds. A significant fraction is willing to forgo money to not receive such information. In particular, subjects in Eil and Rao (2011) who came to believe they were among the highest IQ performers are willing to give up part of their earnings to learn their true rank, whereas those who believed they were among the bottom require a subsidy to receive the information. As predicted by the theory, motivated reasoning and information avoidance go hand in hand, being two strategies for self-esteem management.

3.2.2 Subsequent literature

In a similar vein, Bolte and Fan (2023) show that subjects neglect more strongly the fact that information on their relative IQ rank comes from the same source if the news is positive compared to negative – they exhibit self-serving correlation neglect. Again, this pattern is significantly less pronounced when the news is not ego-relevant.

A similar asymmetric effect occurs in the context of disclosure games, as Hagenbach and Saucet (2025) show. In their experiment, when senders who have private information about a receiver's IQ but also an incentive to deceive them disclose (or not) their information strategically, receivers display more skepticism towards self-threatening than towards self-serving information. When the information is not ego-relevant, no such asymmetry is present. The authors further show that the results on IQ are well explained by receivers carrying either more or fewer steps of iterative reasoning (elimination of dominated strategies), depending on whether these steps take them in an ego-boosting or ego-damaging direction.

In line with the predictions of the anticipatory-utility version of the theory, asymmetric updating appears to be sensitive to the delay until uncertainty resolution (or the lack thereof). As Drobner (2022) shows, the pattern of asymmetric updating is stronger when

participants do not expect to eventually be told of their relative IQ than when they know it will be. This helps explain the sometimes mixed evidence on asymmetric updating. For instance, Coutts (2019) conduct an experiment in which subjects know that they will learn the truth at the end, and finds no evidence that they respond more to positive news than negative ones. From a self-image perspective, this is in fact expected, as the anticipation of full revelation reduces almost to zero the period during which favorably distorted beliefs could be "savored". Further supporting the self-image model, Drobner and Goerg (2024) provide evidence that, when subjects expect no resolution of uncertainty (but not when they do), they interpret signals received in a self-serving way, judging IQ as more or less relevant to study and job performance depending on whether these signals were favorable or unfavorable.

Buser, Gerhards, and Van Der Weele (2018) investigate individual heterogeneity in the degree to which subjects display asymmetric updating. They enrich the design of Möbius et al. (2022) by having subjects take three different IQ tests and examining how they update in each case. This allows the authors to investigate whether asymmetric updating about ego-relevant information can be considered a relatively stable individual trait. They do not find significant correlations across tasks, indicating that such behavior may be specific to the decision-context.

Several papers have investigated the role of motivated belief updating in domains other than intelligence.[13] In Eil and Rao (2011), in addition to IQ, the experiment also contains a treatment where subjects do a speed-dating exercise and are subsequently rated on their attractiveness. They then receive signals about their attractiveness ranking in their group, and show asymmetric updating as in the case of intelligence.

In the moral domain, Konow (2000) shows that stated judgments about what is a fair allocation differ when a subject chooses selfishly as a dictator and when they judge sharing between two other subjects. While these statements are not incentivized, they are anonymous, and the author proposes dissonance reduction as the intervening mechanism. Heese and Chen (2025) offer subjects information about whether a profitable decision harms a re-

[13] See also Sharot and Garrett (2016) for an overview of the evidence of the good news/bad news asymmetry in psychology and neuroscience.

ceiver in a dictator game. The authors find that subjects selectively "fish for good news": they stop earlier if the available news indicates that their action does not harm the receiver than when it indicates harm. This motivated depth of information search parallels the motivated depth of iterative reasoning in Hagenbach and Saucet (2025).

In Gneezy et al. (2020), who build on the design of Babcock et al. (1995), participants act as financial advisors recommending investments to clients, with one option (A) offering a commission but yielding lower returns for the client. When advisers learn about the incentive before evaluating the investments, they are significantly more likely to recommend the self-serving option than when they learn only afterwards, as long as the risk structure of the assets' returns allows some scope for rationalization and self-deception. Revealing further evidence of motivated beliefs, in those conditions the advisers are also more likely to choose investment A for themselves when informed of the incentive prior to evaluating the assets.

Bosch-Rosa, Gietl, and Heinemann (2024) study motivated beliefs in the context of investors with moral hazard and limited liability. Limited liability naturally encourages investors (managers, banks, etc.) to take more risk at others' expense (shareholders, the state for bailouts), but doing so can make it more difficult to maintain a positive self-image. To resolve the tradeoff in a self-serving way, investors may hold motivated beliefs that the risk of losses is low. Indeed, Bosch-Rosa, Gietl, and Heinemann (2024) find that, holding signal strength about loss risk constant, for the same signal, investors state higher (incentivized) likelihoods of project success under limited liability when it leads to loss sharing than under either full liability, or limited liability without loss externalities. At the same time, they invest more in the risky asset, and the authors estimate that one third of the increase is due to motivated optimism, which thus acts as a "multiplier" of the standard incentive effect. The fact that there is no significant distortion when there are no potential adverse consequences also allows the authors to attribute the belief distortion to self-image concerns rather than to anticipatory utility about payoffs.

In a separate experiment, both the losses and the gains from investments are shared with others. On average across all subjects, the authors find no differences in beliefs between limited and full liability. They do, however, find evidence of motivated beliefs among those

who showed concern for others by previously giving a positive amount in a dictator game. The self-image perspective can rationalize this heterogeneity: a person who fully knows that they are selfish or has no self-image concern, hence no incentive to distort beliefs. In contrast, those who want to signal to themselves that they are altruistic will both give in the dictator game and distort their beliefs in the investment game.

As mentioned in Section 2, besides affective reasons such as pure self-esteem and anticipatory utility, people may hold motivated beliefs about themselves and the consequences of their actions ((θ, e) in the model outlined earlier) for instrumental reasons.

The first one is that certain beliefs are more conducive to effort and perseverance, either because they help overcome temptations (the Protestant work ethic being an example), or because they make it more compatible with certain valued identities. In Chen and Schildberg-Hörisch (2019), participants perform a real-effort "slider" task, then report their belief distribution over their productivity. Overconfident individuals consistently exert more effort in subsequent rounds. Conversely, when they receive debiasing feedback, their effort declines significantly. Bönisch et al. (2024) explicitly test a version of the Bénabou and Tirole (2002) model, in which belief distortions help increase effort motivation in the face of a self-control problem. Drawing on a large dynamic student survey, the authors collect repeated measures of students' expectations about how much studying will improve their exam performance, then test the prediction that optimism about the value of effort should peak just ahead of exams. Students' average believed returns to effort increase by about 20% as exams approach, then drop off sharply afterwards. By controlling for ability, historical effort, and external signals, the analysis rules out information acquisition or changing actual returns as explanations for these shifts, pointing instead to the self-motivation role of belief distortions.

In a rare study of how sustaining a religious identity induces motivated beliefs, Wang, Wang, and Ye (2023) take advantage of a year in which the very high-stakes Chinese national exam falls during the Ramadan period. A sample of Muslim students in China preparing for the exam are presented with a graph clearly showing that fasting significantly impairs test performance. When asked to simply read off the graph the magnitude of the effect,

subjects underestimate it by a large amount, even though their answers are incentivized and the information is right in front of their eyes. In contrast, for another sample who were previously given writings by Muslim clerics who used Quranic reasoning to explain that pausing the fast during exams is religiously permissible, underestimations occur half as often.

A second functional reason for inflated self-image is that persuading oneself helps persuade others, a mechanism that Hippel and Trivers (2011) argue conferred evolutionary advantages. Schwardmann and Van Der Weele (2019) and Soldà et al. (2022) provide causal evidence of such strategic self-deception. In both experiments, participants complete a task and estimate their relative performance. Some are informed beforehand that they will have to persuade another person of their ability, others learn this only after (incentivized) beliefs are elicited. In both studies, participants anticipating a need for persuasion form more optimistic beliefs about their performance, and these inflated beliefs indeed make them more persuasive in a face-to-face interaction. Soldà et al. (2022) also show that, as a mechanism for arriving at inflated beliefs, participants anticipating persuasion choose feedback signals that are more likely to be positive than negative (how well they answered easy rather than hard questions) and, here again, under-update to negative signals.

Finally, given the relevance of motivated updating, an important question concerns the extent to which people anticipate that they will self-servingly distort their beliefs, a form of metacognition previously discussed in Section 2. In this case, they may engage in ex-ante behavior that either limits or enhances the scope for motivated thinking. To investigate this question, Saccardo and Serra-Garcia (2023) design an experiment in which subjects take the role of advisors, recommending products to uninformed receivers. The advisors can choose whether to first see the incentives they receive for their recommendation, which enables motivated reasoning, or the quality of the product, which does not. They find that people are quite heterogeneous in their cognitive strategies, with half choosing to see the incentive first and half the quality. Importantly, actively enabling motivated reasoning does not impede engaging in it, as advisors choosing to receive the incentive first indeed make more self-serving recommendations. Once again, self-deception proves effective in enhancing self-image.

3.3 Motivated memory

Memory is central to the very notion of identity, as it underlies both the fundamental sense that we remain the same person throughout life and what we learn over time about this person.[14]

Human memory, however, is inherently imperfect; the process through which experiences are encoded, stored, and later retrieved is subject to substantial information loss, uncertainty, and ambiguity. These frictions provide "cover" for a self-image motivated agent to shade their memories in an identity-enhancing direction, and these memory distortions in turn influence later decisions –for instance, by inducing overconfidence. A student or employee may thus recall more easily their successes than their failures, a religious person their good deeds and adherence to precepts than their lapses, a group member their adherence to its norms than their straying from them, etc. For an individual without self-image concerns, in contrast, memory may be imperfect, but errors in both directions would be equally likely.

Following both psychological insights and economic models formalizing the determinants and consequences of motivated recall (Bénabou and Tirole, 2002, 2004, 2011; Hagenbach and Koessler, 2022; Fudenberg, Lanzani, and Strack, 2024), a number of recent studies have investigated the phenomenon both in the lab and in the field. Amelio and Zimmermann (2023) provide a recent review, so here we focus on explaining the core experimental paradigm, its gradual extensions, and some recent applications,

Before doing so, let us note that *some form of motivated attention or memory is necessarily involved* in motivated belief updating and all other identity-protective strategies we discussed earlier, except for the relatively rare case of completely *ex-ante* information aversion. For example, consider a selfish agent who, in order to preserve a positive self-image, avoids information or a choice situation, or when faced with a decision, overrules their selfish desires and chooses altruistically. Equivalently, consider a weak-willed agent who, for self-image preservation, avoids potential information sources (temptations) or acts like a

[14] "For, since consciousness always accompanies thinking, and 'tis that which makes every one to be what he calls self, and thereby distinguishes himself from all other thinking things: in this alone consists personal identity, i.e., the sameness of a rational being; and as far as this consciousness can be extended backwards to any past action or thought, so far reaches the identity of that person." — Locke (1975).

disciplined individual would in the same circumstances. For this to affect beliefs, it must be that their first impulse (deep preference) and the real reason they acted the way they did are subsequently no longer perfectly accessible in memory or awareness.

3.3.1 Experimental paradigm

Zimmermann (2020) builds on the classical paradigm of motivated belief updating and adds a recall stage, together with several novel treatments. Subjects take an IQ test, then receive a noisy signal about their relative rank in a group. As usual, noisiness enables causal identification of the effect of the signals' valence (positive/negative). Importantly, the timing of posterior elicitation was randomly varied between immediately after the signal and one month later. In the first case, subjects update in a rational manner: positive signals increase beliefs in a high rank, while negative signals decrease beliefs. After one month's delay, however, subjects who had received a negative signal barely lower their posterior belief, while those who had received positive signals still show strong increases. In a second experiment, the author demonstrates the role of motivated memory as a mechanism in generating this temporal good news-bad news effect: after one month, recall of the signal and even of the IQ part of the experiment itself is significantly lower when the feedback was negative than when it was positive. This raises the question of whether the asymmetry occurs at the encoding or the retrieval-and-decoding stage of the memory process. Zimmermann (2020) provides evidence for both mechanisms. First, to provide evidence for motivated encoding, before receiving the signal, subjects are informed about the incentivized posterior elicitation that will happen a month later, creating an incentive to store the signal for future recall. The announcement indeed significantly increases updating from negative signals one month later, while having no effect on updating from positive signals. Second, to provide evidence for motivated decoding, no announcement takes place, but the unexpected incentives for subjects to correctly state their rank are increased tenfold. This again has no effect on updating from positive signals after one month, but significantly improves the updating from negative signals. The fact that both treatments only show effects for negative signals provides clear evidence of motivated recall: subjects are already encoding and decoding close

to their capacity level for positive signals, so external manipulations have no effect. For negative signals, however, the tradeoff between incentives for accurate recall and maintaining a positive self-image leads manipulations to have significant, predictable effects.

In a similar vein, Chew, Huang, and Zhao (2020) test subjects' incentivized recall, after several months, of their performance on an IQ test. Participants first solve four Raven's matrices, then several months later they are shown those same questions plus two others, together with the solutions, and asked whether they got the question right, wrong, did not see it, or cannot remember. In addition to forgetting failures more than successes ("positive amnesia"), subjects also tend to misremember failures as successes ("confabulation") and even to say they solved correctly questions they never saw ("delusion").

3.3.2 Subsequent literature

Moving the literature from the lab to the field, Huffman, Raymond, and Shvets (2022) investigate the role of biased recall in explaining why overconfidence persists among managers even when they receive repeated, public feedback. The study focuses on about 230 store managers in a firm, for whom a significant fraction of pay is determined by their rank in quarterly performance tournaments. Rankings are computed according to well-defined, transparent criteria and communicated as part of quarterly performance evaluations. The authors elicit both managers' predictions of their future rank (Q4 2015) and their recollection of past rank (Q2 2015). Since the experimenters are known to have that data, there is no point in lying, and incentives are given for accuracy.

The first result is that nearly half of the managers exhibit overconfidence, predicting better-than-justified future ranks by about 0.5 quintiles on average. This holds whether predictions are compared to ex-post realizations, to reduced-form forecasts based on the performance record, or to a structural Bayesian-learning model. The second key result brings to light the role of systematically biased memory: managers who underperformed tend to misremember their past rank as significantly better, with average errors exceeding 30 places, while top performers recall their rank much more accurately. Approximately 56% of managers exhibit positively skewed recall errors, versus just 24% for negative ones. Further-

more, these distorted memories strongly correlate with inflated predictions: a one-standard-deviation increase in recalled rank corresponds to a half-quintile increase in predicted future rank. A structural model incorporating self-serving recall provides a good fit to the data, outperforming Bayesian models that include both public (evaluation) and private signals.

Using a similar design and data from the Kenya Life Panel Survey, Müller (2022) examines whether parents reconstruct their memories of past fertility preferences to rationalize their current family outcomes. In one round of the survey (2007–09), participants were asked about their desired total family size. In a later round, about ten years later (2018–2021), the number of children to date is observed, and parents are incentivized to recall their earlier stated preference. Among the 1,428 respondents, about 29% already have "excess fertility" by that time, namely, more living children than they had originally reported wanting. Crucially, recall errors are strongly associated with current fertility outcomes: individuals with excess fertility are over 25 percentage points more likely to overstate their past preferences and desires (by one child on average), compared to those with no excess fertility. The degree of misremembering increases with the number of "excess" children and the time elapsed since the undesired birth. Furthering the case for motivated forgetting rather than cognitive limits, monetary incentives for recall have little effect (whereas they work for another "neutral" question), and respondents with excess fertility are significantly less likely to take up an offer to see their past answer, despite a financial incentive to do so (information avoidance).

Thus, as predicted by theories of self-justification and identity protection, individuals adjust their recollections of major life preferences to rationalize outcomes that would otherwise be dissonant, thereby maintaining a coherent and affirming narrative about themselves. Importantly, these biased memories influence future behavior: individuals who recall having wanted more children than they actually did are 10 percentage points more likely to recommend high fertility norms (e.g., 5 or more children) to adolescents.

In Hagenbach, Jacquemet, and Sternal (2025), subjects first receive a signal with a stated level of informativeness about their IQ performance. Two days later, they are reminded of the signal and, for some but not others, of its informativeness; they also state

their posterior beliefs about IQ. Once again, recall is self-serving: those with negative news report significantly lower informativeness when there is no reminder of its true level than when there is, whereas those with positive news display no such difference. Furthermore, absent the reminder but not with it, subjects with negative news manage to maintain excessively high beliefs about their IQ. Similarly, in a condition where the nominal message is negatively correlated with the truth, subjects tend to inflate the informativeness of positive messages that should now be interpreted as negative ones.

As is the case for updating, the effects of motivated memory are not limited to the IQ domain. In the context of prosocial decisions, Carlson et al. (2020) provide evidence that subjects recall a higher degree of giving in dictator games than actually occurred, whereas if a computer randomly makes choices for them, their recall is no longer biased. Saucet and Villeval (2019) similarly show that subjects have a higher recall accuracy when they picked the fair option, and that this effect goes away if a computer randomly selected it. Thus, in both cases, once a threat to people's self-image is removed, they have no difficulty remembering outcomes that maximize their payoff.

In a multi-country study on attitudes relating to the COVID-19 crisis, Sprengholz et al. (2023) show that people surveyed in 2022 adjusted, in the direction of their current perceptions, their recall of past (2020) stated perceptions of risk (infection probability and severity), trust in government and science, frequency of wearing masks and perceived official exaggeration of the pandemic measures. Furthermore, the direction of the bias also differed according to current vaccination status and identification with that status.

Also in the field, Roy-Chowdhury (2022) studies longitudinal data on schoolchildren who are asked to recall their math grades. They find that students, especially low-performing ones, tend to remember their grades as better than they actually were. These memory distortions grow with time since the original grade, and they predict inflated beliefs about academic ability in future periods, controlling for actual performance.

Sial, Sydnor, and Taubinsky (2024) ask participants to recall their gym attendance and find that over 70% overestimate it. Using past attendance as an instrument, they further show that these memory distortions induce optimistic beliefs about one's willpower and a

lower demand for commitment devices.

In Gödker, Jiao, and Smeets (2025), subjects repeatedly invest in either a safe or a risky asset. The authors find that positive returns are over-remembered and negative returns under-remembered. Using random variation in the delay before recall, they further show that these distorted memories cause overestimations in the risky asset's return and own investment ability, together with excessive reinvestment and risk taking.

In Van Der Weele and Von Siemens (2020), subjects make a donation decision and subsequently receive a bracelet. In the treatment condition, they receive a bonus if they return two weeks later and present the intact bracelet, while in the control they receive it just for coming back. As the bracelet likely serves as a reminder of the donation decision and thus potentially increases its signaling value, anticipating this effect may lead subjects to act less selfishly. However, the authors find no significant difference in behavior, and thus no evidence of self-signaling. A second experiment, in which subjects engage in a push-up competition, suggests that motivated memories and rationalizations of past actions (a_2) can reduce the effectiveness of such reminders, as can be seen from Figure 1. Depending on whether participants learned that they performed well or poorly relative to others, they later on altered in a self-serving direction their stated diagnosticity of push-ups for fitness, the importance of fitness for their self-image, and motivations for entering the contest.

3.4 Self-signaling

As noted earlier, a question that naturally arises in the paradigm of Dana, Weber, and Kuang (2007) and subsequent information-aversion experiments is why choosing ignorance does not reveal the agent as selfish, in which case it would be self-defeating. Indeed, a truly altruistic person would always want to know, and conversely, subjects who do avoid information almost always behave according to self-interest afterwards, so remaining uninformed is indeed a signal.

A first reason may be that people succeed in avoiding thinking through, or manage to forget, these relatively sophisticated metacognitive inferences about their motives. Another one is formalized by Grossman and Van Der Weele (2017), building on the self-signaling

model of Bodner and Prelec (2003) and Bénabou and Tirole (2004, 2011), in which agents with imperfect recall of their true motivations infer their types from their actions (see Section 2). Extending the framework to allow for a continuum of generosity types, the authors show that there exists a semi-separating equilibrium in which choosing ignorance indeed lowers self-image but, for a range of relatively ungenerous agents, not enough to justify taking the risk of learning that the state is *Unaligned* and being confronted with either acting prosocially or having self-image fall to its lowest possible level.

In a very different setting, Bursztyn et al. (2020) conducted field experiments with Pakistani men, both in a lab and at home. Respondents are offered a bonus payment of about 20% of their daily wage if they anonymously check a box expressing gratitude toward the U.S. government, which funded the experiments. Between 25% and 33% of them refuse the payment —a costly self-signal of having the "right" political identity. Similarly, Bénabou, Falk, and Henkel (2024) finds that 40% of a sample of German university students refuse to anonymously click on a statement saying "I support the destruction of the environment," even though they fully understand that the sole consequence would be to trigger a donation of €15 to a charity for children with cancer, whereas clicking on "I support the preservation and protection of the environment" yields nothing for this or any other cause.

In a large online field experiment in China, Dubé, Luo, and Fang (2017) test another prediction of the self-signaling model, namely the crowding out of motivation by extrinsic incentives, even in private decisions. Randomized customers of a large cell phone provider are offered (via SMS) discounted movie tickets that, if purchased, will trigger a charitable donation by the company. For a fixed discount, the volume of purchases increases with the amount of the donation, and conversely, when there is no donation, it increases with the amount of the discount –both in line with standard demand theory. When the donation and the discount are bundled, however, varying the amount of the latter amounts to varying a monetary reward offered for a prosocial act (buying the ticket). As shown by Bénabou and Tirole (2006), this "overjustification effect" damages the prosocial image attached to the act, offsetting and potentially dominating the standard price-incentive effect of the discount. Since SMS solicitations and subsequent purchasing decisions are entirely anony-

mous, the image at play here can only be self-image. The authors find that for relatively small donations, discounts increase demand, but for even moderate-sized ones they display a hump-shaped pattern, consistent with greater subsidies increasingly damaging the self-signal of prosociality attached to a purchase, as predicted by the theory. This mechanism is confirmed by a follow-up survey in which, at moderate donation levels, the fraction of subjects reporting "feeling good about myself" as a motivation to purchase declines with the level of the discount.[15]

Gneezy et al. (2012) use three experiments in which subjects can "pay-what-you-want" to provide several types of evidence for self-image concerns. In the first one, informing consumers that half of the price they choose to pay for a good will go to a charity predictably increases the prices set by those who purchase, but reduces the fraction who do buy. This is a form of decision avoidance (how much to pay), related to the exit in dictator games discussed in Section 3.1. In two other experiments, respectively: (i) more consumers buy when the seller offers a discounted price relative to the usual one than when they can name their own; (ii) consumers choose to pay less when they must pay the seller directly than when purchases are anonymous. Both cases are instances of self-signaling. In the first one, paying an exogenously low price becomes uninformative about selfishness. In the second, being monitored increases social-image pressure, but this in turn crowds out the social-signaling value of the price decision.

3.5 Preference for truth-telling

Another important form of self-signaling involves truth-telling. Absent repeated-game and reputational concerns, standard economic agents should always lie if it is to their advantage. However, a robust finding is that people frequently tell the truth even when it means foregoing some gains, and do so even in one-shot, anonymous experimental settings.[16]

[15]Goette and Tripodi (2020) document another form of crowding out predicted by the model. When image concerns are heterogeneous, publicizing good deeds may backfire, as the act then becomes a stronger signal of image seeking rather than altruistic motivation. Studying repeated blood donors in Italy, they find that participants indeed make such inferences when public recognition is present, and that in this condition, total donations are lower than with a simple ask.

[16]When lying is detectable, agents care about the signal it sends even in anonymous interactions, (Dufwenberg and Dufwenberg, 2018; Gneezy, Kajackaite, and Sobel, 2018; Khalmetski and Sliwka, 2019).

Self-image concerns naturally induce truth-telling, as honesty is universally considered a key component of moral character. Lying is condemned by most religions (Christianity, Islam, Buddhism), normative ethical theories (e.g., Kantian ethics), and in the perception of the general population. Accordingly, telling a lie even when it is undetectable by others is a threatening signal to one's identity as a good, moral person.

3.5.1 Experimental paradigm

One of the most frequently used paradigms was introduced by Fischbacher and Föllmi-Heusi (2013). Subjects roll a physical six-sided die in privateand report the result. Their payoff is increasing in the reported number, except for the number six, which results in a zero payoff. While lies are undetectable at the individual level, comparing aggregate reported numbers to the expected outcomes of a die reveals the average extent of lying behavior. Fischbacher and Föllmi-Heusi (2013) estimate that as much as 39% of subjects report honestly and only 22% maximize their income through lying.

A second frequently employed paradigm is a modified cheap-talk sender-receiver game, introduced by Gneezy (2005). A receiver chooses between two options, without knowing the monetary consequences. A sender knows which Option favors the receiver (say, A) and which one favors them (say, B). The sender chooses between two messages: one tells the receiver that Option A maximizes their payoff (the truth), the other that Option B maximizes their payoff (a lie). Gneezy (2005) finds that about 55% of senders choose the truthful message, even though it reduces their payoff. At the same time, subjects display a tradeoff between honesty and self-interest: the more they lose from telling the truth, the more they lie on average.

3.5.2 Subsequent literature

Truth-telling even in unobserved settings has been replicated many times using the paradigm of Fischbacher and Föllmi-Heusi (2013): under low and high incentives, when decisions are repeated, experiments take place online, or coins instead of dice are used, in student and representative populations, and across cultures around the world. A meta-analysis by Abeler,

Nosenzo, and Raymond (2019) covering 90 experiments using the paradigm in 47 countries reveals that, on average, people forgo 38% of their maximal payment by refusing to lie. Direct lying costs, capturing self-image concerns, emerge as one of the principal components explaining the wealth of experimental data.

In the self-signaling model, the informational content of an action, namely what it indicates about the agent's character, depends on everyone's equilibrium play –an endogenous social norm. In line with the theory, Le Maux, Masclet, and Necker (2021) find that subjects' level of honesty in the die-rolling experiment is sensitive to information about how other participants behaved, even though individual choices are unobservable. These results all point to self-image being an important contributor to truth-telling behavior, with agents engaging in self-signaling (specifically, self-interested agents pooling with intrinsically honest ones) to preserve positive beliefs about what kind of person they are.

3.6 Excuses

Another important implication of self-image concerns is that it leads people to seize upon or even actively seek excuses: in order to act selfishly and still maintain a positive self-image, some other rationalization for the action needs to be given.

The literature has documented a rich set of rationales that people use to excuse self-serving behavior. One, previously encountered in the section on information avoidance, is uncertainty or ambiguity: people use the risk that a prosocial action may not realize as an excuse not to choose it. In general, it is well-known that once prosocial actions become uncertain, prosocial behavior decreases, for instance in dictator games (Krawczyk and Le Lec, 2010; Brock, Lange, and Ozbay, 2013; Freundt and Lange, 2017). Such behavior could simply reflect risk-aversion, however. To distinguish between the two, Exley (2016) conducts an experiment in which subjects choose between a certain payoff and a risky lottery that has a positive probability of paying nothing, under four conditions: (i) the certain payoff goes to them, whereas the lottery prize goes to a charity donation; (ii) the certain payoff goes to a donation, whereas the lottery prize goes to them; (iii) they are the sole recipient in either case; (iv) the charity is the sole recipient in either case. From the last two conditions,

one sees that subjects are equally risk-averse for themselves as they are for the charity. In the two conditions involving self-other tradeoffs, however, subjects' risk attitudes reverse: compared to the no-tradeoff conditions, in (i) they display more aversion to charity risk, and in (ii) less aversion to self risk. Intuitively, a sufficiently high risk aversion can serve as an excuse for the self-serving choice in the first case, and a sufficiently low one in the second. Exley (2016) also shows that the asymmetric risk-response pattern is stronger for subjects who prefer to avoid payoff-relevant information in the moral wiggle room game of Dana, Weber, and Kuang (2007), providing further evidence for excuse-driven behavior.

Garcia, Massoni, and Villeval (2020) show that subjects display a similar behavior when the costs and benefits for self and/or charity are ambiguous instead of risky. Interestingly, the magnitudes are similar, suggesting that the mere presence of uncertainty is used as an excuse, irrespective of whether it is due to risk or ambiguity.

Many other features of the choice context are used as excuses. Exley (2020) shows that people also use charity-performance metrics in that way. Relative to a choice between charities, subjects are much more sensitive to performance metrics when this can benefit them, compared to when the choice is between two charities. Exley and Kessler (2024) demonstrate that subjects make errors in simple cognitive tasks, but predominantly in the direction that benefits themselves. When performing simple additions, for instance, they make very few errors when those benefit one charity over another, but a significant number when those benefit them instead of a charity. They find similar directional errors are in estimation tasks prone to correlation neglect (see also Bolte and Fan (2023)) or to anchoring, which all provide the excuse that the mistake may have been due to "innocent" inattention or use of a simplifying heuristic.

Another excuse is to distort one's belief about others' behavior. Di Tella et al. (2015) show that subjects who have more to gain from behaving selfishly toward another player are more likely to believe that this person acted selfishly toward them. This effect appears to be sensitive to the experimental design or measurement, however, as follow-up studies generally found no belief distortions about others' actions (Ging-Jehli, Schneider, and Weber, 2020; Ahumada et al., 2022; Verrina, 2023).

In Hamman, Loewenstein, and Weber (2010), subjects use delegation as an excuse for bringing about selfish allocations. When given the option to delegate their dictator-game decision to a third party, they do so, but systematically seek intermediaries who have a track record of favoring the dictator over the recipient. Such delegation leads to significantly lower transfers, while at the same time, principals report feeling less personally responsible for the outcomes.

These patterns of behavior are again at odds with classical models of prosocial behavior, such as warm-glow or pure altruism, In contrast, using risk, ambiguity, or the possibility of inadvertent mistakes as excuses allows people to maintain a positive self-image while extracting more for themselves. Note that, here also, some element of *self-deception* must be involved. Some of the excuses used are fairly transparent, and moreover, when using an excuse, or a fortiori seeking one out, the individual must remain or quickly become unaware (self-serving memory) of why they really acted the way they did.

3.7 Further applications

3.7.1 Asking and being asked

As mentioned before, a large literature documents the “avoid the ask" phenomenon: people will often incur costs to avoid situations in which they would have made a moral or prosocial decision (e.g., DellaVigna, List, and Malmendier, 2012). On the other side of the interaction, despite expressions like “It can’t hurt to ask,” people often hesitate to make requests—whether for a favor, a loan, a recommendation, or a raise—even when most others would accept if made aware of the need. Indeed, asking can actually “hurt”, in two identity-relevant ways. First, through shame: a request may signal weakness, neediness, or lack of competence, thus damaging social image, and possibly self-image. Second, through rejection: a deliberate refusal reveals that the potential helper does not really value the person in need or the relationship, which can be hurtful or humiliating. To avoid such a loss in self-image, that person may refrain from asking, thereby foregoing significant benefits. The first mechanism primarily involves social signaling of some stigmatized trait (Chandrasekhar, Golub, and Yang, 2018). The second, studied by Bénabou, Jaroszewicz, and Loewenstein

(2025), is a form of information aversion, and it applies even when the request reflects positive characteristics (e.g., a nomination for a prize) rather than negative ones. The authors also show how this reluctance to ask can induce generous helpers to offer unsolicited help. The downside, however, is that a failure to offer can then deter asking even more, leading to "waiting traps" where each party would like the other one to act, but neither does.

3.7.2 Incentivization and hypothetical bias

People's choices in theoretical decisions and those that have real consequences often diverge, a phenomenon known as hypothetical bias. In the context of contingent valuation studies, in particular, several meta-analyses, e.g., List and Gallet (2001) and Murphy et al. (2005), provide robust evidence that hypothetical willingness-to-pay (WTP) generally exceeds by far those elicited using real money. As a more recent example, Rodemeier (2023) similarly shows that stated WTP for carbon offsets is substantially higher than the revealed-preference WTP.

Moral identity concerns provide a rationale for the discrepancy when there is a tradeoff between self-interest and generating positive externalities for others. When choices have little consequences, signaling prosociality by expressing a high WTP for the externality is "cheap" (Bénabou et al., 2023). When real money is at stake, such signaling becomes more costly, and demand for generating the externality thus declines. As such, not only can self-image motivations explain hypothetical bias, they also predict when the bias should be strongest: in decisions involving image-relevant options.

Johansson-Stenman and Svedsäter (2012) test this prediction in an experiment that varies two aspects of a WTP elicitation: first, whether it is hypothetical or has real consequences, and second, whether the elicitation is about an image-relevant good – the preservation of endangered animals – or a restaurant voucher, as a neutral good. For the image-relevant good, WTP is significantly lower in the real-consequence case compared to the hypothetical choice, whereas there is no significant difference between the two for the neutral good.[17] Similarly, Bénabou, Falk, and Henkel (2024) find that subjects' willingness to

[17]Heinicke (2018) similarly finds that hypothetical bias is higher when information given about a product emphasizes its sustainability, which may increase subjects' self-image concerns.

forgo self-payments in order to trigger a donation for a good cause is substantially higher when the choice has no consequence, compared to when there is a chance it actually gets implemented. In contrast, they find that hypothetical bias is absent in the trolley dilemma, an ethical decision where there is no tradeoff between benefits to self and to others, and no universally agreed notion of which is the moral action.[18]

Relatedly, Grossman (2015) runs dictator games in which choices are implemented either with certainty or with probability 1/3. Prosocial choices increase slightly in the latter case, but this difference is statistically insignificant, so that no evidence of self-signaling is found in this case.

3.7.3 Moral cleansing and moral licensing

In settings where moral decisions have to be made repeatedly, self-image concerns may give rise to what are called moral cleansing and moral licensing effects, which are two opposite forms of history-dependence.

Moral cleansing refers to the pattern in which, after a moral transgression such as acting selfishly toward someone else or harming the environment, individuals engage in compensatory or offsetting actions such as charitable donations, helping others, or endorsing prosocial values. Intuitively, having damaged their moral identity, they subsequently try to restore it. In Ploner and Regner (2013), for instance, subjects first roll a die, either publicly or privately, to determine their endowment for a subsequent dictator game. They then decide, as dictators, how much of this amount to share with another participant. The authors find that subjects who took advantage of the anonymity condition to cheat in the first stage, so as to receive a high endowment, are subsequently more generous in the dictator game.

Moral licensing, on the other hand, describes the phenomenon where past good moral behavior gives the individual a perceived "license" to act less morally in subsequent decisions.

[18]Related to hypothetical bias, self-image concerns may also contribute to systematic biases in self-reported survey data (Bursztyn et al., 2025). When answering sensitive questions, respondents may be reluctant to admit – even to themselves – that they hold certain attitudes or engaged in particular behaviors. For instance, evidence has shown that respondents overstate positive behaviors such as the amount of charitable giving (Bekkers and Wiepking, 2011) and physical activity levels (Colley et al., 2018), while downplaying behaviors that are associated with dependence or failure, such as applying for unemployment insurance (Dutz et al., 2021).

Intuitively, those who acted well have built up some self-reputational capital, some of which they can afford to spend later on. In a sequence of sixteen dictator games, Brañas-Garza et al. (2011) show that participants' donations in one round negatively predict giving in the next, thus demonstrating both moral licensing (less giving after generous choices) and moral cleansing (greater giving after selfish choices).

In marketing experiments, Khan and Dhar (2006) find that consumers who merely imagined volunteering for an environmental cause rated themselves as more prosocial, but afterwards were more likely to make indulgent purchases like choosing designer over practical goods. Clot, Grolleau, and Ibanez (2016) similarly show that subjects who imagine engaging in a pro-environmental task are afterwards more likely to (really) appropriate money from a common pool, particularly when "doing" the task was chosen and unrewarded. Clot, Grolleau, and Ibanez (2018) find similar effects on subsequent dictator-game giving, which becomes more selfish. In a field experiment with over 1,500 workers in a large firm, List and Momeni (2021) find that employees' exposure to a corporate social-responsibility message stating that the firm is giving 5% of its wage bill to a charity "on behalf of all the workers who helped us with [a] project" leads to increases in shirking and cheating. In laboratory experiments in China, Hong, Tirole, and Zhang (2025), demonstrate that while public generosity increases with audience size due to greater reputational incentives, subsequent private giving declines.

Whereas moral cleansing and especially moral licensing, which involves a reversal of behavior, are difficult to account for in (typically static) models of identity based on preferences, they arise quite naturally in the presence of self-image concerns, as explained above and formalized in Bénabou and Tirole (2011) and Hong, Tirole, and Zhang (2025).

3.7.4 Political preferences: predistribution versus redistribution

The importance of self-image concerns also extends to the political domain. As documented by Kuziemko, Longuet-Marx, and Naidu (2023) for the United States, less educated voters significantly prefer *predistribution* policies such as public-sector employment, minimum wages, protective tariffs, subsidies to plants that might otherwise close down, etc., to *redis-*

tribution policies such as taxes and transfers. Conversely, those are preferred by educated voters (and economists), who are firmly opposed to predistribution. Similarly, in almost every country, farmers facing declining prices want production subsidies and import restrictions rather than transfers, or especially monetary incentives to keep their land fallow. A natural explanation, universally proffered by the people in question and amply documented by sociologists, is that there is self-worth and dignity attached to working and feeling that one is a productive member of society, rather than living off "handouts".

Conversely, those receiving welfare benefits are seen as having negative characteristics, such as being lazy or failures (Stuber and Schlesinger, 2006; Chase and Walker, 2013; Li and Walker, 2018). This stigma, both social and internalized, may be a factor contributing to the substantial non-take-up of benefits that many welfare programs suffer from (Currie, 2006), even though participation is essentially private. Also very much in line with the cognitive mechanisms we have seen at work in self-image (such as information avoidance and some motivated memory), minimum wages, price supports, and trade protections are much less salient and repeatedly visible (to self and others) than periodic transfers of having low productivity and being dependent on the generosity of others.

Relatedly, Bénabou and Tirole (2009) show how concerns of dignity and pride can lead to bargaining failures and conflicts even when information is symmetric and there are mutual gains from trade. Individuals and groups with such identity concerns will reject "insultingly low" offers even though they make them materially better off, because accepting them can be seen as admitting to oneself (and possibly others) a low productivity, weakness, or fault (e.g., Babcock and Loewenstein (1997) and Bewley (1999)). In contrast, walking away and "slamming the door" or scapegoating the other party for a bad situation can help preserve a sense of self-worth and/or sustain wishful beliefs about one's abilities and future prospects.

4 Collective identity, groupthink, and stereotypes

Most studies on self-image to date have focused on individually formed beliefs about the self, such as one's morality (prosociality, fairness, honesty), intelligence, religiosity, etc.,

and their implications for individual behavior. As mentioned in the introduction, however, identity is also in large part a social phenomenon, involving how groups see themselves and each other, and how this affects and is affected by their interactions. Accordingly, researchers have started to explore the emergence of collective beliefs and identities as an equilibrium phenomenon of *social cognition*.

On the theoretical side, Bénabou (2013) shows how wishful thinking can become contagious in groups, organizations, or markets. Other members' beliefs about the nature of one's group or the environment it faces shape their actions, which in turn affect everyone's future prospects (expected utilities, returns, risks). These prospects, in turn, determine the relative costs and benefits of realistic versus distorted beliefs about the group and/or the self, whether through anticipatory feelings or self-motivation incentives. Depending on the interaction structure, the contagion of motivated beliefs –horizontally among peers or vertically in a hierarchy– can result in either beneficial group morale or harmful groupthink and ensuing crises.

McGee (2024) shows how incorrect group stereotypes can emerge and persist, due to strategic interactions that make motivated beliefs (or belief inculcation by parents) individually optimal. In competitive environments, publicly known negative stereotypes about another group's abilities discourage its efforts and valuably increase one's own, even though the stigmatized group disagrees with the stereotype. In cooperative environments, positive stereotypes about the other group allow one's own to free ride and force the other to work harder. Although incorrect and potentially harmful socially, these beliefs persist as an equilibrium: if enough of the other members of one's group adopt them, it is individually optimal to do so as well.

Lung (2022) models how people use identity cues, like gender or ethnicity, to guide decisions under uncertainty even when these cues are uninformative. Agents have only noisy signals of their abilities (imperfect self-knowledge again) and infer their chances of success partly from how frequently others in their group have succeeded in the past in it. Although non-Bayesian, this "identity-based inference" helps them optimally self-select into more or less competitive or risky activities, such as STEM fields, finance, etc. Over time, these choices

generate feedback loops: when fewer members of a group participate and succeed in some type of career, future agents from that group update down on their own ability and become less likely to enter, thus perpetuating disparities in participation and outcomes.

On the experimental side, Oprea and Yuksel (2022) explore how belief distortions evolve through bilateral communication. They elicit subjects' beliefs about whether they are in the top or bottom half of their group on an IQ test, then allow them to revise these beliefs continuously, in real time. After a while, each subject is matched with a counterpart in the same performance group, and both can observe the evolution of each other's beliefs. Updating is once again highly asymmetric: more pessimistic subjects in a pair significantly raise their belief toward that of the more optimistic counterpart, while there is almost no movement in the other direction. No such pattern occurs when the pairing is based on a random number, removing the scope for self-image concerns. The authors conclude that the data rejects both standard Bayesian updating and confirmation bias, but instead provides evidence of motivated beliefs about whose estimate is the most accurate.

In Kogan, Schneider, and Weber (2025), subjects compete either individually or in groups of six on how well they do in solving a hard combinatorial problem, with the best-performing individual, or group, receiving a prize. In the group condition, members work jointly on the problem, can communicate by chat, and must submit a common response. In both cases, the individual or group receives a noisy signal about its performance, and each subject's beliefs about it are elicited both before and after receiving the signal. The individual and group conditions both display the standard good news-bad news updating asymmetry, to about the same extent. Thus, collective deliberation fails to correct individuals' motivated belief formation. In another condition, an asset market is introduced, in which group members can trade bets with each other on their group winning the competition. Market prices depart significantly from fundamentals in an overoptimistic direction, with subjects displaying even stronger motivated updating than in the individual condition, as well as a reluctance to bet against the preferred group outcome.

Both papers provide evidence for the self-serving nature of group identity, sustained through mutually amplifying belief distortions, and against the common notion that infor-

mation aggregation through communication or markets naturally correct individual biases –the wisdom-of-crowds hypothesis.

5 Experimental methods to reveal self-image concerns

In this section, we discuss the main methods researchers have used in their experimental designs to identify self-image concerns and common threats to identification.

5.1 Exploiting asymmetric predictions

A key feature of the self-image framework is that it predicts asymmetric response patterns: people should update more following positive than negative feedback, remember positive deeds better than negative ones, and selectively avoid information that could threaten rather than bolster their self-view. Accordingly, a principal method has been to study such asymmetric patterns. As described in Section 3, for motivated belief updating researchers investigate subjects' responses to receiving good versus bad news about their own ability (Eil and Rao, 2011; Möbius et al., 2022), for motivated memory they examine the recall of positive versus negative feedback (Chew, Huang, and Zhao, 2020; Zimmermann, 2020), and for information avoidance they study attitudes towards information about others' payoff versus one's own (Kandul and Ritov, 2017; Moradi, 2018).

5.2 Manipulating the scope for self-image concerns

Creating a placebo condition. Another key feature of the self-image framework is that it predicts behavioral effects only when the choice object, information, or decision outcome is relevant to a person's self-image. This allows researchers to design placebo treatments that are structurally identical to the main treatments, but do not carry any information about ability, moral values, or any other desirable attribute. For instance, as discussed previously, Eil and Rao (2011) and Möbius et al. (2022) have placebo conditions in which subjects perform the same updating tasks as in the IQ treatment and receive information that is quantitatively similar, but not about IQ. Similarly, Exley (2016), Exley and Kessler (2023),

and Bosch-Rosa, Gietl, and Heinemann (2024) design control conditions in which, instead of facing a tradeoff between their own utility and that of others, subjects face a tradeoff that either only concerns their own utility, or only others' utility. In such situations, decisions are uninformative about one's generosity, so self-image concerns cannot operate.

Minimizing scope for alternative explanations. Another strategy employed is to create decision situations that are private and fully anonymous, eliminating social-image concerns. For instance, Bursztyn et al. (2020) and Nasim and Stegmann (2022) offer, respectively, Pakistani men and Pakistani school owners an option that has material benefits (monetary payment, free educational materials) but threatens their political identity (thanking the US government, receiving valuable teaching materials from a progressive NGO). The choice is private, so the fact that a significant fraction of respondents turn the offer down can only be attributed to self-image concerns.[19] A related method is to place subjects in a double-blind setting, where neither decisions nor payments can be linked by the experimenter to subjects' identities. For instance, reporting choices in the die-rolling paradigm Fischbacher and Föllmi-Heusi (2013) cannot be observed by anyone except the subject itself. Similarly, Bénabou et al. (2023) use a double-blind payment procedure (Barmettler, Fehr, and Zehnder, 2012) to study the role of self-image concerns in driving differences between direct-elicitation and multiple price list elicitation methods.

5.3 Directly manipulating self-image concerns

Changing the link between actions and self-image relevant values. Several papers use interventions such as information provision to change subjects' beliefs about the association between actions and values (a_2 and θ). For instance, Mechtenberg et al. (2024) conduct an experiment prior to a ballot on whether to prohibit the dehorning of cattle, which animal-rights activists describe as an animal-harming practice. A randomly selected half of subjects is informed about the empirically documented correlation that people who harm animals also tend to harm humans. The other half is informed about the lack of correlation between

[19] One might be worried about subjects not trusting that their choices are truly anonymous. However, Bursztyn et al. (2020) also conduct a treatment where responses are made public. The fraction refusing to accept the bonus actually decreases, so anonymity concerns cannot explain the refusal of the bonus.

empathy with animals and empathy with humans. The information thus changes the signaling value about altruism conveyed by one's vote. In line with the self-signaling model, the authors find a significant treatment effect both on intentions to vote in favor of prohibiting dehorning and reported voting behavior after the ballot took place.

Tonke (2025) uses a similar method in a field experiment on people's propensity to pay their public-utility bills. Across several treatments, in appeals sent to subscribers, he varies whether the appeal emphasizes that paying one's bills is what a "responsible citizen" should do. In a post-intervention survey, he finds that subjects' self-perception of being a responsible citizen is increased after receiving these appeals, while receiving a neutral appeal does not change self-perceptions. Importantly, the identity appeal leads to significantly higher bill payments during both the three-month intervention and the following eight months.

In another field experiment, Ghosal et al. (2022) study an intervention that attempts to change the values that sex workers in India associate with their occupation. In weekly group sessions over eight weeks, the treatment group received a psychological intervention that aimed to improve their impaired self-image through interactive discussion, verbal persuasion, and role playing. For instance, discussions centered around whether the workers could regard themselves as someone trying to make an honest living, which is better than being a thief or a dishonest person. Relative to a control who received no such intervention, the authors find positive effects on self-image, as measured through survey items such as self-worth and shame about their occupation. Importantly, intervention subsequently increased workers' saving and preventive-health behavior.

Changing beliefs about others' characteristics. Another method of changing subjects' beliefs about the association between actions and values is to provide them with information about the characteristics of other people performing these actions. Schneider (2022) studies how consumers' demand for a product is affected by the ideological values of its customer base. He first threatens subjects' self-image by having them take and see the results of an implicit association test (IAT) that has been argued to measure implicit racism. They then face a choice between two consumption goods, after observing the choices made by 10 other individuals. In the treatment, subjects learn that those individuals are right-wing

extremists, while in the control group, they receive no further information. He finds that subjects' willingness to pay for the consumption good that is associated with right-wing extremists through their choices is significantly lower. As the experiment is double blind, the treatment effect is best explained by subjects having uncertainty about their own type, and consumption choices acting as signals about type.[20]

A similar technique is used by Henkel and Zimpelmann (2025) in the context of financial decisions. In their experiment, subjects decide whether to invest in stocks or a safe asset. Prior to their choice, they receive information about the charitable-donation behavior of 10 stockholders and 10 non-stockholders. These 20 individuals are randomly sampled from a larger population, which generates exogenous variation in the difference in donation behavior between stockholders and non-stockholders. The authors first show that subjects change their perceptions of stockholders based on the information: most initially think stockholders are greedy and selfish, which are highly identity-relevant traits, as the authors verify. Second, given information that shows positive donation behavior of stockholders, subjects adjust their beliefs. Importantly, subjects' investment choices are influenced by this identity-relevant information: the less stockholders donate relative to non-stockholders, the less likely subjects are to choose the stock option.

Changing self-image through interventions. Somewhat less common in economics are interventions that directly manipulate self-image or its salience. Such interventions are more common in psychology, for instance in the large literature on self-esteem (e.g., Sweeney and Moyer, 2015). One exception is a natural experiment by Bursztyn et al. (2018), who study the demand for status credit cards in a sample of Indonesian bank customers. They first show that "platinum" credit cards, which signal higher status, have higher demand than control ones that provide the same material benefits but do not carry the "platinum" label. To investigate the role of self-image in driving this demand, they experimentally increase subjects' self-esteem by asking them to describe a recent experience from their personal or

[20]Teyssier, Etilé, and Combris (2015) conduct an experiment in which they elicit subjects' WTP for fair-trade goods. In one condition, subjects learn the WTP of another subject before making their own, private choice. When learning about a WTP that is higher than their own, they do not increase their own WTP, while they significantly adjust downwards if the other WTP is lower.

professional life that made them feel particularly proud. The idea is that higher self-esteem lowers the need for self-reassurance through consumption behavior. The treatment indeed increased subjects' self-esteem and subsequently lowered their demand for platinum credit cards.

Also in the domain of economic status, Bottan et al. (2025) use hypothetical discrete-choice experiments to estimate, on a large representative sample of the US population, people's willingness to pay (WTP) for a higher self image (by reducing other citizens' average income) and social image (by increasing others' perceptions of one's income). In their preferred estimates, the value of self-image is at most 19% of that of social image, albeit with a lot of individual heterogeneity. In our view, their design has certain features that could underestimate self-image and overestimate social image, but it provides an innovative first step at disentangling the two.[21]

Falk (2021) implements a literal self-image manipulation. Participants choose between receiving a monetary reward at the cost of another person experiencing a medically harmless but painful electric shock, or forgoing the reward to prevent the shock. Next to the computer screen where subjects make their decisions, they are presented with either a live feed of themselves, themselves in a mirror, a neutral video feed, or no further content. Accordingly, in the first two treatments, subjects are being watched by themselves when making decisions. This has an effect on behavior: subjects are significantly less likely to selfishly choose the reward and shock to the other person in those two conditions.

5.4 Threats to identifying self-image concerns

Confusion and inattention. One possible confounding factor is that subjects may be confused about, or inattentive to, the experimental instructions. An extensive literature has shown that these factors can influence behavior in many of the experimental paradigms that papers studying self-image concerns are build upon.[22] The fraction of subjects appearing to

[21] Impoverishing the society one lives in has negative externalities that subjects may also value, and incentivizing the social-image measure by having them predict what the average prediction of others' WTP will be (Krupka and Weber (2013) method) could generate coordination on an overestimate if everyone thinks (and thinks that others think) that others are more vain than they themselves are, as is likely the case.

[22] Houser and Kurzban (2002) show how confusion affects behavior in the public-goods game, Koppel et al. (2025) in the dictator game and other social games, Cason and Plott (2014) for the Becker-DeGroot-Marschak

avoid free information may thus be artificially inflated if some of them are inattentive and make random choices. Similarly, when comparing hypothetical and incentivized choices, misunderstanding the incentives may induce differences in choices that are unrelated to self-image concerns.

Two important features of the predictions from self-image models allow researchers to alleviate concerns about confusion and random choice. First, as discussed previously, many predictions involve asymmetric behavioral patterns, whereas confusion-induced random choice would lead to symmetric errors. For instance, subjects being confused about a classical balls-and-urns updating task would not lead them to update differently in response to positive versus negative news. Similarly, inattention to IQ signals cannot explain why subjects remember positive better than negative ones. Second, using placebo conditions is an effective way to address concerns about confusion or inattention. A well-designed placebo is structurally identical to the main treatment, ensuring that any misunderstanding or inattentiveness – whether related to the choice object, the information provided, or the decision outcome – affects behavior in both treatment and placebo equally. Accordingly, comparing outcomes across the two conditions effectively controls for these confounds.

Experimenter demand. Another threat to the proper identification of self-image effects is experimenter-demand effects – the concern that participants may try to guess the experimenter's objective from the instructions, and then act accordingly. Hence, instead of self-signaling values with their behavior, they may be engaging in social signaling. Such an effect is more likely to occur when the experimental design makes the goal of the experiment apparent, for instance by providing subjects with morally loaded information and then measuring their updating behavior.

However, in many of the settings discussed in Section 3, experimenter-demand effects are unlikely to occur. This is because thus such experiments often rely on comparisons that are not inferable from the instructions, or because demand effects would actively work against the hypothesis. In Zimmermann (2020), for instance, subjects may infer that the experimenter expects them to update their prior: after all, they have received some infor-

(BDM) method, and Danz, Vesterlund, and Wilson (2022) for the binarized scoring rule.

mation. However, as the experiment shows, after one month, subjects only update after receiving positive, but not negative information. In general, papers that explicitly study the scope of experimenter- demand have shown that such effects often have little impact on responses, even when such a demand is explicitly stated, or questions are hypothetical (De Quidt, Haushofer, and Roth, 2018; Mummolo and Peterson, 2019; Danz et al., 2023)

As experimenter-demand effects are a general concern in experimental economics, the literature has also developed several solutions to deal with them. One is employing a double-blind procedure (Barmettler, Fehr, and Zehnder, 2012), which makes it clear to subjects that the experimenter is interested in their personal and "natural" behavior. Another is the use of an obfuscated follow-up (Haaland and Roth, 2020). Using an independent layout, invitation, and design, the researcher recontacts participants after some time to measure different outcomes. If the obfuscation is successful and subjects perceive no connection between the main experiment and follow-up, demand effects cannot drive a treatment effect found in the latter. For instance, Henkel and Zimpelmann (2025) use this method to show that self-image-relevant information persistently changed attitudes with respect to stock investments. Lastly, one can add a condition that contains an explicit demand in order to benchmark the influence of potential demand effects (De Quidt, Haushofer, and Roth, 2018). For instance, Schneider (2022)'s study on consumption goods includes a robustness treatment with the added sentence "We expect that participants who are shown these instructions will specify a lower maximum price than they normally would." He finds that the sentence has no effect on demand.

6 Conclusion

As shown in this survey, self-image concerns pervade many aspects of people's cognition and behavior. They lead to information avoidance, asymmetric recall, rationalizations, and belief updating after identity-threatening versus identity-enhancing news, and to self-signaling actions. Identity-protective beliefs, in turn, have important effects on a host of economically important behaviors, among which: (i) altruism, charitable giving, and environmental

responsibility; (ii) lying, conflicts of interest, and other moral choices; (iii) avoidance or delegation of decisions; (iv) effort, willingness to compete, and behavior in strategic interactions; (v) financial risk taking and investment; (vi) health and fertility decisions; (vii) failures of the “wisdom of crowds", as individual belief distortions can be mutually amplifying, resulting in biased group decisions; (viii) inefficient bargaining failures and conflicts; (ix) political attitudes toward redistributive transfers versus price and wage supports.

There are many fruitful directions for further research. A first one, as on many other topics, is that of more field experiments, which are still a minority compared to those conducted online or in the laboratory. A second one, on which there is a nascent literature, is the emergence and often fact-proof persistence of collective self-image and group identities, arising through complementarities and contagion in individuals’ motivated views of themselves and their environment.

References

Abeler, Johannes, Daniele Nosenzo, and Collin Raymond (2019). "Preferences for Truth-Telling." *Econometrica* 87 (4): 1115–53. [35]

Adena, Maja, and Steffen Huck (2020). "Online Fundraising, Self-Image, and the Long-Term Impact of Ask Avoidance." *Management Science* 66 (2): 722–43. [19]

Ahumada, Beatriz, Yufei Chen, Neeraja Gupta, Kelly Hyde, Marissa Lepper, Will Mathews, Neil Silveus, Lise Vesterlund, Taylor Weidman, Alistair Wilson, K Pun Winichakul, and Liyang Zhou (2022). "Well Excuse Me! Replicating and Connecting Excuse-Seeking Behaviors." *Working Paper*, [37]

Akerlof, George A, and Rachel E Kranton (2005). "Identity and the Economics of Organizations." *Journal of Economic Perspectives* 19 (1): 9–32. [1, 2]

Akerlof, George A, and Rachel E Kranton (2010). "Identity economics: How our identities shape our work, wages, and well-being." In *Identity Economics*. Princeton University Press. [1]

Akerlof, George A. (1989). "The Economics of Illusion." *Economics and Politics* 1 (1): 1–15. [2]

Akerlof, George A., and William T. Dickens (1982). "The Economic Consequences of Cognitive Dissonance." *American Economic Review* 72 (3): 307–19. [2]

Akerlof, George A., and Rachel E. Kranton (2000). "Economics and Identity." *Quarterly Journal of Economics* 115 (3): 715–53. [1]

Alger, Ingela, and Jörgen W. Weibull (2013). "Homo Moralis: Preference Evolution under Incomplete Information and Assortative Matching." *Econometrica* 81 (6): 2269–302. [4]

Amelio, Alessandro, and Felix Zimmermann (2023). "Motivated Memory in Economics—A Review." *Games* 14 (1): 15. [6, 27]

Andreoni, James, Justin M. Rao, and Hannah Trachtman (2017). "Avoiding the Ask: A Field Experiment on Altruism, Empathy, and Charitable Giving." *Journal of Political Economy* 125 (3): 625–53. [20]

Arendt, Hannah (1972). "Lying in Politics: Reflections on the Pentagon Papers." In *Crises of the Republic: Lying in Politics, Civil Disobedience, On Violence, Thoughts on Politics and Revolution*. New York: Harcourt Brace Jovanovich, 1–47. Quotation on p. 6: "The deceivers began by deceiving themselves." [9]

Aronson, Elliot (1969). "The Theory of Cognitive Dissonance: A Current Perspective." In *Advances in Experimental Social Psychology*. Leon Berkowitz, ed. vol. 4, Academic Press, 1–34. [2]

Asch, Solomon E. (1955). "Opinions and Social Pressure." *Scientific American* 193 (5): 31–35. [1]

Babcock, Linda, and George Loewenstein (1997). "Explaining Bargaining Impasse: The Role of Self-Serving Biases." *Journal of Economic Perspectives* 11 (1): 109–26. [4, 42]

Babcock, Linda, George Loewenstein, Samuel Issacharoff, and Colin Camerer (1995). "Biased Judgments of Fairness in Bargaining." *American Economic Review* 85 (5): 1337–43. [24]

Barmettler, Franziska, Ernst Fehr, and Christian Zehnder (2012). "Big Experimenter Is Watching You! Anonymity and Prosocial Behavior in the Laboratory." *Games and Economic Behavior* 75 (1): 17–34. [14, 46, 51]

Battaglini, Marco, Roland Bénabou, and Jean Tirole (2005). "Self-Control in Peer Groups." *Journal of Economic Theory* 123 (2): 105–34. [5]

Baumeister, Roy F. (2022). *The Self Explained: Why and How We Become Who We Are*. 1st. New York, NY: Guilford Press, 420. [1, 2]

Bekkers, René, and Pamala Wiepking (2011). "Accuracy of Self-Reports on Donations to Charitable Organizations." *Quality & Quantity* 45 (6): 1369–83. [40]

Bem, Daryl J. (1972). "Self-Perception Theory." In *Advances in Experimental Social Psychology*. Leonard Berkowitz, ed. New York: Academic Press, 1–62. [2]

Bénabou, Roland (2013). "Groupthink: Collective Delusions in Organizations and Markets." *Review of Economic Studies* 80 (2): 429–62. Earlier version as NBER WP No. 14764 (2009). [43]

Bénabou, Roland (2015). "The economics of motivated beliefs." *Revue d'économie politique* 125 (5): 665–85. Jean-Jacques Laffont Lecture. [6]

Bénabou, Roland, Armin Falk, and Luca Henkel (2024). "Ends versus Means: Kantians, Utilitarians and Moral Decisions." *Working Paper*, [4, 33, 39]

Bénabou, Roland, Armin Falk, Luca Henkel, and Jean Tirole (2023). "Eliciting Moral Preferences under Image Concerns: Theory and Experiment." *Working Paper*, [39, 46]

Bénabou, Roland, Ania Jaroszewicz, and George Loewenstein (2025). "It hurts to ask." *European Economic Review* 171: 104911. [38]

Bénabou, Roland, and Jean Tirole (2002). "Self-Confidence and Personal Motivation." *Quarterly Journal of Economics* 117 (3): 871–915. [2, 3, 25, 27]

Bénabou, Roland, and Jean Tirole (2004). "Willpower and personal rules." *Journal of Political Economy* 112 (4): 848–86. [2, 3, 27, 33]

Bénabou, Roland, and Jean Tirole (2006). "Incentives and Prosocial Behavior." *American Economic Review* 96 (5): 1652–78. [3, 33]

Bénabou, Roland, and Jean Tirole (2009). "Over My Dead Body: Bargaining and the Price of Dignity." *American Economic Review* 99 (2): 459–65. [4, 42]

Bénabou, Roland, and Jean Tirole (2011). "Identity, Morals, and Taboos: Beliefs as Assets." *Quarterly Journal of Economics* 126 (2): 805–55. [2, 3, 5, 9, 13, 27, 33, 41]

Bénabou, Roland, and Jean Tirole (1, 2016). "Mindful Economics: The Production, Consumption, and Value of Beliefs." *Journal of Economic Perspectives* 30 (3): 141–64. [2, 6]

Bertrand, Marianne, Emir Kamenica, and Jessica Pan (2015). "Gender Identity and Relative Income within Households." *Quarterly Journal of Economics* 130 (2): 571–614. [2]

Bewley, Truman F. (1999). *Why Wages Don't Fall During a Recession*. Cambridge, MA: Harvard University Press, 527. [42]

Bisin, Alberto, and Thierry Verdier (2001). "The Economics of Cultural Transmission and the Dynamics of Preferences." *Journal of Economic Theory* 97 (2): 298–319. [1]

Bodner, Ronit, and Drazen Prelec (2003). "Self-Signaling and Diagnostic Utility in Everyday Decision Making." In *The Psychology of Economic Decisions*. Isabelle Brocas and Juan D. Carrillo, ed. Oxford, UK: Oxford University Press, 105–26. [3, 33]

Bolte, Lukas, and Tony Q. Fan (2023). "Motivated Mislearning: The Case of Correlation Neglect." *Working Paper*, [3, 22, 37]

Bönisch, Felix, Tobias König, Sebastian Schweighofer-Kodritsch, and Georg Weizsäcker (2024). "Beliefs as a Means of Self-Control? Evidence from a Dynamic Student Survey." *Working Paper*, [25]

Bonomi, Giampaolo, Nicola Gennaioli, and Guido Tabellini (2021). "Identity, Beliefs, and Political Conflict." *Quarterly Journal of Economics* 136 (4): 2371–411. [2]

Borisova, Polina, and Nikhil Vellodi (2024). "A Theory of Self-Prospection." *Working Paper*, [2]

Bosch-Rosa, Ciril, Daniel Gietl, and Frank Heinemann (2024). "Risk Taking Under Limited Liability and Moral Hazard: Quantifying the Role of Motivated Beliefs." *Management Science*, (4): mnsc.2021.03947. [24, 45]

Bottan, Nicolas L., Ricardo Perez-Truglia, Hitoshi Shigeoka, and Katsunori Yamada (2025). "Feeling Rich or Looking Rich? Quantifying Self-Image and Social-Image Motives." *Working Paper*, [49]

Brañas-Garza, Pablo, Marisa Bucheli, Maria Paz Espinosa, and Teresa García-Muñoz (2011). "Moral Cleansing and Moral Licenses: Experimental Evidence." *Economics & Philosophy* 27 (2): 199–212. [41]

Brekke, Kjell A., Snorre Kverndokk, and Karine Nyborg (2003). "An Economic Model of Moral Motivation." *Journal of Public Economics* 87 (9-10): 1967–83. [1]

Broberg, Tomas, Tore Ellingsen, and Magnus Johannesson (2007). "Is Generosity Involuntary?" *Economics Letters* 94 (1): 32–37. [20]

Brock, J. Michelle, Andreas Lange, and Erkut Y Ozbay (2013). "Dictating the Risk: Experimental Evidence on Giving in Risky Environments." *American Economic Review* 103 (1): 415–37. [36]

Bursztyn, Leonardo, Michael Callen, Bruno Ferman, Saad Gulzar, Ali Hasanain, and Noam Yuchtman (2020). "Political Identity: Experimental Evidence on Anti-Americanism in Pakistan." *Journal of the European Economic Association* 18 (5): 2532–60. [33, 46]

Bursztyn, Leonardo, Bruno Ferman, Stefano Fiorin, Martin Kanz, and Gautam Rao (2018). "Status Goods: Experimental Evidence from Platinum Credit Cards." *Quarterly Journal of Economics* 133 (3): 1561–95. [48]

Bursztyn, Leonardo, Ingar Haaland, Nicolas Röver, and Christopher Roth (2025). "The Social Desirability Atlas." *Working Paper*, [4, 40]

Buser, Thomas, Leonie Gerhards, and Joël Van Der Weele (2018). "Responsiveness to Feedback as a Personal Trait." *Journal of Risk and Uncertainty* 56 (2): 165–92. [23]

Cain, Daylian M., Jason Dana, and George E. Newman (2014). "Giving Versus Giving In." *Academy of Management Annals* 8 (1): 505–33. [20]

Caplin, Andrew, and John V. Leahy (2001). "Psychological Expected Utility Theory and Anticipatory Feelings." *Quarterly Journal of Economics* 116 (1): 55–79. [2]

Cappelen, Alexander W., Benjamin Enke, and Bertil Tungodden (2025). "Universalism: Global Evidence." *American Economic Review* 115 (1): 43–76. [4]

Carlson, Ryan W., Michel André Maréchal, Bastiaan Oud, Ernst Fehr, and Molly J. Crockett (2020). "Motivated Misremembering of Selfish Decisions." *Nature Communications* 11 (1): 2100. [11, 31]

Carrillo, Juan D., and Thomas Mariotti (2000). "Strategic Ignorance as a Self-Disciplining Device." *Review of Economic Studies* 67 (3): 529–44. [3, 13]

Cason, Timothy N., and Charles R. Plott (2014). "Misconceptions and Game Form Recognition: Challenges to Theories of Revealed Preference and Framing." *Journal of Political Economy* 122 (6): 1235–70. [49]

Chandrasekhar, Arun G., Benjamin Golub, and He Yang (2018). "Signaling, Shame, and Silence in Social Learning." *Working Paper*, [38]

Chase, Elaine, and Robert Walker (2013). "The Co-construction of Shame in the Context of Poverty: Beyond a Threat to the Social Bond." *Sociology* 47 (4): 739–54. [42]

Chen, Si, and Hannah Schildberg-Hörisch (2019). "Looking at the Bright Side: The Motivation Value of Overconfidence." *European Economic Review* 120: 103302. [25]

Chew, Soo Hong, Wei Huang, and Xiaojian Zhao (2020). "Motivated False Memory." *Journal of Political Economy* 128 (10): 3913–39. [3, 10, 29, 45]

Cialdini, Robert B., and Noah J. Goldstein (2004). "Social Influence: Compliance and Conformity." *Annual Review of Psychology* 55 (1): 591–621. [1]

Clot, Sophie, Gilles Grolleau, and Lisette Ibanez (2016). "Do Good Deeds Make Bad People?" *European Journal of Law and Economics* 42 (3): 491–513. [41]

Clot, Sophie, Gilles Grolleau, and Lisette Ibanez (2018). "Shall We Pay All? An Experimental Test of Random Incentivized Systems." *Journal of Behavioral and Experimental Economics* 73: 93–98. [41]

Colley, Rachel C., Gregory Butler, Didier Garriguet, Stephanie A. Prince, and Karen C. Roberts (2018). "Comparison of self-reported and accelerometer-measured physical activity in Canadian adults." *Health Reports* 29 (12): 3–15. [40]

Conrads, Julian, and Bernd Irlenbusch (2013). "Strategic Ignorance in Ultimatum Bargaining." *Journal of Economic Behavior & Organization* 92 (8): 104–15. [18]

Cooley, Charles H. (1902). *Human Nature and the Social Order*. New York: Charles Scribner's Sons. [4]

Coutts, Alexander (2019). "Good News and Bad News Are Still News: Experimental Evidence on Belief Updating." *Experimental Economics* 22 (2): 369–95. [23]

Currie, Janet (2006). "The Take-up of Social Benefits." In *Poverty, the Distribution of Income, and Public Policy*. Alan Auerbach, David Card, and John Quigley, ed. New York: Russell Sage, 80–148. [42]

Damgaard, Mette Trier, and Christina Gravert (2018). "The Hidden Costs of Nudging: Experimental Evidence from Reminders in Fundraising." *Journal of Public Economics* 157 (1): 15–26. [20]

Dana, Jason, Daylian M. Cain, and Robyn M. Dawes (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. [19]

Dana, Jason, Roberto A. Weber, and Jason Xi Kuang (2007). "Exploiting Moral Wiggle Room: Experiments Demonstrating an Illusory Preference for Fairness." *Economic Theory* 33 (1): 67–80. [10, 14, 15, 32, 37]

Danz, David, Marissa Lepper, Guillermo Lezama, Priyoma Mustafi, Lise Vesterlund, Alistair Wilson, and K. Pun Winichakul (2023). "Simon Doesn't Say: Minimal Qualitative Distortions from Experimenter Demand." *Working Paper*, [51]

Danz, David, Lise Vesterlund, and Alistair J. Wilson (2022). "Belief Elicitation and Behavioral Incentive Compatibility." *American Economic Review* 112 (9): 2851–83. [50]

De Quidt, Jonathan, Johannes Haushofer, and Christopher Roth (2018). "Measuring and Bounding Experimenter Demand." *American Economic Review* 108 (11): 3266–302. [51]

DellaVigna, Stefano, John A. List, and Ulrike Malmendier (2012). "Testing for Altruism and Social Pressure in Charitable Giving." *Quarterly Journal of Economics* 127 (1): 1–56. [20, 38]

Di Tella, Rafael, Ricardo Perez-Truglia, Andres Babino, and Mariano Sigman (2015). "Conveniently Upset: Avoiding Altruism by Distorting Beliefs about Others' Altruism." *American Economic Review* 105 (11): 3416–42. [37]

Dillenberger, David, and Philipp Sadowski (2012). "Ashamed to Be Selfish." *Theoretical Economics* 7 (1): 99–124. [19]

Drobner, Christoph (2022). "Motivated Beliefs and Anticipation of Uncertainty Resolution." *American Economic Review: Insights* 4 (1): 89–105. [22]

Drobner, Christoph, and Sebastian J. Goerg (2024). "Motivated Belief Updating and Rationalization of Information." *Management Science* 70 (7): 4583–92. [23]

Dubé, Jean-Pierre, Xueming Luo, and Zheng Fang (2017). "Self-Signaling and Prosocial Behavior: A Cause Marketing Experiment." *Marketing Science* 36 (2): 161–86. [33]

Dufwenberg, Martin, and Martin A. Dufwenberg (2018). "Lies in Disguise – A Theoretical Analysis of Cheating." *Journal of Economic Theory* 175 (5): 248–64. [34]

Dutz, Deniz, Ingrid Huitfeldt, Santiago Lacouture, Magne Mogstad, Alexander Torgovitsky, and Winnie Van Dijk (2021). "Selection in Surveys: Using Randomized Incentives to Detect and Account for Nonresponse Bias." *Working Paper*, [40]

Duval, T. Shelley, and Robert A. Wicklund (1972). *A Theory of Objective Self-Awareness*. New York: Academic Press. [2]

Eil, David, and Justin M Rao (2011). "The Good News-Bad News Effect: Asymmetric Processing of Objective Information about Yourself." *American Economic Journal: Microeconomics* 3 (2): 114–38. [10, 11, 21–23, 45]

Engel, Christoph (2011). "Dictator Games: A Meta Study." *Experimental Economics* 14 (4): 583–610. [14]

Enke, Benjamin, and Thomas Graeber (2023). "Cognitive uncertainty." *Quarterly Journal of Economics* 138 (4): 2021–67. [5]

Enke, Benjamin, Ricardo Rodríguez-Padilla, and Florian Zimmermann (2023). "Moral Universalism and the Structure of Ideology." *Review of Economic Studies* 90 (4): 1934–62. [4]

Epley, Nicholas, and Thomas Gilovich (2016). "The Mechanics of Motivated Reasoning." *Journal of Economic Perspectives* 30 (3): 133–40. [20]

Epperson, Raphael, and Andreas Gerster (2024). "Willful Ignorance and Moral Behavior." *Working Paper*, [18]

Exley, Christine L, and Judd B Kessler (2023). "Information Avoidance and Image Concerns." *The Economic Journal* 133 (656): 3153–68. [16, 17, 45]

Exley, Christine L, and Judd B Kessler (2024). "Motivated errors." *American Economic Review* 114 (4): 961–87. [3, 37]

Exley, Christine L. (2016). "Excusing Selfishness in Charitable Giving: The Role of Risk." *Review of Economic Studies* 83 (2): 587–628. [4, 36, 37, 45]

Exley, Christine L. (2020). "Using Charity Performance Metrics as an Excuse Not to Give." *Management Science* 66 (2): 553–63. [37]

Falk, Armin (2021). "Facing Yourself - A Note on Self-Image." *Journal of Economic Behavior and Organization* 186: 724–34. [49]

Feiler, Lauren (2014). "Testing Models of Information Avoidance with Binary Choice Dictator Games." *Journal of Economic Psychology* 45 (12): 253–67. [16]

Fershtman, Chaim, Uri Gneezy, and Moshe Hoffman (2011). "Taboos and Identity: Considering the Unthinkable." *American Economic Journal: Microeconomics* 3 (2): 139–64. [3]

Fischbacher, Urs, and Franziska Föllmi-Heusi (2013). "Lies in Disguise—An Experimental Study on Cheating." *Journal of the European Economic Association* 11 (3): 525–47. [35, 46]

Freundt, Jana, and Andreas Lange (2017). "On the Determinants of Giving under Risk." *Journal of Economic Behavior & Organization* 142 (10): 24–31. [36]

Friston, Karl (2010). "The Free-Energy Principle: a unified brain theory?" *Nature Reviews Neuroscience* 11 (2): 127–38. [2]

Fudenberg, Drew, Giacomo Lanzani, and Philipp Strack (2024). "Selective-Memory Equilibrium." *Journal of Political Economy* 132 (12): 3978–4020. [27]

Garcia, Thomas, Sebastien Massoni, and Marie Claire Villeval (2020). "Ambiguity and Excuse-Driven Behavior in Charitable Giving." *European Economic Review1* 124: 103412. [37]

Ghosal, Sayantan, Smarajit Jana, Anandi Mani, Sandip Mitra, and Sanchari Roy (2022). "Sex Workers, Stigma, and Self-Image: Evidence from Kolkata Brothels." *Review of Economics and Statistics* 104 (3): 431–48. [47]

Ging-Jehli, Nadja R., Florian H. Schneider, and Roberto A. Weber (2020). "On Self-Serving Strategic Beliefs." *Games and Economic Behavior* 122 (7): 341–53. [37]

Gneezy, Ayelet, Uri Gneezy, Gerhard Riener, and Leif D. Nelson (2012). "Pay-What-You-Want, Identity, and Self-Signaling in Markets." *Proceedings of the National Academy of Sciences* 109 (19): 7236–40. [34]

Gneezy, Uri (2005). "Deception: The Role of Consequences." *American Economic Review* 95 (1): 384–94. [35]

Gneezy, Uri, Agne Kajackaite, and Joel Sobel (2018). "Lying Aversion and the Size of the Lie." *American Economic Review* 108 (2): 419–53. [34]

Gneezy, Uri, and Aldo Rustichini (2011). "Taboos and Identity: Considering the Unthinkable." *American Economic Journal: Microeconomics* 3 (2): 139–64. [13]

Gneezy, Uri, Silvia Saccardo, Marta Serra-Garcia, and Roel van Veldhuizen (2020). "Bribing the Self." *Games and Economic Behavior* 120: 311–24. [24]

Gödker, Katrin, Peiran Jiao, and Paul Smeets (2025). "Investor Memory." *Review of Financial Studies* 38 (6): 1595–640. [32]

Goette, Lorenz, and Egon Tripodi (2020). "Does Positive Feedback of Social Impact Motivate Prosocial Behavior? A Field Experiment with Blood Donors." *Journal of Economic Behavior & Organization* 175 (7): 1–8. [34]

Golman, Russell, David Hagmann, and George Loewenstein (2017). "Information Avoidance." *Journal of Economic Literature* 55 (1): 96–135. [6, 13]

González-Jiménez, Víctor (2022). "Social status and motivated beliefs." *Journal of Public Economics* 211: 104662. [3]

Graham, Jesse, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P. Wojcik, and Peter H. Ditto (2013). "Moral Foundations Theory." In *Advances in Experimental Social Psychology*. vol. 47, Elsevier, 55–130. [4]

Graham, Jesse, Brian A. Nosek, Jonathan Haidt, Ravi Iyer, Spassena Koleva, and Peter H. Ditto (2011). "Mapping the Moral Domain." *Journal of Personality and Social Psychology* 101 (2): 366–85. [4]

Grossman, Gene M, and Elhanan Helpman (2020). "Identity Politics and Trade Policy." *Review of Economic Studies* 88 (3): 1101–26. [2]

Grossman, Zachary (2014). "Strategic Ignorance and the Robustness of Social Preferences." *Management Science* 60 (11): 2659–65. [17]

Grossman, Zachary (2015). "Self-Signaling and Social-Signaling in Giving." *Journal of Economic Behavior & Organization* 117 (9): 26–39. [40]

Grossman, Zachary, and Joël J. Van Der Weele (2017). "Self-Image and Willful Ignorance in Social Decisions." *Journal of the European Economic Association* 15 (1): 173–217. [32]

Haaland, Ingar, and Christopher Roth (2020). "Labor Market Concerns and Support for Immigration." *Journal of Public Economics* 191 (June): [51]

Hagenbach, Jeanne, Nicolas Jacquemet, and Philipp Sternal (2025). "The Motivated Memory of Noise." *Games and Economic Behavior* 152: 257–75. [30]

Hagenbach, Jeanne, and Frédéric Koessler (2022). "Selective Memory of a Psychological Agent." *European Economic Review* 142: 104012. [27]

Hagenbach, Jeanne, and Charlotte Saucet (2025). "Motivated skepticism." *Review of Economic Studies* 92 (3): 1882–919. [3, 22, 24]

Hamman, John R., George Loewenstein, and Roberto A. Weber (2010). "Self-Interest through Delegation: An Additional Rationale for the Principal-Agent Relationship." *American Economic Review* 100 (4): 1826–46. [11, 38]

Heese, Carl, and Si Chen (2025). "Fishing for Good News: Motivated Information Acquisition." *Journal of Political Economics: Microeconomics* forthcoming: [23]

Heidhues, Paul, Botond Kőszegi, and Philipp Strack (2023). "Misinterpreting yourself." *Working Paper*, [4]

Heinicke, Franziska (2018). "Self-Image and Hypothetical Bias in Willingness to Pay Elicitation: The Case of a Private Good." *Working Paper*, [39]

Henkel, Luca, and Christian Zimpelmann (2025). "Proud to Not Own Stocks: How Identity Shapes Financial Decisions." *Working Paper*, [4, 48, 51]

Hestermann, Nina, and Yves Le Yaouanq (2021). "Experimentation with self-serving attribution biases." *American Economic Journal: Microeconomics* 13 (3): 198–237. [4]

Hippel, William von, and Robert Trivers (2011). "The evolution and psychology of self-deception." *Behavioral and Brain Sciences* 34 (1): 1–56. [3, 26]

Hong, Fuhai, Jean Tirole, and Chen Zhang (2025). "Moral Licensing: Prosocial Behavior in Public and Private Spheres." Working Paper. [3, 41]

Houser, Daniel, and Robert Kurzban (2002). "Revisiting Kindness and Confusion in Public Goods Experiments." *American Economic Review* 92 (4): 1062–69. [49]

Huffman, David B., Collin Raymond, and Julia Shvets (2022). "Persistent Overconfidence and Biased Memory: Evidence from Managers." *American Economic Review* 112 (10): 3141–75. [10, 29]

Johansson-Stenman, Olof, and Henrik Svedsäter (2012). "Self-Image and Valuation of Moral Goods: Stated versus Actual Willingness to Pay." *Journal of Economic Behavior & Organization* 84 (3): 879–91. [39]

Kandul, Serhiy, and Ilana Ritov (2017). "Close Your Eyes and Be Nice: Deliberate Ignorance behind pro-Social Choices." *Economics Letters* 153: 54–56. [17, 45]

Khalmetski, Kiryl, and Dirk Sliwka (2019). "Disguising Lies—Image Concerns and Partial Lying in Cheating Games." *American Economic Journal: Microeconomics* 11 (4): 79–110. [34]

Khan, Uzma, and Ravi Dhar (2006). "Licensing Effect in Consumer Choice." *Journal of Marketing Research* 43 (2): 259–66. [41]

Knutsson, Mikael, Peter Martinsson, and Conny Wollbrant (2013). "Do People Avoid Opportunities to Donate?" *Journal of Economic Behavior & Organization* 93 (9): 71–77. [20]

Kogan, Shimon, Florian H. Schneider, and Roberto A. Weber (2025). "Self-Serving Biases in Beliefs about Collective Outcomes." *Games and Economic Behavior*, Journal Pre-proof; available online. [44]

Konow, James (2000). "Fair Shares: Accountability and Cognitive Dissonance in Allocation Decisions." *American Economic Review* 90 (4): 1072–91. [23]

Koppel, Lina, David Andersson, Magnus Johannesson, Eirik Strømland, and Gustav Tinghög (2025). "Comprehension in Economic Games." *Journal of Economic Behavior & Organization* 234: 107039. [16, 49]

Köszegi, Botond (2006). "Ego utility, overconfidence, and task choice." *Journal of the European Economic Association* 4 (4): 673–707. [2, 3]

Kőszegi, Botond, George Loewenstein, and Takeshi Murooka (2022). "Fragile self-esteem." *Review of Economic Studies* 89 (4): 2026–60. [3]

Krawczyk, Michal, and Fabrice Le Lec (2010). "'Give Me a Chance!' An Experiment in Social Decision under Risk." *Experimental Economics* 13 (4): 500–11. [36]

Krupka, Erin L., and Roberto A. Weber (2013). "Identifying Social Norms Using Coordination Games." *Journal of the European Economic Association* 11 (3): 495–524. [49]

Kunda, Ziva (1990). "The Case for Motivated Reasoning." *Psychological Bulletin* 108 (3): 480–98. [2, 20]

Kuziemko, Ilyana, Nicolas Longuet-Marx, and Suresh Naidu (2023). ""Compensate the Losers?" Economic Policy and Partisan Realignment in the US." NBER Working Paper 31794. National Bureau of Economic Research. [41]

Larson, Tara, and C. Monica Capra (2009). "Exploiting Moral Wiggle Room: Illusory Preference for Fairness? A Comment." *Judgment and Decision Making* 4 (6): 467–74. [16]

Lazear, Edward P., Ulrike Malmendier, and Roberto A. Weber (2012). "Sorting in Experiments with Application to Social Preferences." *American Economic Journal: Applied Economics* 4 (1): 136–63. [19]

Le Maux, Benoit, David Masclet, and Sarah Necker (2021). "Monetary Incentives and the Contagion of Unethical Behavior." *Working Paper*, [36]

Leary, Mark R., and Deborah L. Downs (1995). "Interpersonal functions of the self-esteem motive: The self-esteem system as a sociometer." In *Efficacy, Agency, and Self-Esteem*. Michael H. Kernis, ed. New York: Plenum Press, 123–44. [4]

Li, Mianguan, and Robert Walker (2018). "On the Origins of Welfare Stigma: Comparing Two Social Assistance Schemes in Rural China." *Critical Social Policy* 38 (4): 667–87. [42]

Lindquist, Samuel, Silvia Saccardo, and Marta Serra-Garcia (2025). "Belief Management and Unethical Behavior." SSRN Working Paper, April 30, 2025. Chapter in *Research Handbook on Unethical Behavior*, forthcoming. [6]

List, John A, and Craig A Gallet (2001). "What Experimental Protocol Influence Disparities Between Actual and Hypothetical Stated Values?" *Environmental and Resource Economics* 20: 241–54. [39]

List, John A, and Fatemeh Momeni (2021). "When corporate social responsibility backfires: Evidence from a natural field experiment." *Management Science* 67 (1): 8–21. [41]

Locke, John (1975). *An Essay Concerning Human Understanding*. Peter H. Nidditch, ed. New Edition. Oxford: Clarendon Press. Originally published in 1690. [27]

Loewenstein, George (1987). "Anticipation and the Valuation of Delayed Consumption." *The Economic Journal* 97 (387): 666. [2]

Lung, Caroline W. Liqui (2022). "Optimal Self-Screening and the Persistence of Identity-Driven Choices." *Working Paper*, [43]

Matthey, Astrid, and Tobias Regner (2011). "Do I Really Want to Know? A Cognitive Dissonance-Based Explanation of Other-Regarding Behavior." *Games* 2 (1): 114–35. [16]

McGee, Dan (2024). "Stereotypes and Strategic Discrimination." *Working Paper*, [43]

Mechtenberg, Lydia, Grischa Perino, Nicolas Treich, Jean-Robert Tyran, and Stephanie Wang (2024). "Self-Signaling in Voting." *Journal of Public Economics* 231: 105070. [46]

Mijović-Prelec, Danica, and Dražen Prelec (2010). "Self-deception as self-signaling: A model and experimental evidence." *Philosophical Transactions of the Royal Society B: Biological Sciences* 365 (1538): 227–40. [3]

Möbius, Markus M., Muriel Niederle, Paul Niehaus, and Tanya S. Rosenblat (2022). "Managing Self-Confidence: Theory and Experimental Evidence." *Management Science* 68 (11): 7793–817. [11, 21–23, 45]

Momsen, Katharina, and Markus Ohndorf (2020). "When Do People Exploit Moral Wiggle Room? An Experimental Analysis of Information Avoidance in a Market Setup." *Ecological Economics* 169: 106479. [18]

Momsen, Katharina, and Markus Ohndorf (2023). "Information Avoidance: Self-image Concerns, Inattention, and Ideology." *Journal of Economic Behavior & Organization* 211 (7): 386–400. [17]

Moradi, Homayoon (2018). "Selfless Ignorance: Too Good to Be True." *Working Paper*, [17, 45]

Morin, Alain (2011). "Self-Awareness Part 1: Definition, Measures, Effects, Functions, and Antecedents: Self-Awareness." *Social and Personality Psychology Compass* 5 (10): 807–23. [2]

Müller, Maximilian W. (2022). "Selective Memory around Big Life Decisions." *Working Paper*, [30]

Müller, Maximilian W., Joan Hamory Hicks, Jennifer Johnson-Hanks, and Edward Miguel (2022). "The Illusion of Stable Fertility Preferences." *Population Studies*, [10]

Mummolo, Jonathan, and Erik Peterson (2019). "Demand Effects in Survey Experiments: An Empirical Assessment." *American Political Science Review* 113 (2): 517–29. [51]

Murphy, James J, P Geoffrey Allen, Thomas H Stevens, and Darryl Weatherhead (2005). "A Meta-Analysis of Hypothetical Bias in Stated Preference Valuation." *Environmental and Resource Economics* 30: 313–32. [39]

Nasim, Sanval, and Andreas Stegmann (2022). "Political Identity and Foreign Aid Efficacy: Evidence from Pakistani Schools." *Working Paper*, [46]

Oh, Suanna (2023). "Does Identity Affect Labor Supply?" *American Economic Review* 113 (8): 2055–83. [2]

Oprea, Ryan, and Sevgi Yuksel (2022). "Social Exchange of Motivated Beliefs." *Journal of the European Economic Association* 20 (2): 667–99. [44]

Ploner, Matteo, and Tobias Regner (2013). "Self-Image and Moral Balancing: An Experimental Analysis." *Journal of Economic Behavior & Organization* 93 (9): 374–83. [40]

Rodemeier, Matthias (2023). "Willingness to Pay for Carbon Mitigation: Field Evidence from the Market for Carbon Offsets." *Working Paper*, [39]

Ross, Lee, and Richard E. Nisbett (1991). *The Person and the Situation: Perspectives of Social Psychology*. McGraw-Hill Series in Social Psychology. New York: McGraw-Hill, 286. [7]

Roy-Chowdhury, Vivek (2022). "Self-Confidence and Motivated Memory Loss: Evidence From Schools." *Working Paper*, [31]

Saccardo, Silvia, and Marta Serra-Garcia (2023). "Enabling or Limiting Cognitive Flexibility? Evidence of Demand for Moral Commitment." *American Economic Review* 113 (2): 396–429. [26]

Santos-Pinto, Luís, and Joel Sobel (2005). "A Model of Positive Self-Image in Subjective Assessments." *American Economic Review* 95 (5): 1386–402. [4]

Saucet, Charlotte, and Marie Claire Villeval (2019). "Motivated memory in dictator games." *Games and economic Behavior* 117: 250–75. [11, 31]

Schneider, Florian H. (2022). "Signaling Ideology through Consumption." *Working Paper*, [47, 51]

Schwardmann, Peter, and Joël Van Der Weele (2019). "Deception and Self-Deception." *Nature Human Behaviour* 3 (10): 1055–61. [26]

Serra-Garcia, Marta, and Nora Szech (2022). "The (In)Elasticity of Moral Ignorance." *Management Science* 68 (7): 4815–34. [18]

Sharot, Tali, and Neil Garrett (2016). "Forming Beliefs: Why Valence Matters." *Trends in Cognitive Sciences* 20 (1): 25–33. [2, 23]

Shayo, Moses (2009). "A Model of Social Identity with an Application to Political Economy: Nation, Class, and Redistribution." *American Political Science Review* 103 (2): 147–74. [1, 2, 5]

Shayo, Moses (2020). "Social identity and economic policy." *Annual Review of Economics* 12 (1): 355–89. [1, 2]

Sherif, Muzafer (1935). "A Study of Some Social Factors in Perception." *Archives of Psychology*, (27): [1]

Sial, Afras Y., Justin R. Sydnor, and Dmitry Taubinsky (2024). "Biased Memory and Perceptions of Self-Control in Exercise." *Working Paper*, [31]

Smári, Jakob, Daníel Ólason, and Ragnar P. Ólafsson (2008). "Self-Consciousness and Similar Personality Constructs." In *The SAGE Handbook of Personality Theory and Assessment*. Gregory John Boyle, Gerald Matthews, and Donald H. Saklofske, ed. 1st ed. Los Angeles, CA: SAGE Publications, 486–505. [2]

Smith, Adam (1759). *The Theory of Moral Sentiments*. London: A. Millar. [4]

Soldà, Alice, Changxia Ke, Lionel Page, and William von Hippel (2022). "Strategically Delusional." *Experimental Economics* 25: 1009–36. [26]

Sprengholz, Philipp, Luca Henkel, Robert Böhm, and Cornelia Betsch (2023). "Historical Narratives about the COVID-19 Pandemic Are Motivationally Biased." *Nature* 623 (7987): 588–93. [31]

Stuber, Jennifer, and Mark Schlesinger (2006). "Sources of Stigma for Means-Tested Government Programs." *Social Science & Medicine* 63 (4): 933–45. [42]

Sweeney, Allison M., and Anne Moyer (2015). "Self-Affirmation and Responses to Health Messages: A Meta-Analysis on Intentions and Behavior." *Health Psychology* 34 (2): 149–59. [48]

Taber, Charles S., and Milton Lodge (2006). "Motivated skepticism in the evaluation of political beliefs." *American Journal of Political Science* 50 (3): 755–69. [3]

Tajfel, Henri (1978). "Interindividual Behaviour and Intergroup Behaviour." In *Differentiation Between Social Groups: Studies in the Social Psychology of Intergroup Relations*. Henri Tajfel, ed. London: Academic Press, 27–60. [1]

Tajfel, Henri, and John C. Turner (1979). "An Integrative Theory of Intergroup Conflict." In *The Social Psychology of Intergroup Relations*. Stephen Worchel and William G. Austin, ed. Brooks/Cole, 33–47. [1]

Teyssier, Sabrina, Fabrice Etilé, and Pierre Combris (2015). "Social- and Self-Image Concerns in Fair-Trade Consumption." *European Review of Agricultural Economics* 42 (4): 579–606. [48]

Tonke, Sebastian (2025). "Shaping Identity: Evidence from a Large-scale Field Experiment." *Journal of Political Economy Microeconomics* Forthcoming: [47]

Trachtman, Hannah, Andrew Steinkruger, Mackenzie Wood, Adam Wooster, James Andreoni, James J. Murphy, and Justin M. Rao (2015). "Fair Weather Avoidance: Unpacking the Costs and Benefits of "Avoiding the Ask"." *Journal of the Economic Science Association* 1 (1): 8–14. [20]

Van Der Weele, Joël J., and Ferdinand A. Von Siemens (2020). "Bracelets of Pride and Guilt? An Experimental Test of Self-Signaling." *Journal of Economic Behavior & Organization* 172 (4): 280–91. [32]

Verrina, Eugenio (2023). "Upset but (Almost) Correct: A Conceptual Replication of Di Tella, Perez-Truglia, Babino and Sigman (2015)." *Journal of the Economic Science Association* 9 (2): 327–36. [37]

Vu, Linh, Margarita Leib, Ivan Soraperra, Joël J. Van Der Weele, and Shaul Shalvi (2023). "Ignorance by Choice: A Meta-Analytic Review of the Underlying Motives of Willful Ignorance and Its Consequences." *Psychological Bulletin* 149 (9-10): 611–35. [6, 15]

Wang, Ao, Shaoda Wang, and Xiaoyang Ye (2023). "When Information Conflicts with Obligations: The Role of Motivated Cognition." *Economic Journal* 133 (654): 2533–52. [25]

Zimmermann, Florian (2020). "The Dynamics of Motivated Beliefs." *American Economic Review* 110 (2): 337–63. [10, 28, 45, 50]