# Virtue Discounting:
## Observability Reduces Moral Actors' Perceived Virtue

Gordon T. Kraft-Todd*[a], Max Kleiman-Weiner[b], Liane Young[a]
[a]Department of Psychology and Neuroscience, 140 Commonwealth Ave, Boston College, Chestnut Hill, MA 02467, USA
[b]School of Engineering and Applied Sciences, Harvard University, 33 Kirkland St., Cambridge, MA 02138, USA
*Corresponding author: gordon.kraft-todd@bc.edu; Cell: 978-621-6120; Fax: 617-552-0523

**Abstract**
Performing prosociality in public presents a paradox: only by doing so can people demonstrate their virtue and also influence others through their example, yet observers may derogate actors' behavior as mere "virtue signaling." We introduce the term *virtue discounting* to refer broadly to the reasons that people think actors' behavior is less attributable to their virtue. We investigate the *observability* of actors' behavior as one such reason, and explore observers' motivational inferences for their behavior using the comparison of *generosity* and *fairness* as a case study among virtues. Across 14 studies (7 preregistered, total $N$=9,360), we show that public actors are perceived as less morally good than private actors, and that this effect is stronger for generosity compared to fairness (i.e. *differential virtue discounting*). Exploratory factor analysis suggests that three types of motives—principled, reputation-signaling, and norm-signaling—affect virtue discounting. Using structural equation modeling, we show that observability's effect on actors' moral goodness ratings is largely explained by inferences that actors have less principled motivations. Further, we leverage experimental evidence to provide stronger causal evidence of these effects. We discuss theoretical and practical implications of our findings, as well as future directions for research on the social perception of virtue.

**Keywords**
virtue; observability; motivation; generosity; fairness

## Introduction

*A [person] of the highest virtue does not keep to virtue and that is why [they have] virtue.*
*A [person] of the lowest virtue never strays from virtue and that is why [they are] without virtue.*
*—Lao Tzu, Tao Te Ching (Chapter 38, L.1-4; Tzu & Lau, 1963)*

Public prosociality presents a paradox: For an actor's virtue to be known—and for their example to be followed—their morally good behavior must be observed by others; yet when an actor's morally good behavior is observable, others may doubt the actor's virtue. Resolving this paradox is particularly perplexing for early adopting advocates of uncommon prosocial behaviors (e.g. installing residential solar panels, buying electric vehicles, adopting vegan diets): can such an individual effectively demonstrate the moral worth of their behavior while avoiding the denigration of their moral character?

To illustrate this paradox, imagine browsing a printed list of people who donated the same amount to a charitable cause. The first two names listed are one of your friend's and "Anonymous"; whose behavior is more indicative of their virtue? If you think your friend appears less virtuous by comparison, your intuitions are aligned with ancient philosophical arguments (Maimonides, 1170) borne out in recent empirical work (De Freitas et al., 2019). Still, your (less virtuous) friend is likely to have more prosocial influence on you than the (more virtuous) anonymous giver (Smith et al., 2015). From a consequentialist standpoint, your friend actually caused *more* overall good; why, then, do we think they're *less* virtuous?

This paradox has long been puzzled over in the context of charitable giving—a prime example of behavior expressing the virtue of *generosity*—but are the dynamics the same when considering behaviors expressing other virtues (e.g. *fairness*)? This question is of theoretical interest for understanding social perception of prosociality, but it is also of practical interest for its downstream consequences: we can be more effective advocates if we can understand which types of behaviors face the paradox of public prosociality (and why).

### *"Virtue discounting" — theoretical background of a novel organizing concept*

We define a *virtue* as: a quality of individuals valued by their culture and expressed through a stable pattern of properly motivated behavior (Kraft-Todd et al., 2022). Our definition of virtue brings together various relevant theoretical approaches and empirical findings on virtue ethics, social perception, moral cognition, and evolutionary psychology. First, following most work in the Western philosophical and empirical traditions, many fundamental aspects of our concept of virtue can be traced to Aristotle's *Ethics* (Aristotle, 1999). Namely, that virtue is trait-like, i.e. a stable, dispositional characteristic of individuals; that it is expressed through behaviors; and that it requires "proper" motivation (Cokelet & Fowers, 2019). Proper motivation evokes Aristotle's concept of *eudaimonia* (often translated as "human flourishing") which he meant as an end in itself (rather than a means to some other end; Aristotle, 1999), interpreted in recent work as, "one's understanding of the virtue is itself motivating" (Cokelet & Fowers, 2019). Second, though some work in this tradition has attempted to identify a definitive taxonomy of virtues (Peterson & Seligman, 2004), we take the view that cultures will differ in the virtues they value (Snow, 2019) in accordance with their social norms (Boyd & Richerson, 1992). Third, the psychology of virtue perception is a subset of social perception in which we perceive others' traits from their behaviors (Tamir & Thornton, 2018), integrating motivational inferences (Carlson et al., 2022) to make "person-centered" characterological moral judgments

of others (Pizarro & Tannenbaum, 2012). Fourth, we believe that virtue perception functions according to the logic of costly signaling theory (Zahavi, 1975), in that we make virtue judgments using actors' observable phenotypes (i.e. their behaviors) to make inferences about actors' (difficult to directly observable) underlying traits (i.e. their virtue). Finally, we believe that a key aspect of our virtue perception is analogous to discounting in causal attribution (Morris & Larrick, 1995); in that we jointly evaluate the relative contribution of multiple causes for actors' behavior, discounting those which evidence (or lack thereof) leads us to believe they are less likely.

Synthesizing this work, we introduce the term *virtue discounting* to describe phenomena wherein observers are less likely to think that actors' behavior is attributable to their virtue. In accordance with our definition of virtue, this information may regard: actors' "proper" motivation, the consistency (or stability) of their behavior, and the cultural specificity of values prescribing certain virtues. In the present work, we focus on proper motivation (and consider others in the General Discussion). Specifically, we focus on the role of *observabilit*y, i.e. the degree to which actors' behavior is visible to uninvolved third parties (as opposed to recipients, interaction-partners, etc.), as a cause of virtue discounting.

Our *virtue discounting* terminology is useful in bringing together various lines of research showing this effect across domains such as social media, marketing, and charitable giving. The observability of virtue has become notoriously prominent in discussions of behavior on social media. The concept "virtue signaling" recently entered the cultural lexicon (Bartholomew, 2015; alternatively, sometimes referred to as "performative virtue") to pejoratively describe actors who invest minimal effort to widely broadcast their support for a cause (e.g. "outrage"; Crockett, 2017; Spring et al., 2018). Predating social media, the related (but importantly distinct; Tosi & Warmke, 2020) phenomenon "moral grandstanding"—i.e. "the use of moral talk for self-promotion"—has also gained increasing scholarly attention (Grubbs et al., 2019). Grandstanding involves people attempting to "impress others with their moral qualities…by saying something in public moral discourse", and so online platforms facilitate grandstanding by making it easy for people to participate in such discourse. Further, given rampant idealized self-presentation on visual social media such as Instagram (Harris & Bardey, 2019), users have ample opportunity to post images cultivating an image of themselves as possessing many virtues, such as cleanliness, serenity, faith, perseverance, wisdom, etc. In the business literature, research on "greenwashing" (e.g. Laufer, 2003) investigates skeptical interpretations of firms' eco-friendly products and investments. Although such activity is sometimes honestly intended to demonstrate firms' commitment to environmentalism, it is often derogated as a mere exercise in marketing and public relations. Similarly, in charitable giving, "conspicuous compassion" (West, 2004) is a term applied to outward signals of empathy or solidarity (most notably, ribbons demonstrating support for e.g. breast cancer research or veterans) that may more reliably demonstrate signalers' desire for recognition than their concern for the cause. Why, across this work, are observers less likely infer actors' behavior as a signal of their virtue?

Observability is a central component of the theory of indirect reciprocity (Nowak & Sigmund, 1998), which explains how cooperation can evolve among unrelated individuals when there are rules (e.g. social norms) governing individuals' behavior (Ohtsuki & Iwasa, 2006) and information regarding individuals' (non-)adherence to these rules can be stored and transmitted (i.e. through their "reputations"). Implications of this theory are borne out in field experiments (for a review, see: Kraft-Todd et al., 2015) showing that making people's behavior observable activates people's reputational concerns, causing them to follow social norms supporting

prosocial behavior. Despite the efficacy of observability for motivating such behavior (e.g. Yoeli et al., 2013), observers may be less likely to infer that the behavior it inspires is a signal of actors' virtue because actors are not properly motivated. Thus, observers may infer (correctly) that actors are motivated by reputational benefit, rather than (referencing domains discussed above), e.g. genuine outrage at injustice, desire for environmental conservation, or concern for others' welfare. In other words, observability introduces an ulterior motive (reputational concerns) that taint observers' inferences of actors' virtue because actors' behavior is not (properly) motivated by pursuit of the virtue as an end in itself. Accordingly, we formulate our first preregistered hypothesis "virtue discounting": *people will discount actors' virtue when actors' behavior is observable.*

*Generosity vs. fairness as a case study in virtue discounting*

Much of the empirical work we have reviewed so far demonstrates virtue discounting in the context of the virtue *generosity*. We have previously argued that *fairness* is well-suited as a comparison virtue to generosity, and we provided novel evidence that generosity and fairness differ across many psychological dimensions, as well as in the natural language people use to describe example behaviors of each virtue (Kraft-Todd et al., 2022). We build on much prior work distinguishing these virtues: in virtue ethics (e.g. as "natural" vs. "artificial" virtues; Hume, 1902); in recent theoretical arguments distinguishing these virtues' functions (i.e. generosity enables individuals to make cooperative partnerships, whereas fairness enables individuals to avoid punishment from their cooperative partners; Shaw, 2016); and in empirical work demonstrating how these virtues can be operationally distinguished and shown to come into conflict (e.g. Kleiman-Weiner, Shaw, et al., 2017; Shaw & Olson, 2012). Taken together, fairness may serve as a valuable comparison to generosity for investigating virtue discounting across virtues.

As in our previous work (Kraft-Todd et al., 2022), we employ a "narrow" definition of both virtues. We define *generosity* as "giving an abundance of one's money or time", capturing most uses of the term, but not, e.g. generosity of attitudes (Gulliford & Roberts, 2018). We define *fairness* as "treating others equally and fairly, without bias", capturing recent work on impartiality (Shaw, 2013), but not, e.g. work on fairness that focuses on the need of recipients (i.e. charity; Niemi & Young, 2017). Thus, our case study might be more precisely described as: "a case study of generosity vs. fairness (as impartiality)." To avoid confusion: we will henceforth use the terms *generosity* and *impartiality* when discussing our experiments—because these are the terms we use in our stimuli—but to accurately refer to the respective virtues, we will use the terms *generosity* and *fairness*.

Pertinent to the present investigation, there is also reason to believe that generosity and fairness might be discounted to different degrees. Fairness (as a component of *justice*) has long been considered enforceable through law whereas generosity has not (Schneewind, 1990). Consequently, there may be greater plausible deniability in attributing actors' fairness (vs. generosity) to a desire to conform to formal regulations (and social expectations) rather than a desire to improve one's reputation. Further complicating the prosocial interpretation of generous behaviors, although generosity can be a signal of cooperative intent (i.e. willingness to provide benefits to others), it can also be a signal of wealth (Barclay, 2016), and thus individuals may show generosity so that others think they are rich. Also, the ideal of fairness (as impartiality) has a specific numerical connotation—i.e. to treat others with zero bias—whereas the ideal of generosity is effectively without ceiling (i.e. to be maximally generous, one can always give

more). As a result, it is easier to coordinate around a (categorical) norm of fairness than a (continuous) norm of generosity. There is a wealth of evidence from mathematical models and empirical studies (Yoeli & Hoffman, 2022) suggesting that plausible deniability and categorical norms are two important mechanisms motivating human behavior and how it is perceived. In the present context, interpreting fairness compared to generous behaviors seems simpler, both because of the categorical nature of applicable norms, and also because of greater plausible deniability in the attribution of selfish motivations. Thus, we arrive at our second preregistered hypothesis, "differential virtue discounting": *people will discount fairness (as impartiality) less than generosity*.

*Motivational inference in perceptions of virtue*

To this point, we have discussed "motivation" in a limited sense; i.e. as either the proper motivation requisite for virtue or as the motivation to benefit one's reputation. Yet, prior work has explored many distinct motivations for prosocial behavior (Carlson et al., 2022; Kodipady et al., 2021; Narvaez & Snow, 2019; Reiss & Havercamp, 1998). To better understand the motives observers attribute to prosocial actors and how these contribute to virtue discounting, we focus on six motivations suggested by this literature: *self-presentation*, *norm-signaling*, *self-benefit*, *other-benefit*, *moral rules*, and *virtue identification*.

We opened with a paradox of public prosociality that highlights the tension between two motivations. Pulling in one direction (and to put a finer point on "reputational concerns"), we call the motivation to affect others' impression of oneself *self-presentation*, following work documenting how people manage others' impression of them through "self-presentation strategies" (Jones & Pittman, 1982). Pulling in the other direction, we call the motivation to lead by example *norm-signaling*, deriving from work on the role individuals have to influence social norms in a "grassroots" manner (Tankard & Paluck, 2016). Although the idea of "leading by example" is a topic of increasing theoretical (Henrich, 2009) and empirical (Kraft-Todd et al., 2018) interest, we are unaware of previous work investigating "leading by example" as a *motivation* (though see: Kodipady et al., 2021).

We briefly alluded to another pair of conflicting motivations in our discussion of observable generosity. A number of recent reviews have emphasized how fundamental the motivation of *other-benefit*, i.e. the desire to improve the welfare of others, is for motivating prosocial behavior (e.g. Keltner et al., 2014). The emphasis of these reviews stands in contrast to a historical bias, informed by the economic and evolutionary literatures (e.g. Friedman, 1953), to a reductive focus on the motivation of *self-benefit*, i.e. the desire to improve one's own welfare.

Finally, we explore two identity-relevant motivations. Research on social norms has distinguished types of social rules guiding behavior, among which is the class of *moral rules* that individuals believe should guide behavior regardless of others' expectations (i.e. as opposed to "socially-dependent" social rules that rely on others' expectations, moral rules are "socially-independent"; Bicchieri, 2006; Levine et al., 2020). This motivation interestingly differs from the previously discussed motivations because it is not outcome-oriented; instead, it derives from an individual's socially-independent moral values. Another socially-independent motivation for virtuous behavior derives from actors' sense of identity, as explored in research on, e.g. "moral identity" (Aquino & Reed, 2002), "moral consistency" (Kleiman-Weiner, Saxe, et al., 2017; Mullen & Monin, 2016), and "self-concept maintenance" (Mazar et al., 2008). In the context of the present work, we study whether individuals' identification with the specific virtue under

investigation might be a motivation for virtuous behavior, and call this *virtue identification*.[1] This motivation most closely approximates "proper" motivation in our definition of virtue.

In keeping with previous work, our third preregistered hypothesis is: *(differential) virtue discounting will be explained by observers' inferences that public actors have more selfish motivations than private actors*. In the present work, however, we extend this previous work which typically invokes "selfish" motivations (generally) and reputational concerns (specifically) as drivers of virtue discounting (e.g. Raihani & Power, 2021). We do not intend our investigation to be an exhaustive account of all possible motivational inferences, but our preregistered exploratory factor analyses and structural equation models shed new light on the social perceptions underlying virtue discounting.

**General Methods**

We describe the general procedure in common across experiments here, and summarize more fine-grained detail in Table 1. In the Supplement, we provide justifications for changes in experiment design across experiments (Section 2), and also provide complete experimental instructions (Section 9). All measures, manipulations, and exclusions in all experiments are disclosed across the main text and Supplement. All experiments were conducted online using Qualtrics survey software. A convenience sample of participants were recruited using the crowdsourcing tool Cloud Research and Amazon Mechanical Turk ("mTurk"; Arechar et al., 2017; Berinsky et al., 2012). We excluded duplicate Amazon worker IDs and IP addresses to prevent analyzing multiple observations per participant (as well as participants who dropped out prior to assignment to condition), yielding a final sample of $N$=9,360 participants (51.2% female, *average age*=37.7 years). Informed consent was obtained from all participants, who completed experiments in *mean*=5 minutes and were paid *mean*=$0.89 for their participation. For conciseness, we abbreviate references to specific experiments as "E#" (e.g. E3=Experiment 3).

After providing consent and entering their mTurk ID, we randomly assigned participants to one between-subjects condition. We crossed our primary manipulations in all experiments in a 2 (virtue: generosity vs. impartiality) x 2 (observability: public vs. private) factorial design. First, participants read text delivering our virtue manipulation, adapted from Merriam-Webster.com (see Table 1). We asked participants to imagine that they know someone (henceforth: "the actor") whom we named using a list of common female names in the US (because we were not interested in the effect of actor gender on the dependent variables, we used all female names). We then told participants that the actor engaged in a set of three behaviors demonstrating [generosity/impartiality]. Then, participants read text delivering our observability manipulation (see Table 1, Columns 7 and 8).

[1] We note that although we treat *moral rule* and *virtue identification* motivations as "socially-independent", this may only apply to the actors' proximate psychology, as both motivations are likely shaped through cultural norm internalization (Henrich, 2016).

**Table 1. Key elements of stimuli.** *Shown are characteristics and language of key elements of stimuli bearing on hypotheses in all experiments (for complete experimental instructions, see Supplement Section 9).*

| | | | | Stimuli language | | | |
|---|---|---|---|---|---|---|---|
| | | Stimuli information | | Virtue (example behaviors) | | Observability | |
| Experiment | N | Motive stipulated? | Example behaviors | Generosity | Impartiality | Private | Public |
| 1 | 389 | Yes | (none) | [Virtue definition]: "Generosity usually means giving an abundance of one's money or time" | [Virtue definition]: Impartiality usually means treating everyone equally and fairly, without bias | Though she is [G/I] when she is with others, she is especially [G/I] when no one is watching since she knows that acting in this way is consistent with her values. | She is especially [G/I] when others are watching her act since she knows that her reputation for being [G/I] will improve. |
| 2 | 394 | | | | | | |
| 3 | 394 | | | | | | |
| 4 | 388 | | Experimenter-generated | - volunteered at a homeless shelter<br>- donated money to charities like Doctors without Borders<br>- donated blood during a blood drive | - made sure everyone at a social gathering receives the same amount of food<br>- divided work evenly among all participants in a group project<br>- made auditions or job applications blind so that subtle, unconscious biases against particular genders or ethnicities don't enter into the decision-making process | | |
| 5 | 393 | | | | | Though she is [G/I] when she is with others, she is even [G/I] when no one is watching. | She is especially [G/I] when others are watching her act. |
| 6 | 393 | | Participant-generated | - bought a friend an expensive gift<br>- gave a waiter a large tip<br>- stayed late to help a coworker | - gave her children equal allowances<br>- conducted a blind audition<br>- drew names from a hat for a project at work | She did these things in private; therefore, other people did not know that she did them. | She did these things in public; therefore, other people knew that she did them. |
| 7 | 582 | No | | | | | |
| 8 | 577 | | | | | | |
| 9 | 386 | | | | | | |
| 10 | 663 | Yes* | | | | | |
| 11 | 394 | No | | - gave someone a hand carrying groceries<br>- shared food with friends<br>- gave someone praise | - stayed out of an argument<br>- divided food by cutting and letting the other person pick which piece they want<br>- helped to moderate when your friends had a disagreement | | |
| 12 | 377 | | | | | | |
| 13 | 1770 | Yes* | | - bought someone a meal<br>- donated blood<br>- volunteered at an animal shelter | - divided food by cutting it and letting the other person pick which piece they wanted<br>- learned to pronounce others' names regardless of their country of origin<br>- listened to both parties in a conflict equally. | | |
| 14 | 2260 | | | | | | |

We generated the behaviors we used as stimuli in a rigorous "bottom-up" manner (Kraft-Todd et al., 2022). First, we randomly assigned participants to virtue condition (generosity vs. impartiality) and asked them to provide behaviors that demonstrated the virtue using free-response text. Second, we recruited an independent sample (also randomly assigned to virtue condition) to rate a subset of these participant-generated behaviors on nine underlying dimensions impacting moral judgment. Finally, to avoid idiosyncrasies of any specific behavior, we selected a set of three behaviors that we described hypothetical actors engaging in to create a general impression of the actors as demonstrating each virtue in participants' minds. We note that using data from our previous work (Kraft-Todd et al., 2022), we find no significant differences in participants' ratings between the sets of behaviors we used as stimuli in the generosity and impartiality conditions of each experiment across three dimensions: moral goodness, descriptive normativity, and the extent to which behavior is indicative of the actor's consistency across situations. This evidence speaks against alternative explanations of our effects, e.g. based on the diagnosticity of the behavior (Mende-Siedlecki et al., 2013) for actors' virtue.

Following the stimuli, we presented participants with the two primary dependent measures and six secondary dependent measures (see Table 2) in randomized order. All dependent measures were answered on 100-point unmarked slider scales with extreme anchors labeled (and midpoints labeled for primary dependent measures). At the end of each experiment, we presented participants with basic demographic questions in randomized order (see Supplement Table 2 for summary of participant demographics by experiment).

**Table 2. Dependent measures.** *Shown are labels (Column 2) and item wording (Column 3) for dependent measures used in all experiments.*

| | *Item* | *Wording* |
|---|---|---|
| Primary dependent measures | Moral goodness | "How morally good is [the actor]?" |
| | Trait ratings | "How [generous/impartial] is [the actor]?" |
| Secondary dependent measures (motivational inferences) | Moral rule | "...because she thinks it is the right thing to do?"* |
| | Virtue identification | "...because she wants to be [generous/impartial]?"* |
| | Other-benefit | "...because she wants to benefit others?"* |
| | Self-presentation | "...because she is trying to make others think she is [generous/impartial]?"* |
| | Self-benefit | "...because she thinks she will personally benefit from acting this way?"* |
| | Norm-signalling | "...because she wants others to be [generous/impartial], and she is trying to lead by example?"* |

*Preceded by: "How much do you think [name] is motivated to act [generously/impartially]..."

We conducted analyses using STATA (16.1) and R software (4.1.2). We obtained effect sizes (Cohen's D) through use of an online calculator (Lenhard & Lenhard, 2016). For regression analyses, we compute pairwise comparisons of estimated marginal cell means corrected for multiple comparisons using Scheffe's adjustment (Winer et al., 1991; though results are equivalent using Bonferroni correction). Structural equation models were constructed using standardized variables (Hayes, 2013), and indirect effects are calculated using the multivariate delta method (Sobel, 1982) with bootstrapped standard errors (UCLA: Statistical Consulting Group, 2021). Prior to conducting E8, we conducted a power analysis using the Superpower package in R (Lakens & Caldwell, 2021), using data from E7. With a desired effect size of $d$=.30 (for the virtue*observability interaction; see Supplement Section 3 for effects by experiment), our sample size of $N$=100 per cell was powered at 83.25% with an alpha level of .05. For E1-7, we used the heuristic of recruiting $N$=100 per cell. Sample size for each experiment was determined before any data analysis. Sensitivity analyses were conducted using G*Power (3.1.9.7) software (Faul et al., 2007).

Below we present these preregistered analyses using data aggregated across all of our experiments, though we also provide precise preregistered analyses in the Supplement (see Supplementary Tables 4-7). Although our preregistered hypotheses include both of our primary dependent measures (i.e. moral goodness ratings and trait ratings), these measures were highly correlated (across all experiments: $r$=.74, $p$<.001), so we present moral goodness rating results here for conciseness. Results do not qualitatively differ if trait ratings are used instead (see Supplementary Tables 5 and 7). E10, E11, E13, and E14 included preregistered exclusions for participants who failed attention checks (average failure rate across studies=12.2%). To maintain consistency in our results across experiments and to be maximally inclusive of data we collected, the results presented below include participants who failed attention checks ($N$=720, 7.7% of total $N$). Results for these experiments are robust to excluding these participants (Supplementary Table 8).

Preregistrations were conducted for: E8 (https://aspredicted.org/FLG_VRD), E9 (https://aspredicted.org/NUP_ESB), E10 (https://aspredicted.org/FCO_HXM), E11 (https://aspredicted.org/2Y2_YKT), E12 (https://aspredicted.org/AES_HGO), E13 (https://aspredicted.org/G6S_S35), and E14 (https://aspredicted.org/SHW_9PN). All data and

code are publicly available at: (https://osf.io/sud3m/?view_only=380a169770b9474f93d2b5b73adc7410).

**Analysis 1. Virtue discounting affects generosity more than impartiality**

The purpose of Analysis 1 is to test for evidence of our first two hypotheses (preregistered in E8-14): 1) virtue discounting: people will discount actors' virtue when actors' behavior is observable; and 2) differential virtue discounting: people will discount fairness (as impartiality) less than generosity. To do so, we examine participants' ratings of actors' moral goodness in our hypothetical vignettes across conditions, using multivariate regression (collapsing across experiments) as well as meta-analysis (using the "metan" package in R; Olivoto & L'ucio, 2020).

We also provide an exploratory test of the directionality implied by our *virtue discounting* terminology (as discussed in the introduction), i.e. that participants discount public virtue rather than reward private virtue. In two experiments (E7 and E8, see Supplement Section 5) we added an additional "baseline" observability condition that provided no information about observability (i.e. we simply omitted the line, "She did these things in [public/private]; therefore, other people [knew/did not know] that she did them" from our experimental instructions; see Supplement Section 9). Crucially, we note that the example behaviors used as stimuli in these experiments were not rated differently on the dimension of "potential for anonymity" in our previous data (Kraft-Todd et al., 2022), supporting our interpretation that participants did not perceive differences across virtues in the extent to which these behaviors were known to uninvolved third-parties. By comparing participants' moral goodness ratings in the public and private observability conditions to the baseline observability condition, we provide a test of whether public virtue is discounted, or instead, whether private virtue is rewarded.

*Methods*

We use data from all conditions in all experiments in which we manipulated observability ($N$=8,969; 51.2% female, average age=37.8 years). First, we conduct a multivariate regression analysis predicting moral goodness and trait ratings by our virtue manipulation (generosity vs. impartiality), observability manipulation (public vs. private), and their interaction, with Experiment as a covariate. Second, we conduct a random-effects meta-analysis on the effect sizes from the virtue*observability interaction.[2]

*Results*

As predicted in our first preregistered hypothesis (*virtue discounting*), we find a significant effect of our observability manipulation, collapsed across virtue condition, on moral goodness ratings ($F$(1,8954)=593.51, $p$<.001, $d$=.51) such that public actors ($m$=69.80, 95% CI [69.21, 70.39]) are perceived as significantly less morally good than private actors ($m$=79.73, 95% CI [79.18, 80.28]). Sensitivity analysis revealed that (at $p$=.05 and power=80%) the minimum detectable effect size for this test is $d$=.06.

[2] Note that we only tested generosity in E10 so this experiment is not included in this analysis.

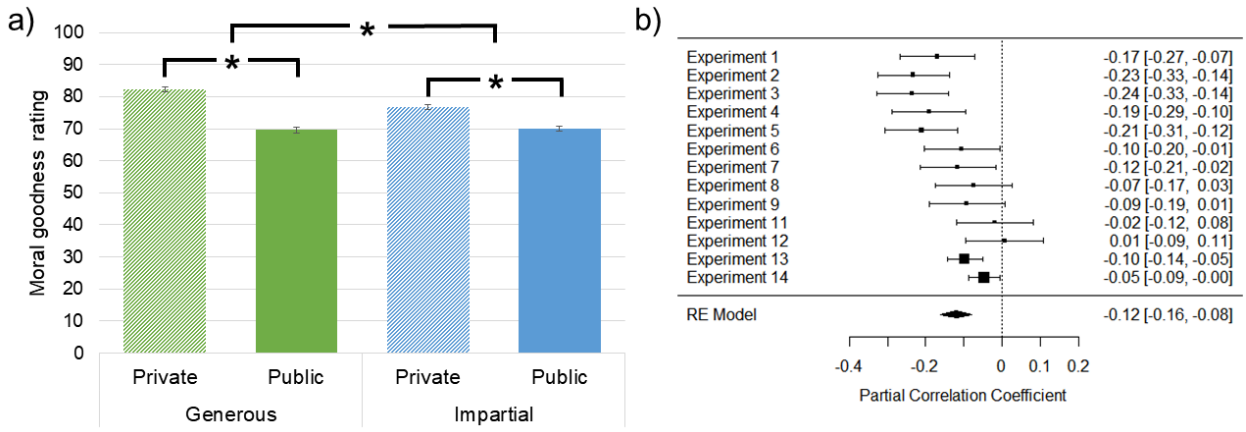


**Fig 1. Virtue discounting (i.e. public actors are rated less morally good than private actors) affects generosity more than impartiality.** Shown are **a)** means (with 95% CIs) of moral goodness ratings (0-100 unmarked slider) collapsed across all 14 experiments, as a function of whether the actor is said to have engaged in generous (green) or impartial (blue) behaviors and whether the actor is said to have engaged in behaviors publicly (solid) or privately (lines). Significant contrasts denoted with (*). From left to right: private generosity *N*=2,396, public generosity *N*=2,428, private impartiality *N*=2,079, public impartiality *N*=2,069. **b)** Random effects meta-analysis of the virtue*observability interaction on moral goodness ratings across 13 experiments (E10 only investigated generosity). Effect sizes are shown as Cohen's d; error bars indicate 95% CIs. The relative sizes of the boxes indicate the weighting assigned to the experiments by the meta-analysis.

As predicted in our second preregistered hypothesis (*differential virtue discounting*), we also find a significant interaction between virtue and observability on moral goodness ratings ($F(1,8952)=50.31$, $p<.001$, $d=.15$; see Figure 1). Sensitivity analysis revealed that (at $p=.05$ and power=80%) the minimum detectable effect size for this test is $d=.06$. Publicly generous actors are perceived as significantly less morally good ($m=69.60$, 95% CI [68.75, 70.44]) than privately generous actors ($m=82.24$, 95% CI [81.50, 82.98], Scheffe's $t=-22.80$, $p<.001$, $d=.64$). Publicly impartial actors ($m=70.04$, 95% CI [69.22, 70.86]) are also perceived as significantly less morally good than privately impartial actors ($m=76.83$, 95% CI [76.03, 77.63], Scheffe's $t=-11.46$, $p<.001$, $d=.36$). Providing the crucial test of our preregistered differential virtue discounting hypothesis, a Wald test ($\chi 2(1)=51.49$, $p<.001$) reveals that the effect of observability is greater for generosity (*contrast*=-12.63, 95% CI [-14.18, -11.09]) than impartiality (*contrast*=-6.85, 95% CI [-8.52, -5.18]).

We also provide multiple demonstrations of the robustness of our effects and our interpretation. A random-effects meta-analysis on the 13 virtue*observability interaction effect sizes reveals that generosity is discounted to a greater extent than impartiality ($d=-1.01$, 95% CI [-1.36, -.66], $Z=5.64$, $p<.001$; see Figure 1b). We observe that there is evidence of heterogeneity in effect size across experiments ($\chi 2(10)=153.58$, $p<.001$), and hypothesize that this is due to differences in the designs of our experiments (see General Methods). Note that across all experiments, the virtue*observability interaction is either significant, such that generosity is discounted more than impartiality, or it is not significant; it is never the case that impartiality is discounted more than generosity. We further demonstrate the robustness of our effects by replicating these results in nearly all meaningful subsets of our experiments; importantly, including among our preregistered experiments (see Supplement Section 4).

Finally, we provide evidence in support of the directionality implied by our *virtue discounting* terminology. Comparing participants' moral goodness ratings in the public and

private observability conditions to the baseline observability condition, we find that, compared to generous actors described without observability information (i.e. the baseline condition), privately generous actors are perceived as equivalently morally good (*contrast*=-3.09, 95% CI [-8.38, 2.20], Scheffe's *t*=-1.95, *p*=.579), whereas publicly generous actors are perceived as significantly less morally good (*contrast*=-5.54, 95% CI [-10.83, -.25], Scheffe's *t*=-3.49, *p*=.033). We interpret this result as showing that, consistent with our terminology, observers discount public virtue. In these experiments, impartial actors' virtue was not discounted (comparing public to private observability conditions: *contrast*=-2.47, 95% CI [-7.80, 2.85], Scheffe's *t*=-1.55, *p*=.792), so this comparison cannot be made for impartiality.

*Discussion*

Here we provide robust evidence in support of both of our first two preregistered hypotheses. First, our results demonstrate *virtue discounting*; participants rated public actors as less morally good than private actors. Second, our results demonstrate *differential virtue discounting*; participants discounted public (compared to private) generosity to a greater extent than they discounted public (compared to private)impartiality. These results are consistent with much previous work showing virtue discounting of generosity (e.g. Lin-Healy & Small, 2012). Further, it is the first demonstration, to our knowledge, of virtue discounting of impartiality. Finally, we provide evidence in support of the directionality implied by our *virtue discounting* terminology. Namely, participants rate publicly generous actors as less morally good than privately generous actors and also generous actors described without observability information (i.e. the baseline condition), but they do not rate the latter two types of actors differently. This result implies that *public virtue is discounted* (although without this evidence it might have been argued that *private virtue is rewarded*). Next, we turn to our analyses showing motivational inferences as a mechanism of observability in virtue discounting.

**Analysis 2. Virtue discounting due to observability can be explained by motivational inferences**

The purpose of Analysis 2 is to explore the role of motivational inferences as a mechanism of observability in virtue discounting, testing our third preregistered hypothesis (in E8, E9, and E11-13) that *virtue discounting will be explained by observers' inferences that public actors have more selfish motivations than private actors.*. It would follow from our previous results to explore how motivational inferences mediate the virtue*observability interaction, i.e. accounting for the *differential virtue discounting* effect (this analysis presented in Supplement Section 6). An important result of that analysis, however, is that there are not qualitative differences in the motivational inferences driving virtue discounting of generosity compared to impartiality (i.e. it is not the case that different types of motivations explain virtue discounting for each virtue, only the degree to which observers infer these motivations). We acknowledge that this may be due to the fact that our exploration of motivational inferences was not exhaustive (and comment on this issue in greater detail in the General Discussion). That is, a similar investigation employing a broader range of motivational inferences might find that there *are* distinct motivational inferences driving virtue discounting in generosity compared to impartiality. Though we admit this possibility, here we present an analysis of how motivational inferences explain *virtue discounting*, i.e. collapsing across generosity and impartiality.

Additionally, we believe that our investigation of the mechanism of observability in virtue discounting could be of greater theoretical interest (i.e. than such an analysis of differential

virtue discounting) because it could suggest a generalizable mechanism explaining virtue discounting via motivational inferences across an even broader range of virtues. Since we only use two virtues as stimuli, further research will be needed before such conclusions could be drawn. Still, we believe our analysis is an important first step towards such work.

*Methods*

We use data from all experiments in which we measured motivational inferences (E6-9 and E11-13; *N*=4,087; 49.9% female, average age=37.9 years). We present correlations among these items in Supplement Section 7. Here we present an exploratory factor analysis (EFA; preregistered in E8, E9, and E11-13) of motivational inference items to better understand their latent structure, and then fit a multiple mediation model using these factor scores. We constructed generalized structural equation models to compute the mediation results (see General Methods for more details), and present alternative model specifications in Supplement Section 8.

*Results*

We begin with an EFA (preregistered in E8, E9, and E11-13) of our six motivational inference items using iterated principal factors and oblique rotation. The analysis yielded two factors explaining 95.5% of the variance. Factor 1 explained 61.0% of the variance, and items with high loadings (>.6) were: *moral rule*, *virtue identification*, and *other-benefit*. Following our previous work (Kraft-Todd et al., 2022), we labeled Factor 1 "principled" because these motivations pertain either to actors' moral beliefs/identity (i.e. moral rule and virtue identification) or prosociality (other-benefit). Factor 2 explained 34.4% of the variance, and we labeled it "reputation-signaling" due to high loadings (>.6) by the items: *reputational benefit* and *self-benefit*.

We note, however, that our norm-signaling item loads almost equivalently on these factors (*principled*=.39; *reputation-signaling*=.45). Because norm-signaling did not load uniquely on either factor, and also because we are specifically interested in this novel construct, we therefore conduct a second EFA omitting this item, with the intention to use the resulting factor scores in addition to norm-signaling ratings as mediators in our subsequent analysis. This second EFA (also using iterated principal factors and oblique rotation) retains the same two factors, on which all items load as described above. We observe that they also explain similar proportions of the variance among motivational inference items (*principled*=69.2%; *reputation-signaling*=28.0%), and that these factors are moderately and negatively correlated (*r*=-.46).

Next, we examine the mediation of observability on moral goodness ratings by the two motivational inference factor scores plus the norm-signaling item (collapsed across virtue, controlling for Experiment and the covariance among mediators). Participants infer that, compared to private actors, public actors have significantly lower *principled* motivation (*b*=-.56, 95% CI [-.62, -.50], *p*<.001; see Figure 2), and significantly higher *reputation-signaling* motivation (*b*=.67, 95% CI [.62, .73], *p*<.001) as well as *norm-signaling* motivations (*b*=.28, 95% CI [.22, .34], *p*<.001). Next, we find that participants' moral goodness ratings are significantly associated with their motivational inferences, such that higher *principled* (*b*=.53, 95% CI [.50, .55], *p*<.001) and *norm-signaling* (*b*=.11, 95% CI [.08, .13], *p*<.001) inferences are associated with higher moral goodness ratings, while higher *reputation-signaling* inferences are associated with lower moral goodness ratings (*b*=-.06, 95% CI [-.08, -.03], *p*<.001).

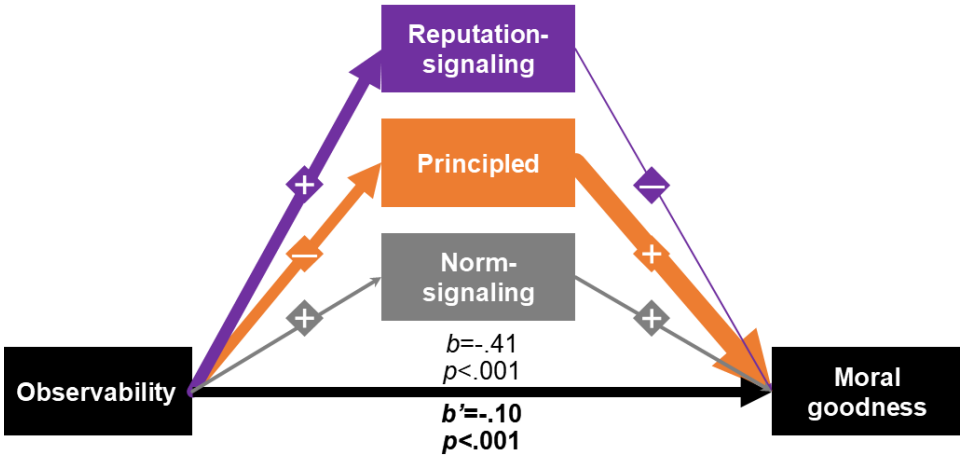


**Fig 2. The effect of observability on virtue discounting is explained by inferences that actors have lower principled motivations.** Shown is a generalized structural equation model (showing multiple mediation) of the effect of observability on moral goodness ratings by motivational inference factor scores and norm-signaling item collapsed across virtue (*N*=4,087). Bottom arrow (black) represents the effect of the observability manipulation (public compared to private condition) on moral goodness ratings with (b) and without (b') mediators as covariates. From left-to-right, the first set of arrows represents the effect of the observability manipulation on mediators, and the second set of arrows represents the correlation of mediators with moral goodness ratings. Line thickness represents correlation strength; "+" and "-" represent correlation direction; all variables standardized for this analysis.

Finally, we turn to the mediation results. Restating the *virtue discounting* result we present in Analysis 1, the total effect of observability on moral goodness ratings is significant, such that participants rate public actors as less morally good than private actors (*b*=-.41, 95% CI [-.46, -.35], *p*<.001). Although the direct effect of observability on moral goodness (i.e. accounting for indirect effects through motivational inferences) is greatly reduced in magnitude, it remains significant (*b'*=-.10, 95% CI [-.14, -.06], *p*<.001), implying partial mediation (72.6% of the total effect). Calculating indirect effects as percent of total effect, *principled* motivation accounts for 72.3% of this mediation, *reputation-signaling* motivation accounts for 9.3%, and *norm-signaling* motivation accounts for -7.3%.

*Discussion*

Here we provide evidence consistent with an explanation of observability's effect on virtue discounting through motivational inferences. In short, participants' motivational inferences (i.e. *reputation-signaling* factor score, *principled* motivation factor score, and *norm-signaling* item) mediate 72.6% of the effect of our observability manipulation (i.e. public vs. private conditions) on moral goodness ratings (collapsed across virtue condition). Although our *reputation-signaling* and *principled* motivational inference factors intuitively seem like conceptual opposites, we note that they are interestingly only moderately negatively correlated (*r*=-.46). Further, our finding that the *norm-signaling* motivational inference item does not uniquely load on either of our two other motivational inference factors suggests that this construct may have a unique role in contributing to an explanation of virtue discounting (Kodipady et al., 2021).

The results of this analysis only weakly support our third preregistered hypothesis (i.e. that virtue discounting will be explained by observers' inferences that public actors have more selfish motivations than private actors). Despite the finding that an increase in ("selfish")

*reputation-signaling* motivational inferences account for 9.3% of the virtue discounting effect, 72.3% of the effect was explained through a decrease in *principled* motivational inferences. Our mediation result is robust to alternative model specifications, including one in which we model the mediating effect of the *reputation-signaling* factor score through the *principled* motivation factor score (see Supplement Section 8). That is, it might have been the case that participants perceived publicly virtuous actors to have greater *reputation-signaling* motivation, and therefore participants perceived them to have lower *principled* motivation, but our alternative model specification does not support such an account. This finding stands somewhat at odds with previous work showing that selfish motivations account for virtue discounting (e.g. Newman & Cain, 2014), although it should be noted that previous work did not simultaneously test multiple motivations that might explain this effect.

**Analysis 3. Observability does not cause virtue discounting when observers know actors' motivations**

The purpose of Analysis 3 is to provide stronger causal evidence that motivational inferences explain why observability causes virtue discounting. Our results from Analysis 2 suggest that observability engenders ambiguity in observers' motivational inferences; i.e. observers can attribute different—and even conflicting—motivations to publicly virtuous actors. To reiterate, participants inferred that, compared to private actors, public actors had lower *principled* motivation and higher *reputation-signaling* motivation (both associated with participants rating actors as less morally good). Yet, participants also inferred that public actors had higher *norm-signaling* motivation, which is associated with participants rating actors as *more* morally good.

Rather than rely on correlational mediation (as in Analysis 2), here we leverage a design feature of many of our experiments (E1-6, and some conditions in E10 and E14) in which we explicitly manipulated actors' motivation alongside observability. By comparing the effect of our observability manipulation among conditions in which we stipulate actors' motivation to those in which we *do not* (as in E7-9, E11-13, and other conditions in E10 and E14), we can test the "motivational ambiguity hypothesis" (preregistered in E10 and E14): *the main effect of observability (i.e. public vs. private) on ratings of moral goodness will be substantially reduced when we stipulate actors' motivation compared to when we do not.* To put this hypothesis plainly: observers may discount virtue because they are uncertain about actors' motivations (as in our "no motive" conditions, and presumably in real life). Explicitly providing people with information about actors' motivations, therefore, should drastically reduce the effect of observability on perceptions of actors' moral goodness.

*Methods*

We use data from all conditions in all experiments in which we manipulated observability (same as Analysis 1; $N$=8,969; 51.2% female, average age=37.8 years), subgrouping by manipulations of actors' motivation in our hypothetical vignettes (*principled*, $N$=2,151 vs. *reputation-signaling*, $N$=2,150 vs. no motivation stipulated, $N$=4,671). Although details varied by experiment (see Table 1), our manipulations of actor motive generally stipulated that actors either simultaneously had *principled* but not *reputation-signaling* motivations or *reputation-signaling* but not *principled* motivations. For conditions in which we did not stipulate actor motivation, we simply omitted language regarding actors' motivation.

As in Analysis 2, here we are primarily interested to explore the observability*motive interaction collapsed across virtues. Still, because it is possible that this effect would differ by virtue, we begin by conducting a regression analysis predicting moral goodness ratings by the interaction of the observability manipulation (public vs. private), the actor motive manipulation (reputation-signaling vs. principled vs. none; coded as an indicator variable with "none" as the holdout condition), and our virtue manipulation (generosity vs. impartiality), with Experiment as a covariate. We observe that the three-way interaction terms are not significant (virtue*observability*reputation-signaling: $p$=.339; virtue*observability*principled: $p$=.806), implying the observability*motive interaction does not differ by virtue. We therefore include the virtue manipulation as a covariate in the results presented here (these results are robust to simply omitting the virtue manipulation from the model presented below, see Supplement Section 3). As a result, the regression analysis presented here predicts moral goodness and trait ratings by the interaction of the observability manipulation (public vs. private) and the actor motive manipulation (reputation-signaling vs. principled vs. none), with Experiment and the virtue manipulation as covariates.

*Results*

Overall, we find a significant interaction between observability and motive on moral goodness ratings ($F$(2,8952)=42.11, $p$<.001, $d$=.19). Sensitivity analysis revealed that (at $p$=.05 and power=80%) the minimum detectable effect size for this test is $d$=.07. When we do not stipulate actor motivation, we again observe *virtue discounting*; i.e. public actors are rated as less morally good than private actors (*contrast*=-6.61, 95% CI [-8.30, -4.91], $t$=-12.96, $p$<.001, see Figure 4).

Consistent with our hypothesis (preregistered in E10 and E14), we observe that public and private actors are not rated differently on moral goodness when we stipulate that actors have *reputation-signaling* motivation (*contrast*=-1.75, 95% CI [-5.12, 1.62], $t$=-1.73, $p$=.702) or *principled* motivation (*contrast*=2.80, 95% CI [-.59, 6.19], $t$=2.75, $p$=.182).

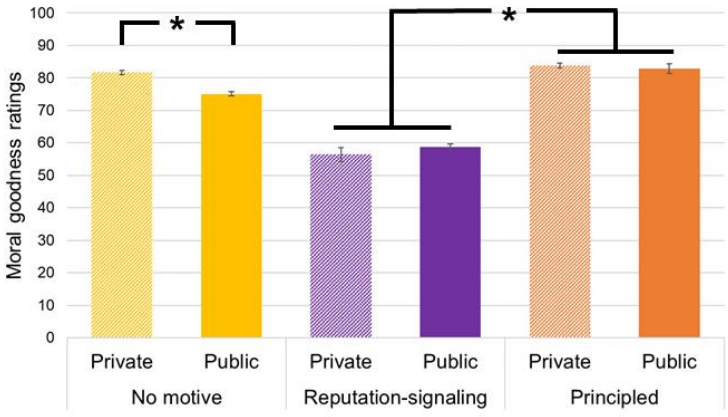


**Fig 4. Observability does not affect actors' perceived virtue when their motivation is stipulated.** Shown are means (with 95% CIs) of trait ratings (0-100 unmarked slider), as a function of whether the actor is said to have engaged in behaviors privately (lines) or publicly (solid) and whether the actor's motivation is stipulated as: none (yellow), reputation-signaling (purple), principled (orange). Significant contrasts denoted with (*). From left to right: private no motive $N$=2,311; public no motive $N$=2,360; private reputation-signaling $N$=485; public reputation-signaling $N$=1,665; private principled $N$=1,679; public principled $N$=472.

*Discussion*

In this analysis, we show that observability does not affect participants' ratings of actors' moral goodness when we stipulate actors' motivation. Specifically, when we describe actors as having *principled* (but not *reputation-signaling*) motivations or *reputation-signaling* (but not *principled*) motivations, participants' moral goodness ratings do not differ by observability condition (i.e. public vs. private). This analysis provides stronger causal evidence of the account we propose in Analysis 2: information about actors' motivations impacts observers' ratings of actors' moral goodness. When this information is not explicit—as in the design of Experiments investigated in Analysis 2 and presumably in real life—people use the observability of actors' behavior as a proxy to infer actors' motivation. When people know actors' motivation (as in the design of Experiments investigated in this analysis), however, this information determines their moral judgment, regardless of the observability of actors' behavior.

## General Discussion

Across three analyses martialing data from 14 experiments (seven preregistered, total $N$=9,360), we provide robust evidence of *virtue discounting*. In brief, we show that when actors' behavior is observable, people are less likely to attribute this behavior to actors' virtue. In Analysis 1—which includes a meta-analysis of all experiments we ran—we show that observability causes virtue discounting (supporting our first preregistered hypothesis), and that this effect is larger in the context of generosity compared to fairness (as impartiality; supporting our second preregistered hypothesis). In Analysis 2, we provide suggestive evidence that participants' motivational inferences mediate a large portion (72.6%) of the effect of observability on their ratings of actors' moral goodness (supporting our third preregistered hypothesis). In Analysis 3, we show that when we experimentally manipulate actors' motivation, observability loses its significant effect on participants' judgments of actors' moral goodness (providing stronger evidence supporting our third preregistered hypothesis). We now consider the contributions of our findings to the empirical literature and to theoretical accounts of "virtue signaling", as well as limitations of the present investigation. Finally, we conclude with practical implications for effective prosocial advocacy.

*Contributions to the literature*

Our intention in coining the term *virtue discounting* is to provide an umbrella concept describing phenomena wherein people are less likely to think that actors' behavior is attributable to their virtue. Our studies, as well as the prior research cited, focuses on the observability of actors' behavior as one reason that people discount others' virtue. Yet we intend that *virtue discounting* encompass other reasons documented in past work, such as actors' intentionality (Weiner & Kukla, 1970), actors' causal responsibility (Pizarro et al., 2003), the benefit to recipients (Klein & Epley, 2014), the cost to actors (Janoff-Bulman et al., 2009), actors' decision speed (Jordan, Hoffman, Nowak, et al., 2016), the frequency of the behavior (Futamura, 2018), and the obligatory nature of the behavior (McManus et al., 2020). In accordance with our definition of virtue, most of these speak to whether an actor has "proper" motivation, and an exciting direction of future research would be to investigate whether motivational inferences serve as a common mechanism across these proximate moderators of moral praise (Anderson et al., 2020; in a sense, the inverse of virtue discounting). Further, other elements of our definition of virtue may also contribute to virtue discounting, including inferences regarding the consistency (or stability) of behavior (as instantiated, e.g., in observers making dispositional vs.

situational attributions to actors' behavior; Kim et al., 2020) as well as heterogeneity in culturally-specific values (as instantiated, e.g., in observers perceiving actors from different cultures or identity groups).

Consistent with the majority of studies demonstrating virtue discounting (e.g. Newman & Cain, 2014), we show robust evidence of this effect in the context of *generosity* (Analysis 1). To this literature, the present investigation adds novel evidence that virtue discounting also occurs in the context of *fairness* (as impartiality), though to a lesser extent. Another avenue for future research builds off our finding of heterogeneity in differential virtue discounting across experiments. Our stimuli were generated by a rigorous, bottom-up method (Kraft-Todd et al., 2022) that involved participants rating example behaviors of each virtue on nine features known to impact moral judgment. It is possible that the heterogeneity we observe is explainable via these feature ratings (potentially, independent of virtue), although our current data are not powered to conclusively test this possibility.

Our investigation of motivational inferences as a mechanism of observability on ratings of moral goodness (Analysis 2) yield insights into: 1) the structure of motivational inferences; 2) a relatively novel motivational construct; and 3) mechanistic explanations of virtue discounting via motivational inferences.

First, following prior work (e.g. Lin-Healy & Small, 2012), we hypothesized that virtue discounting will be explained by observers attributing selfish motivations to publicly virtuous actors. Our exploratory factor analysis of motivational inference items yielded two factors that we labeled *principled* and *reputation-signaling*, and although these factors intuitively seem like conceptual opposites, they are only moderately negatively correlated ($r$=-.46). It may be compelling to broadly conceptualize the construct of "motivation" as a bipolar scale with the endpoints "selfish" and "selfless", but here, we show that the conceptually "selfish" *reputation-signaling* motivation factor and the conceptually "selfless" *principled* motivation factor are less strongly (though still negatively) correlated than one might expect. It is intuitively plausible that actors might be motivated both to help others *and* to have others think well of them. Future work might therefore eschew a simplistic unidimensional conceptualization of motivational inferences, instead taking a more pluralistic approach and extending the present work to include other motivations for prosocial behavior suggested by the literature (Narvaez & Snow, 2019; Reiss & Havercamp, 1998).

Second, we were surprised to find that our *norm-signaling* motivational inference item was not uniquely captured by either of the two factors resulting from our exploratory factor analysis. One previous study has demonstrated the importance of norm-signaling motivational inferences in explaining perceptions of individuals sharing their gender pronouns (Kodipady et al., 2021), although we are unaware of other work exploring this construct. Although it might be argued that the mediation demonstrated by this item (i.e. explaining the effect of observability on moral goodness ratings) was relatively minor (-7.3%), it is worth noting that this was roughly equivalent in magnitude to the effect of our *reputation-signaling* motivation factor (9.3%; technically, an "opposing mediation"; Kenny et al., 1998), which represents the most frequently cited motivational explanation of virtue signaling in previous work.

Third, perhaps the most puzzlingly counterintuitive finding we present is that virtue discounting is largely explained by observers' inferences that public actors have lower *principled* motivations (accounting for 72.3% of the effect), and that *reputation-signaling* inferences mediate only 9.3% of this effect. Building on prior work, we expected that the effect of observability on virtue discounting would be explained by inferred selfish motivations; instead, it

appears that it is actually a decrease in (conceptually "selfless") *principled* motivational inferences, rather than an increase in (conceptually "selfish") *reputation-signaling* motivational inferences that explains virtue discounting in our paradigm. We rule out a possible alternative explanation, that *reputation-signaling* motivational inferences mediate the effect of *principled* motivational inferences on moral goodness ratings (see Supplement Section 8). Still, we believe more research is needed before placing great confidence in this conclusion. For example, previous demonstrations of virtue discounting often employ within-subjects comparisons of public vs. private actors (Lin-Healy & Small, 2012), but we compare public vs. private actors in between-subjects designs. Consistent with the literature on "joint vs. separate evaluation" (Hsee et al., 1999), it could be the case that *reputation-signaling* motivational inferences would emerge as the primary mechanism in a within-subjects design (see, e.g., a paradigm eliciting such reversals of moral judgment; McManus et al., 2020).

Although more research is needed to better understand which specific motivations contribute to virtue discounting, our results from Analysis 3 strengthen our interpretation that motivational inferences can explain the effect of observability in virtue discounting. Additional insight about the role of motivational inferences in virtue discounting is provided by comparing our results from Analysis 3 with our "baseline" (i.e. no observability information) conditions in Analysis 1. Taken together, these results suggest that people assume that actors have *principled* motivations, both when these actors' behavior is conducted privately and also when information about the observability of their behavior is absent. It is unclear whether this pattern of results might be attributed to a simple "behavior-motivation congruency heuristic" (i.e. perhaps people tend to think that the valence of actors' motivation matches their behavior) or reflects some sort of generalized trust (i.e. people assume others have good intentions).

We conclude these considerations by contrasting our results, and the literature on virtue discounting to which they most directly contribute, with a puzzlingly divergent set of results. In many situations, observers react to virtuous behavior with positive valence emotions (see work on, e.g. elevation; Haidt, 2003), and further, that people often emulate virtuous behavior when they observe it (Thomson & Siegel, 2017). Yet, beyond the virtue discounting literature we discuss here, other work also suggests that such "do-gooders" are frequently derogated for their morally-motivated behavior (e.g. Sparkman & Attari, 2020). We propose a fascinating question for future research to explore: what individual- and/or situational-factors moderate such celebration (Bai et al., 2019) versus derogation (Minson & Monin, 2012) responses in social comparison (Mussweiler, 2003)?

*"Virtue signaling"*

Although we have discussed costly signaling theory (Zahavi, 1975) throughout our investigation, we believe further consideration of this theory as it applies to "virtue signaling" and virtue discounting in light of the present investigation is warranted. Echoing the pejorative nature of accusations of "virtue signaling" (Bartholomew, 2015), we propose that most virtuous behaviors are not sufficiently costly to serve, technically, as honest signals of virtue. In previous work, we have demonstrated that (continuous) perceptions of the costliness of actors' behavior contributes to categorical perceptions of them (i.e. as "heroic" or "not heroic"; Kraft-Todd & Rand, 2019). Much remains to be discovered, however, in the perception of virtuous behaviors—and crucially, their costliness—as "signals" of virtue. For example, it would be interesting to investigate the confidence with which people infer actors' virtuous traits from actors' behavior

across a range of behaviors and virtues, and whether such inferences are correlated with perceptions of those behaviors' costliness.

Orthogonal to the sufficient costliness of virtuous behavior, there is additional nuance to consider in the "signaling" aspect of observability. First, in our manipulation of observability (comparing public to private actors), we assume that observers are somehow able to learn about actors' private behavior. This process is non-trivial in the real world, and previous research suggests that observers perceive actors as more virtuous when actors share their good deeds compared to when observers have no knowledge of these behaviors (Berman et al., 2014). Second, observability may admit of degrees, rather than the binary distinction we impose upon it here (i.e. public vs. private), as other research suggests that "subtle" signals can sometimes be preferable to more obvious signals (Bliege Bird et al., 2018).

Finally, we propose an alternative explanation for differential virtue discounting motivated by costly signaling theory. Recent work substantiates a "false-signaling" account for our strong disapproval of hypocrites (Jordan et al., 2017). In sum, people are particularly sensitive to mismatches between actors' public-facing self-presentation of virtue and their perception of actors' true character, which might constitute an "anti-hypocrisy bias." If people have expectations that certain behaviors should be more honest reflections of actors' self-perceived character than others, we might expect greater discounting of these virtues. Consider this result from Analysis 2: the correlation of the *principled* and *reputation-signaling* motivation factors differs by virtue (*impartiality*: $r$=-.35; *generosity*: $r$=-.54). One interpretation is that perceptions of "selfish" vs. "selfless" motivations are more unidimensional (i.e. seen as more essentialized, or binary) for generosity than impartiality (see further discussion of this point above), consistent with the "expectations of honest signaling" account we propose here. In other words, if observers think that actors should only engage in generous (compared to impartial) behaviors in a way that authentically demonstrates their self-perceived character, then any hint of inauthenticity (e.g. behaving publicly rather than privately) might trigger their anti-hypocrisy bias and contribute to virtue discounting.

*Limitations*

The present investigation is not without shortcomings. We believe some caution is warranted regarding the mechanistic account of the effect of observability on moral goodness through motivational inferences given a comparison of our results across Analyses 2 and 3. Although we demonstrate substantial mediation (75.9%) of the effect of observability in Analysis 2, we do not demonstrate full mediation (i.e. the direct effect of observability on moral goodness is still significant), implying that there are likely other mechanisms that we do not measure. In Analysis 3, when we stipulate that virtuous actors have *reputation-signaling* (and not *principled*) motivations, participants' ratings of actors' moral goodness are reduced to a greater extent than when we merely describe public actors (with no motive stipulated). Our manipulation of actor motivation thus may have been an overdetermination of our proposed mechanism, which could imply, for example, that potential mechanisms we failed to consider have a milder effect on moral goodness ratings, or that there are individual differences in our proposed motivational inference mechanism. It could also be the case that observers infer that public actors have mixed motives, e.g. *both* principled and reputation-signaling motivations (see Analysis 2, correlation of these factors, $r$=-.46), which is obscured by our manipulations in Analysis 3 (e.g. we describe actors having reputation-signaling *but not* principled motivations).

The generalizability of our findings may be limited because we investigated only two virtues, measured only six items of motivational inference, and our methods relied on exclusively hypothetical vignette scenarios administered to convenience samples. There is a substantial literature investigating other virtues (e.g. Peterson & Seligman, 2004)—and despite ambiguity in virtue concepts and their operationalization (McGrath, 2014)—we leave it to future work to examine whether the virtue discounting effect generalizes to other virtues (see, e.g. research on trustworthiness; Jordan, Hoffman, Bloom, et al., 2016). Similarly, we also recognize that motivation (like virtue) is a multidimensional construct (e.g. Reiss & Havercamp, 1998) and that previous work has suggested that distinct virtues are driven by distinct motivations (Narvaez & Snow, 2019). Future work might also measure a greater range of motivations for each virtue (see, e.g. Kodipady et al., 2021). Due to the limitations of drawing generalizable conclusions about moral judgment from hypothetical vignette studies (Feldman-Hall et al., 2012) and online convenience samples (Simons et al., 2017), we encourage future work to employ more diverse methods. We also expect that different cultures may discount virtue differently. Fundamental to our definition of virtue is that the value of a trait that constitutes virtue derives from the social norms of the culture under investigation. Therefore, in addition to merely replicating our results in non-WEIRD samples (Henrich et al., 2010), we propose that our results are likely to vary in accordance with the value that different cultures place on different virtues.

*Practical implications and conclusion*

Finally, our findings yield insights about effective advocacy of real-world virtuous behaviors. We started with a paradox of public prosociality: while others can only infer your virtue and learn from your prosocial example by observing your behavior, they simultaneously may be less likely to attribute your behavior to your virtue when your behavior is observable. Our results are consistent with this paradox of public prosociality, particularly when we consider the "opposing mediation" (Kenny et al., 1998; i.e., the negative mediation effect) of our *norm-signaling* item on moral goodness ratings. To reiterate: the effect of observability on motivation inference is in the same direction for *norm-signaling* and *reputation-signaling* motivations, whereas the effect of motivation inference on moral goodness is in the same direction for *norm-signaling* and *principled* motivations. In other words, observable actors can be perceived positively if observers infer that actors have *norm-signaling* motivations; yet, observably virtuous actors also can be perceived negatively if observers infer that actors have low *principled* motivations (and also high *reputation-signaling* motivations). This pattern of results suggests that, to be perceived positively and potentially influence observer behavior, actors should consider how to convincingly convey *norm-signaling* motivations *without* engendering perceptions of low *principled* motivations (and also high *reputation-signaling* motivations).

Further, following our preceding discussion regarding a costly signaling account of virtue discounting, it could be the case that for virtues where "selfish" and "selfless" motivations are perceived more dichotomously, actors need to more clearly demonstrate their "selfless" motivations to avoid observers discounting their virtue. This concern echoes our second preregistered hypothesis ("differential virtue discounting"), which suggests that practitioners (as well as researchers) should carefully examine which virtues they are displaying in their advocacy, and consider whether these are particularly susceptible to virtue discounting and diluting their overall message.

If prosocial advocates can credibly demonstrate that they are motivated by *principles* and *norm-signaling*, and not by *reputation-signaling*—particularly in the context of some virtues

(e.g. generosity)—they may be able to resolve the paradox of public prosociality. One might wonder, "how can prosocial advocates credibly demonstrate their motivation?" In addition to a direction for future research, we leave this question to prosocial advocates as a prompt for reflection. The answer will undoubtedly be contingent on the individual, audience, and setting, but finding it may be a key to catalyzing the contagion of prosociality.

**Acknowledgments**
We would like to thank the members of [anonymized for peer review] and the reviewers for their generous feedback.

**Disclosure statement**
The authors declare no conflict of interest.

**Funding**
This research was made possible by funding by the John Templeton Foundation, The Virtue Project at [anonymized for peer review], and NSF award #[anonymized for peer review].

**Open Practices**

Preregistrations were conducted for: E8 (https://aspredicted.org/FLG_VRD), E9 (https://aspredicted.org/NUP_ESB), E10 (https://aspredicted.org/FCO_HXM), E11 (https://aspredicted.org/2Y2_YKT), E12 (https://aspredicted.org/AES_HGO), E13 (https://aspredicted.org/G6S_S35), and E14 (https://aspredicted.org/SHW_9PN). All data and code are publicly available at: (https://osf.io/sud3m/?view_only=380a169770b9474f93d2b5b73adc7410).

## References

Anderson, R. A., Crockett, M. J., & Pizarro, D. (2020). A Theory of Moral Praise. *Trends in Cognitive Sciences*, *24*(9), 694–703. https://doi.org/10.1016/j.tics.2020.06.008

Aquino, K., & Reed, A. (2002). The self-importance of moral identity. *Journal of Personality and Social Psychology*, *83*(6), 1423.

Arechar, A. A., Kraft-Todd, G. T., & Rand, D. G. (2017). Turking overtime: How participant characteristics and behavior vary over time and day on Amazon Mechanical Turk. *Journal of the Economic Science Association*, 1–11. https://doi.org/10.1007/s40881-017-0035-0

Aristotle. (1999). *Nicomachean Ethics*. Hackett Publishing Company, Inc.

Bai, F., Ho, G. C. C., & Yan, J. (2019). Does virtue lead to status? Testing the moral virtue theory of status attainment. *Journal of Personality and Social Psychology*. PsycARTICLES. https://doi.org/10.1037/pspi0000192

Barclay, P. (2016). Biological markets and the effects of partner choice on cooperation and friendship. *Current Opinion in Psychology*, *7*, 33–38. https://doi.org/10.1016/j.copsyc.2015.07.012

Bartholomew, J. (2015, April 18). The awful rise of "virtue signalling." *The Spectator*.

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk. *Political Analysis*, *20*(3), 351–368. https://doi.org/10.1093/pan/mpr057

Berman, J. Z., Levine, E. E., Barasch, A., & Small, D. A. (2014). The Braggart's Dilemma: On the Social Rewards and Penalties of Advertising Prosocial Behavior. *Journal of International Marketing*.

Bicchieri, C. (2006). *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press.

Bliege Bird, R., Ready, E., & Power, E. A. (2018). The social significance of subtle signals. *Nature Human Behaviour*, *2*(7), 452–457. https://doi.org/10.1038/s41562-018-0298-3

Boyd, R., & Richerson, P. J. (1992). How microevolutionary processes give rise to history. In M. H. Niteki & D. V. Nitecki (Eds.), *Evolution and History*. State University of New York Press.

Carlson, R. W., Bigman, Y. E., Gray, K., Ferguson, M. J., & Crockett, M. J. (2022). How inferred motives shape moral judgements. *Nature Reviews Psychology*, *1*(8), 468–478. https://doi.org/10.1038/s44159-022-00071-x

Cokelet, B., & Fowers, B. J. (2019). Realistic virtues and how to study them: Introducing the STRIVE-4 model. *Journal of Moral Education*, *48*(1), 7–26. https://doi.org/10.1080/03057240.2018.1528971

Crockett, M. J. (2017). Moral outrage in the digital age. *Nature Human Behaviour*, *1*(11), 769–771. https://doi.org/10.1038/s41562-017-0213-3

De Freitas, J., DeScioli, P., Thomas, K. A., & Pinker, S. (2019). Maimonides' ladder: States of mutual knowledge and the perception of charitability. *Journal of Experimental Psychology: General*, *148*(1), 158–173. PsycARTICLES. https://doi.org/10.1037/xge0000507

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Feldman-Hall, O., Mobbs, D., Evans, D., Hiscox, L., Navrady, L., & Dalgleish, T. (2012). What we say and what we do: The relationship between real and hypothetical moral choices. *Cognition*, *123*(3), 434–441. https://doi.org/10.1016/j.cognition.2012.02.001

Friedman, M. (1953). *Essays in Positive Economics*. University of Chicago Press.

Futamura, I. (2018). Is extraordinary prosocial behavior more valuable than ordinary prosocial behavior? *PLOS ONE*, *13*(4), e0196340. https://doi.org/10.1371/journal.pone.0196340

Grubbs, J. B., Warmke, B., Tosi, J., James, A. S., & Campbell, W. K. (2019). Moral grandstanding in public discourse: Status-seeking motives as a potential explanatory mechanism in predicting conflict. *PLOS ONE*, *14*(10), e0223749. https://doi.org/10.1371/journal.pone.0223749

Gulliford, L., & Roberts, R. C. (2018). Exploring the "unity" of the virtues: The case of an allocentric quintet. Theory & Psychology, 28(2), 208–226. https://doi.org/10.1177/0959354317751666

Haidt, J. (2003). Elevation and the positive psychology of morality. In C. L. M. Keyes & J. Haidt (Eds.), *Flourishing: Positive psychology and the life well-lived.* (pp. 275–289). American Psychological Association.

Harris, E., & Bardey, A. C. (2019). Do Instagram Profiles Accurately Portray Personality? An Investigation Into Idealized Online Self-Presentation. *Frontiers in Psychology*, *10*. https://www.frontiersin.org/articles/10.3389/fpsyg.2019.00871

Hayes, A. F. (2013). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach.* Guilford Press.

Henrich, J. (2009). The evolution of costly displays, cooperation and religion: Credibility enhancing displays and their implications for cultural evolution. *Evolution and Human Behavior*, *30*(4), 244–260.

Henrich, J. (2016). *The Secret of Our Success: How Culture Is Driving Human Evolution, Domesticating Our Species, and Making Us Smarter*. Princeton University Press.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2–3), 61–83.

Hsee, C. K., Loewenstein, G. F., Blount, S., & Bazerman, M. H. (1999). Preference reversals between joint and separate evaluations of options: A review and theoretical analysis. *Psychological Bulletin*, *125*(5), 576–590. PsycARTICLES. https://doi.org/10.1037/0033-2909.125.5.576

Hume, D. (1902). *Enquiries Concerning Human Understanding and Concerning the Principles of Morals*. Clarendon Press.

Janoff-Bulman, R., Sheikh, S., & Hepp, S. (2009). Proscriptive versus prescriptive morality: Two faces of moral regulation. *Journal of Personality and Social Psychology*, *96*(3), 521–537. https://doi.org/10.1037/a0013779

Jones, E. E., & Pittman, T. S. (1982). Toward a general theory of strategic self-presentation. *Psychological Perspectives on the Self*, *1*(1), 231–262.

Jordan, J. J., Hoffman, M., Bloom, P., & Rand, D. G. (2016). Third-party punishment as a costly signal of trustworthiness. *Nature*, *530*(7591), 473–476. https://doi.org/10.1038/nature16981

Jordan, J. J., Hoffman, M., Nowak, M. A., & Rand, D. G. (2016). Uncalculating cooperation is used to signal trustworthiness. *Proceedings of the National Academy of Sciences*, *113*(31), 8658–8663. https://doi.org/10.1073/pnas.1601280113

Jordan, J. J., Sommers, R., Bloom, P., & Rand, D. G. (2017). Why Do We Hate Hypocrites? Evidence for a Theory of False Signaling. *Psychological Science*, *28*(3), 356–368. https://doi.org/10.1177/0956797616685771

Keltner, D., Kogan, A., Piff, P. K., & Saturn, S. R. (2014). The Sociocultural Appraisals, Values, and Emotions (SAVE) Framework of Prosociality: Core Processes from Gene to Meme. *Annual Review of Psychology*, *65*(1), 425–460. https://doi.org/10.1146/annurev-psych-010213-115054

Kenny, D. A., Kashy, D. A., & Bolger, N. (1998). Data analysis in social psychology. In *Data analysis in social psychology* (4th ed., Vol. 1, pp. 233–265). McGraw-Hill.

Kim, M., Park, B., & Young, L. (2020). The Psychology of Motivated versus Rational Impression Updating. *Trends in Cognitive Sciences*, *24*(2), 101–111. https://doi.org/10.1016/j.tics.2019.12.001

Kleiman-Weiner, M., Saxe, R., & Tenenbaum, J. B. (2017). Learning a commonsense moral theory. *Cognition*, *167*, 107–123. https://doi.org/10.1016/j.cognition.2017.03.005

Kleiman-Weiner, M., Shaw, A., & Tenenbaum, J. B. (2017). *Constructing social preferences from anticipated judgments: When impartial inequity is fair and why?* 676–681.

Klein, N., & Epley, N. (2014). The topography of generosity: Asymmetric evaluations of prosocial actions. *Journal of Experimental Psychology: General*, *143*(6), 2366.

Kodipady, A., Kraft-Todd, G. T., Sparkman, G., Hu, B., & Young, L. (2021). *Beyond Virtue Signaling: Perceived Motivations for Pronoun Sharing*. PsyArXiv. https://doi.org/10.31234/osf.io/s6ct9

Kraft-Todd, G. T., Bollinger, B., Gillingham, K., Lamp, S., & Rand, D. G. (2018). Credibility-enhancing displays promote the provision of non-normative public goods. *Nature*, *563*(7730), 245–248. https://doi.org/10.1038/s41586-018-0647-4

Kraft-Todd, G. T., Kleiman-Weiner, M., & Young, L. (2022). *Operationalizing and Dissociating Virtues from the 'Bottom up': A Case study of Generosity vs. Impartiality*. PsyArXiv. https://doi.org/10.31234/osf.io/3paqs

Kraft-Todd, G. T., & Rand, D. G. (2019). Rare and Costly Prosocial Behaviors Are Perceived as Heroic. *Frontiers in Psychology*, *10*(234). https://doi.org/10.3389/fpsyg.2019.00234

Kraft-Todd, G. T., Yoeli, E., Bhanot, S., & Rand, D. G. (2015). Promoting cooperation in the field. *Current Opinion in Behavioral Sciences*, *3*, 96–101. https://doi.org/10.1016/j.cobeha.2015.02.006

Lakens, D., & Caldwell, A. R. (2021). Simulation-Based Power Analysis for Factorial Analysis of Variance Designs. *Advances in Methods and Practices in Psychological Science*, *4*(1), 2515245920951503. https://doi.org/10.1177/2515245920951503

Laufer, W. S. (2003). Social Accountability and Corporate Greenwashing. *Journal of Business Ethics*, *43*(3), 253–261. https://doi.org/10.1023/A:1022962719299

Lenhard, W., & Lenhard, A. (2016). Calculation of Effect Sizes. *Psychometrica*. https://doi.org/10.13140/RG.2.1.3478.4245

Levine, S., Kleiman-Weiner, M., Schulz, L., Tenenbaum, J., & Cushman, F. (2020). The logic of universalization guides moral judgment. *Proceedings of the National Academy of Sciences*, *117*(42), 26158–26169. https://doi.org/10.1073/pnas.2014505117

Lin-Healy, F., & Small, D. A. (2012). Cheapened altruism: Discounting personally affected prosocial actors. *Organizational Behavior and Human Decision Processes*, *117*(2), 269–274.

Maimonides, M. (1170). Laws of gifts to the poor. *Mishneh Torah*, *10*, 7–14.

Mazar, N., Amir, O., & Ariely, D. (2008). The dishonesty of honest people: A theory of self-concept maintenance. *Journal of Marketing Research*, *45*(6), 633–644.

McGrath, R. E. (2014). Scale- and Item-Level Factor Analyses of the VIA Inventory of Strengths. *Assessment*, *21*(1), 4–14. https://doi.org/10.1177/1073191112450612

McManus, R. M., Kleiman-Weiner, M., & Young, L. (2020). What We Owe to Family: The Impact of Special Obligations on Moral Judgment. *Psychological Science*, *0*(0), 0956797619900321. https://doi.org/10.1177/0956797619900321

Mende-Siedlecki, P., Baron, S. G., & Todorov, A. (2013). Diagnostic Value Underlies Asymmetric Updating of Impressions in the Morality and Ability Domains. *The Journal of Neuroscience*, *33*(50), 19406. https://doi.org/10.1523/JNEUROSCI.2334-13.2013

Minson, J. A., & Monin, B. (2012). Do-Gooder Derogation: Disparaging Morally Motivated Minorities to Defuse Anticipated Reproach. *Social Psychological and Personality Science*, *3*(2), 200–207. https://doi.org/10.1177/1948550611415695

Morris, M. W., & Larrick, R. P. (1995). When one cause casts doubt on another: A normative analysis of discounting in causal attribution. *Psychological Review*, *102*(2), 331–355. https://doi.org/10.1037/0033-295X.102.2.331

Mullen, E., & Monin, B. (2016). Consistency Versus Licensing Effects of Past Moral Behavior. *Annual Review of Psychology*, *67*(1), 363–385. https://doi.org/10.1146/annurev-psych-010213-115120

Mussweiler, T. (2003). Comparison processes in social judgment: Mechanisms and consequences. *Psychological Review*, *110*(3), 472.

Narvaez, D., & Snow, N. (eds. ). (2019). Self, Motivation and Virtue Studies [Special issue]. *Journal of Moral Education*, *48*(1).

Newman, G. E., & Cain, D. M. (2014). Tainted Altruism: When Doing Some Good Is Evaluated as Worse Than Doing No Good at All. *Psychological Science*, *25*(3), 648–655. https://doi.org/10.1177/0956797613504785

Niemi, L., & Young, L. (2017). Who Sees What as Fair? Mapping Individual Differences in Valuation of Reciprocity, Charity, and Impartiality. *Social Justice Research*, *30*(4), 438–449. https://doi.org/10.1007/s11211-017-0291-4

Nowak, M. A., & Sigmund, K. (1998). The Dynamics of Indirect Reciprocity. *Journal of Theoretical Biology*, *194*(4), 561–574. https://doi.org/10.1006/jtbi.1998.0775

Ohtsuki, H., & Iwasa, Y. (2006). The leading eight: Social norms that can maintain cooperation by indirect reciprocity. *J Theor Biol*, *239*(4), 435–444.

Olivoto, T., & L'ucio, A. D. (2020). metan: An R package for multi-environment trial analysis. *Methods in Ecology and Evolution*, *11*(6), 783–789. https://doi.org/10.1111/2041-210X.13384

Peterson, C., & Seligman, M. E. P. (2004). *Character strengths and virtues: A classification and handbook.* American Psychological Association.

Pizarro, D., & Tannenbaum, D. (2012). Bringing character back: How the motivation to evaluate character influences judgments of moral blame. In M. Mikulincer & P. R. Shaver (Eds.), *The social psychology of morality: Exploring the causes of good and evil* (1041017562; 2011-09275-005; pp. 91–108, Chapter xvi, 440 Pages). American Psychological Association, American Psychological Association, Washington, DC; APA PsycBooks®. https://doi.org/10.1037/13091-005

Pizarro, D., Uhlmann, E., & Salovey, P. (2003). Asymmetry in Judgments of Moral Blame and Praise: The Role of Perceived Metadesires. *Psychological Science*, *14*(3), 267–272. https://doi.org/10.1111/1467-9280.03433

Raihani, N. J., & Power, E. A. (2021). No good deed goes unpunished: The social costs of prosocial behaviour. *Evolutionary Human Sciences*, *3*, e40. https://doi.org/10.1017/ehs.2021.35

Reiss, S., & Havercamp, S. M. (1998). Toward a comprehensive assessment of fundamental motivation: Factor structure of the Reiss Profiles. *Psychological Assessment*, *10*(2), 97–106. http://dx.doi.org/10.1037/1040-3590.10.2.97

Schneewind, J. B. (1990). The Misfortunes of Virtue. *Ethics*, *101*(1), 42–63.

Shaw, A. (2013). Beyond "to Share or Not to Share": The Impartiality Account of Fairness. *Current Directions in Psychological Science*, *22*(5), 413–417. https://doi.org/10.1177/0963721413484467

Shaw, A. (2016). Fairness: What it isn't, what it is, and what it might be for. In D. C. Geary & D. B. Berch (Eds.), *Evolutionary Perspectives on Child Development and Education* (pp. 193–214). Springer International Publishing.

Shaw, A., & Olson, K. R. (2012). Children discard a resource to avoid inequity. *Journal of Experimental Psychology: General*, *141*(2), 382–395.

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on Generality (COG): A Proposed Addition to All Empirical Papers. *Perspectives on Psychological Science*, *12*(6), 1123–1128. https://doi.org/10.1177/1745691617708630

Smith, S., Windmeijer, F., & Wright, E. (2015). Peer Effects in Charitable Giving: Evidence from the (Running) Field. *The Economic Journal*, *125*(585), 1053–1071. https://doi.org/10.1111/ecoj.12114

Snow, N. E. (2019). Positive psychology, the classification of character strengths and virtues, and issues of measurement. *The Journal of Positive Psychology*, *14*(1), 20–31. https://doi.org/10.1080/17439760.2018.1528376

Sobel, M. E. (1982). Asymptotic Confidence Intervals for Indirect Effects in Structural Equation Models. *Sociological Methodology*, *13*, 290–312. JSTOR. https://doi.org/10.2307/270723

Sparkman, G., & Attari, S. Z. (2020). Credibility, communication, and climate change: How lifestyle inconsistency and do-gooder derogation impact decarbonization advocacy. *Energy Research & Social Science*, *59*, 101290. https://doi.org/10.1016/j.erss.2019.101290

Spring, V. L., Cameron, C. D., & Cikara, M. (2018). The Upside of Outrage. *Trends in Cognitive Sciences*, *22*(12), 1067–1069. https://doi.org/10.1016/j.tics.2018.09.006

Tamir, D. I., & Thornton, M. A. (2018). Modeling the Predictive Social Mind. *Trends in Cognitive Sciences*, *22*(3), 201–212. https://doi.org/10.1016/j.tics.2017.12.005

Tankard, M. E., & Paluck, E. L. (2016). Norm Perception as a Vehicle for Social Change. *Social Issues and Policy Review*, *10*(1), 181–211. https://doi.org/10.1111/sipr.12022

Thomson, A. L., & Siegel, J. T. (2017). Elevation: A review of scholarship on a moral and other-praising emotion. *The Journal of Positive Psychology*, *12*(6), 628–638. https://doi.org/10.1080/17439760.2016.1269184

Tosi, J., & Warmke, B. (2020). *Grandstanding: The Use and Abuse of Moral Talk*. Oxford University Press.

Tzu, L., & Lau, D. C. (trans). (1963). *Tao Te Ching*. Penguin Books.

UCLA: Statistical Consulting Group. (2021). *Introduction to STATA*. https://stats.idre.ucla.edu/stata/faq/how-can-i-analyze-multiple-mediators-in-stata/

Weiner, B., & Kukla, A. (1970). An attributional analysis of achievement motivation. *Journal of Personality and Social Psychology*, *15*(1), 1–20. https://doi.org/10.1037/h0029211

West, P. (2004). *Conspicuous compassion: Why sometimes it really is cruel to be kind*. The Cromwell Press.

Winer, B. J., Brown, D. R., & Michels, K. M. (1991). *Statistical Principles in Experimental Design* (3rd ed.). McGraw–Hill.

Yoeli, E., & Hoffman, M. (2022). *Hidden Games*. Basic Books.

Yoeli, E., Hoffman, M., Rand, D. G., & Nowak, M. A. (2013). Powering up with indirect reciprocity in a large-scale field experiment. *Proceedings of the National Academy of Sciences*, *110*(Supplement 2), 10424–10429. https://doi.org/10.1073/pnas.1301210110

Zahavi, A. (1975). Mate selection—A selection for a handicap. *Journal of Theoretical Biology*, *53*(1), 205–214.