Differential discounting of virtue signaling:
Public virtue is perceived less favorably than private virtue for generosity but not impartiality

Gordon T. Kraft-Todd*[a], Max Kleiman-Weiner[b], Liane Young[a]
[a]Department of Psychology and Neuroscience, McGuinn 300, 140 Commonwealth Ave, Boston College, Chestnut Hill, MA 02467 ; [b]Department of Psychology, Harvard University, Cambridge, MA 02138
*Corresponding author: gordon.kraft-todd@bc.edu; Cell: 978-621-6120; Fax: 617-552-0523

Abstract
There is a paradox in our desire to be seen as virtuous. If we are modest in showing our virtue, others will not be able to see this quality in us; yet, if we show off our virtue, others may think that we do so only for social credit. Here, we investigate how *virtue signaling* works across two distinct virtues—generosity and impartiality—in 3 online studies (*N*=2,413). We demonstrate the novel phenomenon of *differential virtue discounting*, revealing that participants perceive actors who demonstrate virtue in public to be less virtuous than actors who demonstrate virtue in private, and, critically, that this effect is greater for generosity than impartiality. Further, we provide evidence that motivational attributions are a mechanism of differential virtue discounting using both correlational and experimental mediation paradigms. Specifically, we find that two factors of motivational attributions—reputational motivation and moral motivation—explain the effect of observability on perceptions of virtue. Our supplement includes all data collected (inclusive of main text, 12 studies, total *N*=6,126) and 7 analyses as robustness checks and demonstrations of boundary conditions. We discuss how these findings and our novel terminology can shed light on open questions in the social perception of reputation and motivation.

Word Count: 10,015
Author note: All data and Study code for all studies have been made publicly available via the Open Science Framework and can be accessed at https://osf.io/xhcd3/?view_only=d82d7b84b58d419187b9445f8b130c74. All stimuli and complete experimental instructions can be found in SI. Preregistrations available at: Study 1 (https://aspredicted.org/ZRN_YTE), Study 2b (http://aspredicted.org/blind.php?x=zw4ee5), Study 2c (http://aspredicted.org/blind.php?x=s6v2bx), Study 3 (https://aspredicted.org/FCO_HXM), and SI Study 8 (https://aspredicted.org/AES_HGO).

When you see a list of people who donated to a non-profit, you might think that the people whose names are listed are less generous than those who chose to be listed as "anonymous". If this is the case, your intuitions would be aligned with research demonstrating that people perceive charitable donations as less morally good when they infer ulterior motives, such as selfish material, emotional, or reputational benefits (Barclay & Willer, 2007; Lin-Healy & Small, 2012; Newman & Cain, 2014). But does this pattern hold for virtues other than generosity, such as impartiality? Here we demonstrate the novel phenomenon of *differential virtue discounting*, revealing that public (compared to private) acts of impartiality are not discounted to the same extent as public (compared to private) generosity. Further, we show how observers' motivational inferences can mechanistically contribute to this effect. Finally, we introduce terminology that helps to explicitly organize relevant prior work.

Review of concepts and previous research

We begin with the concept of *virtue signaling*, which has only recently entered the cultural lexicon (Bartholomew, 2015), and we define as a public act of prosocial behavior intended to demonstrate the actor's trait virtue (also sometimes referred to as "performative virtue"). The term was coined (and is often used) to pejoratively describe observers' perceptions of behaviors enabled by social media where actors invest minimal effort to widely broadcast their support for a cause (e.g. "outrage"; Crockett, 2017; Spring et al., 2018). What motivates actors to engage in such behavior? Individuals demonstrate concern for their reputations, i.e., how they are perceived by others (Emler, 1990), and actively engage in managing others' impressions of them (Jones & Pittman, 1982). A reputation for virtue, specifically, can grant individuals higher social status (Bai, 2017), and high social status can in turn lead to greater wealth (Henrich & Gil-White, 2001) and well-being (Marmot et al., 1984). It is therefore no surprise that people want others to think they are virtuous. Consistent with this logic, there is ample evidence that when individuals' behavior is observable to others, individuals are more likely to behave prosocially; e.g. in the lab (Milinski et al., 2002) and in the field (Kraft-Todd et al., 2015; Yoeli et al., 2013). Social media effectively increases the observability of people's behavior—or at least, their self-presentation of it—and so it follows that people would be more likely to share virtuous behavior on these platforms. Why then, do observers of such behavior sometimes describe it as "virtue signaling"?

We introduce the term *virtue discounting* to describe the devaluing of virtuous behavior (as the pejorative connotation of virtue signaling implies) often as a result of observers' inferring actors' ulterior, selfish motivations because the behavior is seen as performative. We coin "virtue discounting" as a neologism of "temporal discounting" (Frederick et al., 2002), which describes devaluing future rewards in favor of present rewards. Both terms describe the dimensional valuation of a concept ("reward" in temporal discounting and "virtue" in virtue discounting). The dimension of valuation in temporal discounting is time (from present to future), and we propose that the dimension of valuation in virtue discounting, is perceived actor motivation (from selfish to selfless). Our conceptualization unites thematically related prior work—including that regarding charitable giving cited above—as well as work on "conspicuous compassion" (West, 2004), which is most often applied to charitable giving ("conspicuous donation behavior"; Grace & Griffin, 2006). Other studies document similar phenomena related to reputation management; for example: when observers know that an actor knows that they are being observed they infer that the actor is motivated by improving their reputation (rather than by their authentic trait prosociality; e.g. Barclay & Willer, 2007); and actors will forgo learning the cost of cooperation

to be perceived as more trustworthy (Jordan, Hoffman, Nowak, et al., 2016). Accordingly, we hypothesize that the mechanism of virtue discounting is observers' inferring that actors behaving virtuously actually have selfish motivations. Therefore, pejorative accusations of "virtue signaling" (or related terminology) would be instances of "virtue discounting" in which observers believe that actors have selfish, ulterior motives for their conspicuous displays of virtue. In these cases, it is likely that actors are motivated by a desire to improve their reputations, and accordingly we propose that the motivational mediator of "virtue signaling" is "reputation signaling".

The virtue of *generosity*—which we define as trait willingness to confer benefits to others at cost to oneself—features most prominently in historical discussion (De Freitas et al., 2019; Maimonides, 1170) and contemporary research on the phenomenon of virtue signaling (Bereczkei et al., 2010; Bliege Bird et al., 2001; Gurven et al., 2000; Lin-Healy & Small, 2012; Newman & Cain, 2014). Yet *virtue* is not a unitary construct; in treatments both ancient (Aristotle, c350BCE/1999) and modern (Graham et al., 2011; Peterson & Seligman, 2004), it is considered a collection of conceptually distinct morally admirable traits. Relevant to the present investigation, there are a few examples of research demonstrating virtue signaling focused on other virtues, e.g. trustworthiness (Jordan, Hoffman, Bloom, et al., 2016) and impartiality (Kleiman-Weiner et al., 2017). Still, we are unaware of any research directly comparing virtue signaling across virtues as we do here.

Specifically, we contrast generosity with *impartiality*—which we define as trait desire to treat others equally and without bias—following previous work that conceptually disambiguates these virtues (Shaw, 2013). We believe impartiality is a particularly interesting virtue to contrast with generosity because both virtues are often explained via appeals to promoting cooperation (Fehr et al., 2008), although recent work has begun to dissociate the evolutionary functions of these virtues (Shaw, 2016). Further, there is mounting empirical evidence that distinguishes them: for example, in a resource distribution paradigm where participants must divide a pot with an odd number of resources (Shaw & Olson, 2012), rather than distributing the entire pot (the generous choice), they choose to destroy a resource so that the pot is divided equally (the impartial choice). Also, people engage in behavior that is actually biased—giving fewer resources to a deserving friend—in order to not appear biased to others, because giving to a friend may appear nepotistic (Shaw et al., 2018).

Hypotheses and design strategy

We build on work disambiguating generosity and impartiality by employing a novel "bottom-up" two-stage method for differentiating virtues that has two key design strengths: first, it avoids the introduction of experimenter bias in stimulus generation; and second, it reduces ambiguity in participants' comprehension of abstract concepts by operationalizing these concepts with concrete examples (this method has been previously employed in studies in other domains that also investigate perceptions of abstract concepts such as "heroism"; Kraft-Todd & Rand, 2019). In Stage 1, we ask participants to provide behaviors that are examples of the virtues (and provide a definition for each). In Stage 2, we ask an independent sample to rate the behaviors on 9 dimensions of interest. We interpret the mean ratings of all participant-generated example behaviors of each virtue on these 9 dimensions as capturing general perceptions of these virtues, and comparing these aggregate ratings allows us to observe differences in how the virtues are perceived.

We hypothesize that generosity and impartiality (as well as other virtues) might be further distinguished by the extent to which they are discounted, i.e., that we may observe *differential virtue discounting*. In other words, observers may infer greater selfish motivations for some virtues than others. In keeping with the analogy to temporal discounting, where previous work has shown individual differences in this process (e.g. Shamosh et al., 2008), we hypothesize that there will similarly be *virtue*-level differences in virtue discounting. Further, we hypothesize that there will be situational moderators of virtue discounting (analogous to, e.g., scarcity for temporal discounting; Griskevicius et al., 2013). Specifically—motivated by previously cited work on both the effect of observability on prosocial behavior (Yoeli et al., 2013) as well as "conspicuous compassion" (West, 2004)—we hypothesize that an important moderator of differential virtue discounting will be *observability*, i.e. whether behavior is conducted in public vs. private.

To summarize our hypotheses, we propose that we will observe differential virtue discounting—i.e., that public (compared to private) acts of generosity will be discounted more than public (compared to private) acts of impartiality—and that observers' perceptions that actors are selfishly motivated will be a mechanism of this effect. As we have discussed, there is ample evidence that public displays of generosity are discounted, yet our hypothesis that impartiality will not be discounted to the same extent is motivated by the intuition that actors might receive greater rewards (and therefore have greater selfish motivations) for being perceived as generous compared to impartial. Consider, for example, an observer's desire to interact with an actor who is seen as an exemplar of each virtue: while the observer could at most expect fair treatment from an extremely impartial actor, they might expect special treatment from an extremely generous actor. Thus, if an actor is *perceived as* generous (even if they are not), this might make them more attractive than if they are perceived as impartial. It follows that actors would have greater selfish motivation to appear generous, compared to impartial, and that, anticipating this, observers would be more likely to discount public acts of generosity.

To summarize our argument, public acts of virtue present observers with an inferential puzzle sometimes leading to accusations of virtue signaling: do actors have authentic prosocial motives, or are they (merely) trying to demonstrate their virtue to observers (i.e., motivated to improve their reputation)? Compared to private acts of virtue—where virtue signaling is not a concern—public acts of virtue are sometimes perceived as less morally good (i.e. virtue discounting). Yet, the extent to which observers discount public acts of virtue may depend on the specific virtue being demonstrated (i.e. differential virtue discounting).

Here, we first explore employ a novel method to disambiguate perceptions of two virtues—generosity and impartiality—through multidimensional ratings of participant-generated examples of each virtue (Study 1). Then, we provide evidence to support the differential virtue discounting hypothesis, showing that making actors' behavior observable leads to greater discounting of generosity compared to impartiality, as well as suggestive evidence that observers' inferences about actors' motivations may be a mechanism of this effect (using correlational mediation; Study 2). Finally, we present causal evidence for this mechanism through an experimental mediation paradigm, showing that a direct manipulation of actors' motivations substantially reduces the effect of observability on perceived virtue (Study 3). We discuss the implications of these findings for related literatures, supplemental analyses (see SI), and address open questions for future research.

General Methods

Online experiments were conducted using Qualtrics survey software, a convenience sample of participants were recruited using the crowdsourcing tool Amazon Mechanical Turk (Arechar et al., 2017; Berinsky et al., 2012). Across *k* studies (*k*=3) presented in the main text, we requested *N*=2,525 participants (*N*=100/condition in all Studies except Study 1 as well as baseline observability conditions in Studies 2a and 2b, see SI Analysis 1). 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*=2,413 participants (44.1% female, average age=36.7 years). Inclusive of all studies (*k*=12) in the main text as well as SI (with the same exclusions), our total sample was *N*=6,126 participants, 44.3% female, average age=36.2 years. Informed consent was obtained from all participants, who completed studies in *median*=3 minutes and were paid $0.60 for their participation. At the end of each Study (except SI Study 1), we presented (in randomized order) participants with basic demographic questions (gender, age, race, income, education, and political affiliation; see SI for complete experimental instructions and SI Table 2 for summary of participant demographics in each Study). Preregistrations were conducted for: Study 1 (https://aspredicted.org/ZRN_YTE), Study 2b (http://aspredicted.org/blind.php?x=zw4ee5), Study 2c (http://aspredicted.org/blind.php?x=s6v2bx), Study 3 (https://aspredicted.org/FCO_HXM), and SI Study 8 (https://aspredicted.org/AES_HGO). All data and code are publicly available at: (https://osf.io/xhcd3/).

Data analysis for all Studies was completed using STATA 16 and effect sizes (Cohen's D) were obtained through use of an online calculator (Lenhard & Lenhard, 2016). For multivariate regressions, 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 (Study 2; SI Analysis 3) are 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). Experimental mediation design (Study 3) follows theoretical guidelines (Imai et al., 2013; Pirlott & MacKinnon, 2016) and previous implementation (e.g. Kraft-Todd et al., 2018).

We conduct power analysis using the Superpower package in R software (Lakens & Caldwell, 2021), using data from Study 2a (which was not preregistered). With a desired effect size of *d*=.30 (for the virtue*observability interaction; see SI Table 3), our sample size of *N*=100 per cell was powered at 83.25% with an alpha level of .05. All other studies with the same effect of interest (i.e. the virtue*observability interaction, Studies 2b, 2c, and SI Studies 2-9) similarly had a sample size of *N*=100 per cell.

For concision, we report the studies (*k*=3) bearing most crucially on our hypotheses in main text, though to avoid file-drawer concerns we include all other studies (*k*=9) in the SI (total *k*=12). Importantly, *k*=9 studies (*k*=1 in the main text and *k*=8 in the SI) shared design features—including manipulations and dependent measures—and therefore these demonstrate independent replications. Further, we demonstrate the robustness of our effects employing a meta-analysis on the combined data (SI Analysis 4).

Study 1: Generosity and impartiality are perceived differently

Our primary aim in Study 1 was to understand whether generosity and impartiality are perceived differently across 9 dimensions of interest. To do so, we recruited *N*=114 participants and randomly assigned them to read a dictionary definition (adapted from Meriam-Webster.com) of either generosity ("giving or sharing in abundance") or impartiality ("lack of favoritism toward one side or another"; see SI Analysis 5). Participants then responded to the prompt: "Please name at least 3 and up to 10 real-life acts of [generosity/impartiality]" using free-response text boxes. We then edited these for semantic content to yield 50 example behaviors of each virtue (see SI Tables 7 and 8).

Methods

We requested *N*=525 participants from mTurk (who did not participate in Supplemental Study 1), though after screening for repeat IP addresses and mTurk IDs (including only the first entry of either) and filtering participants who accepted the HIT on mTurk but neglected to complete the survey, our final sample was *N*=496 participants (38.6% female, average age=37.0 years). We randomly assigned participants to one of two between-subjects conditions, in which they were asked to rate either generous or impartial behaviors. We presented participants with a randomly selected subset of 10 behaviors (presented in randomized order) from the 50 generated for the respective virtue in SI Analysis 5 (as well as 5 additional experimenter-generated behaviors, see SI Tables 7 and 8). Thus, each behavior was rated by an average of *m*=47 participants. Participants rated each behavior on nine dimensions (presented in randomized order): *moral goodness* ("In your opinion, how morally good is it to do this behavior?"); *potential for anonymity* ("How possible is it for someone engaging in this behavior to be anonymous to the recipient(s) of this behavior?"); *prototypicality as demonstration of virtue* ("How much does this behavior exemplify the virtue of [generosity/impartiality]?"); *potential for ulterior motives* ("How likely is it that someone engaging in this behavior does so for ulterior motives?"); *the extent to which the behavior is indicative of the actor's consistency across situations* ("How likely is it that someone engaging in this behavior acts similarly in other situations?"); as well as four which replicated the method of previous work (Kraft-Todd & Rand, 2019): *descriptive normativity* ("In your opinion, how many people in your community do this behavior when they are in the relevant situation?"); *injunctive normativity* ("In your opinion, how much do people in your community think doing this behavior is what you are supposed to do when you are in the relevant situation?"); *benefit to the recipient* ("In your opinion, how much benefit (in terms of money, time, effort, etc.) does the recipient of this behavior receive?"); and *cost to the actor* ("In your opinion, how much cost (in terms of money, time, effort, etc.) does the person who does this behavior incur?"). To discourage explicit consistency in response, all ratings were completed using unmarked sliding scales (i.e., recorded ratings were not displayed to participants). All scales ranged from 0 to 100, were initialized at the midpoint, and had labels on extreme values (e.g. 0—"very little", 100—"very much"; see SI for complete stimuli and experimental instructions).

Results

We use a generalized structural equation model to fit a multivariate, multilevel mixed-effects model to compare each rating (as the dependent measure) across virtues, with target behavior nested within virtue condition, and estimating covariance for each pair of ratings.

Across all behaviors, we find that generous compared to impartial behaviors are perceived as (in decreasing order of mean difference): costlier to actors (*coeff*=10.60, *z*=12.75, *p*<.001; see Figure 1a), more beneficial to recipients (*coeff*=7.08, *z*=8.51, *p*<.001), greater potential for ulterior motives (*coeff*=5.00, *z*=5.38, *p*<.001), greater potential for anonymity (*coeff*=3.97, *z*=2.39, *p*=.017), less prototypical as a demonstration of the virtue (*coeff*=-2.81, *z*=-4.19, *p*<.001), and less injunctively normative (*coeff*=-2.66, *z*=-2.20, *p*=.028). We find that generous and impartial behaviors are not perceived differently on the dimensions of moral goodness (*coeff*=1.34, *z*=1.45, *p*=.148), descriptive normativity (*coeff*=-1.11, *z*=-.90, *p*=.370), and the extent to which the behavior is indicative of the actor's consistency across situations (*coeff*=-.41, *z*=-.61, *p*=.545). It is worth highlighting that across all participant-generated behaviors, generosity and impartiality were not perceived differently on the dimension of moral goodness, implying that (according to this method), these virtues are seen as equally morally good.

To better understand the correlation structure among ratings, we conduct an exploratory factor analysis with iterated principal factors and varimax (orthogonal) rotation. The analysis yielded three factors explaining 95.7% of the variance. Factor 1 explained 57.3% of the variance and the items with high loadings (>.4) were *cost to the actor* and *potential for ulterior motives.* We labeled Factor 1 "signaling value" because cost to the actor is often discussed in applications of costly signaling theory (Zahavi, 1975) to human behavior (e.g. Jordan, Hoffman, Nowak, et al., 2016), and because ulterior motives undermine actors' intended signaling. Factor 2 explained 26.4% of the variance and we labeled it "normativity" due to high loadings (>.4) by the items: *descriptive normativity* and *injunctive normativity.* Factor 3 explained 11.9% of the variance and we labeled it "morality" due to high loadings (>.4) by the items: *moral goodness* and *prototypicality as demonstration of virtue*. We use a generalized structural equation model to fit a multivariate, multilevel mixed-effects model to compare factor scores (as the dependent measure) across virtues, with target behavior nested within virtue condition, and estimating covariance for each pair of factor scores. Across all behaviors, we find that generous compared to impartial behaviors are perceived as higher on the *signaling value* factor (*coeff*=.31, *z*=12.34, *p*<.001; see Figure 1b) and lower on the *normativity* factor (*coeff*=-.13, *z*=-2.02, *p*=.043), consistent with the item-level analysis above. It is interesting to note that *potential for ulterior motives*—the item closest to our proposed mechanism for virtue discounting—loaded on the same factor as *cost to the actor*, and that generosity was perceived as significantly higher than impartiality on this dimension (using either item-level or factor score analysis).

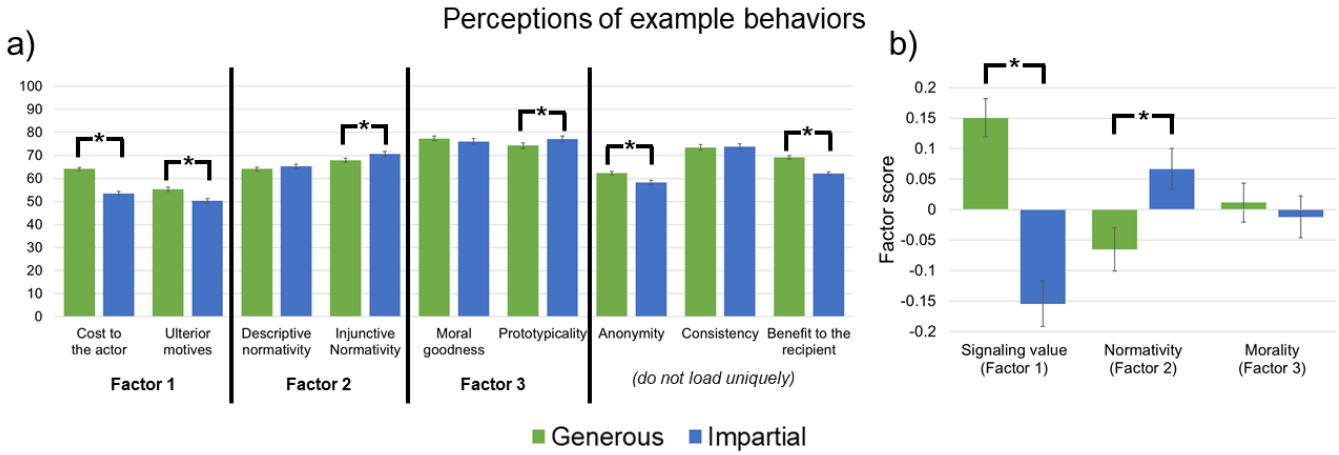

Fig 1. Participant-generated examples of generous and impartial behaviors are perceived differently across many dimensions. Shown are means (with 95% CIs) of ratings (0-100 unmarked slider) across example behaviors of generosity (green) and impartiality (blue) generated by an independent group of participants. Ratings shown by **a)** item and **b)** factor scores. Crucial (i.e., bearing on hypotheses) significant contrasts denoted with (*); *N*=496.

The results of this Study imply that generosity and impartiality—as exemplified by a large number of participant-generated behaviors—are perceived differently along a number of important dimensions. This finding is consistent with prior work conceptually and empirically distinguishing these virtues discussed at greater length in the Introduction (e.g. Shaw, 2013). The result that generosity is perceived as having greater potential for ulterior motives than impartiality is consistent with the direction and proposed mechanism of our differential virtue discounting hypothesis; i.e., that generosity will be discounted more than impartiality, and this will be explained by greater perceptions of ulterior motives for public (compared to private) generosity than public (compared to private) impartiality. We use these ratings to compare the sets of behaviors employed as stimuli in subsequent analyses, highlighting that these include sets of behaviors that are representative of the general differences in perceptions between generosity and impartiality found here (Study 2 and SI Analysis 6) as well as matched across these dimensions (SI Analysis 7).

Study 2: Generosity is discounted but impartiality is not when actors are observable

The purpose of Study 2 is to conceptually replicate previous demonstrations of virtue discounting for generosity (Lin-Healy & Small, 2012; Newman & Cain, 2014), and provide the first investigation of virtue discounting for impartiality. We also extend prior work by investigating observers' inferences of actors' motivations, and use these data to explore potential mechanisms of virtue discounting.

Methods

Here, we use data from 3 online studies (Studies 2a-c, see SI for complete experimental instructions); total *N*=1,201, 46.7% female, average age=35.4 years. We randomly assigned participants to one of 4 between-subjects conditions in a 2 (virtue: generosity vs. impartiality) x 2 (observability: public vs. private) factorial design. After providing consent and entering their mTurk ID, participants read text containing our experimental manipulations. First, they read text

delivering our virtue manipulation: "Some people think that [generosity/impartiality] is a virtue." We then provided a dictionary definition of the virtue adapted from Merriam-Webster.com (e.g. "Generosity usually means giving an abundance of one's money or time"; "Impartiality usually means treating everyone equally and fairly, without bias"). 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 3 behaviors demonstrating [generosity/impartiality] (according to condition). Then, participants read text delivering our observability manipulation: "She did these things in [public/private]; therefore, other people [knew/did not know] that she did them."

We note that the sets of behaviors we employed as stimuli here were representative of the general differences in perceptions between generosity and impartiality found in Study 1 (see Fig. 1). Specifically, consistent with Study 1, between the two sets of behaviors (3 per virtue) we employ as stimuli here, the generous compared to impartial behaviors are perceived as costlier to actors, more beneficial to recipients, less prototypical as a demonstration of the virtue, and less injunctively normative (and that generous and impartial behaviors are not perceived differently on the dimensions of: moral goodness descriptive normativity, and extent to which the behavior is indicative of the actor's consistency across situations). In contrast with Study 1, between the two sets of behaviors we employ as stimuli here, generous compared to the set impartial behaviors employed as stimuli here are not perceived differently on the dimensions of potential for ulterior motives and potential for anonymity.

Following the stimuli, we presented participants with the two primary dependent measures and six secondary dependent measures in randomized order. Primary dependent measures were: *moral goodness* ("How morally good is [the actor]?") and *trait ratings* ("How [generous/impartial] is [the actor]?"). Secondary dependent measures assessed participants' perceptions of the actor's motivation ("How much do you think Jen is motivated to act [generously/impartially]…"): *reputational* ("…because she is trying to make others think she is [generous/impartial]?"); *authentic* ("…because she wants to be [generous/impartial]?"); *norm-signaling* ("…because she wants others to be [generous/impartial], and she is trying to lead by example?"); *moral* ("…because she thinks it is the right thing to do?"); *1st-party benefit* ("How much do you think Jen will personally benefit from behaving this way?"); and *3rd-party benefit* ("How much do you think another person would benefit from interacting with Jen?"). All dependent measures were answered on 100-point unmarked slider scales with extreme anchors labeled (and midpoints labeled for primary dependent measures).

Results

We first use a three-way MANOVA to test for an interaction of our experimental manipulations (virtue and observability) and Study on our primary dependent measures (moral goodness and trait ratings). We do not find a significant interaction of Study with either manipulation nor their interaction (Wilks' lamdas>.99, *p*s>.472), so we use the combined data in further analyses and include Study as a covariate. (We note that Studies 2a and 2b also included a "baseline" observability condition which are excluded in these analyses but are presented in SI Analysis 1).

First, we conduct a multivariate regression analysis to test for evidence of *differential virtue discounting*, i.e., whether the difference between perceptions of public and private displays of virtue is greater for generosity than impartiality. Next, we use exploratory factor analysis to

understand the associations among our secondary dependent measures (motivational attributions). Finally, we employ a structural equation model to test for evidence of our proposed mechanism of differential virtue discounting, i.e., that observers' perceptions of actors' virtue will be explained by observers' inferences about actors' motivations.

As predicted, we find a significant interaction between virtue and observability on trait ratings ($F(1,1195)=17.92$, $p<.001$, $d=.54$), such that public displays of generosity ($m=78.44$, 95% CI [76.65, 80.24]) are rated as demonstrating the virtue significantly less than private displays ($m=87.74$, 95% CI [86.29, 89.19], Scheffe's $t=-6.80$, $p<.001$, $d=.65$), while public and private displays of impartiality are not rated differently (Scheffe's $t=.80$, $p=1.00$, $d=.06$). This result provides support for our *differential virtue discounting* hypothesis: when actors engage in virtuous behaviors in public compared to in private, observers discount the virtue of those behaviors for generosity but not impartiality. We emphasize that we obtain this result employing sets of behaviors for generosity and impartiality that were not rated differently on the dimension of "potential for ulterior motives" (see Study 1), and so the effect cannot be attributed merely to differences between stimuli on this dimension.

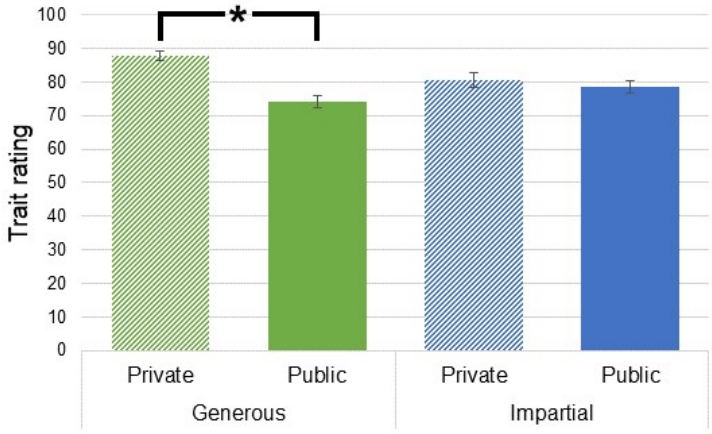

Fig 2. Observability lowers perceived virtue for generosity but not impartiality. Shown are means (with 95% CIs) of trait ratings (0-100 unmarked slider), as a function of whether the actor is said to engage in a set of 3 behaviors that are generous (green) or impartial (blue) and whether this set of behaviors is said to be displayed in public (solid) or private (lines). Crucial (i.e., bearing on hypotheses) significant contrast denoted with (*). From left to right: private generosity $N=300$, public generosity $N=303$, private impartiality $N=298$, public impartiality $N=300$.

Because we only find an effect of observability on virtue judgments for generosity (and not impartiality), we investigate mediation of this effect by our secondary dependent measures only for generosity (see Figure 3). To better understand the correlation structure among our secondary dependent measures (motivational attributions), we conduct a preregistered exploratory factor analysis with varimax (orthogonal) rotation. The analysis yielded two factors explaining 91.9% of the variance. Factor 1 (explaining 54.3% of the variance) we labeled *moral motivation* due to the high loadings (>.5) by the following items: 3rd-party benefit, moral motive, and authentic motive. Factor 2 (explaining 37.6% of the variance) we labeled *reputational motivation* motivations due to high loadings (>.5) by the following items: 1st-party benefit and reputational motive. Conceptually, we believe these factors represent important facets of "selfish motivation" (our proposed mechanism for virtue discounting); we emphasize that they are largely independent (as a result of our orthogonal factor rotation), though weakly related ($r=-.1$,

p<.001). To test for (correlational) mediation of observability on trait virtue judgments by perceived actor motivation, we construct a structural equation model using standardized variables (indirect effects are calculated using the multivariate delta method and standard errors are bootstrapped with 1,000 replications; see SI Analysis 3 for this analysis using all secondary dependent measures rather than factor scores). We first describe correlations in this model, and then the indirect, total, and direct effects in the mediation. Beginning with estimates of model fit, we note that the model accounts for 53.4% of the variance in trait ratings (of generosity).

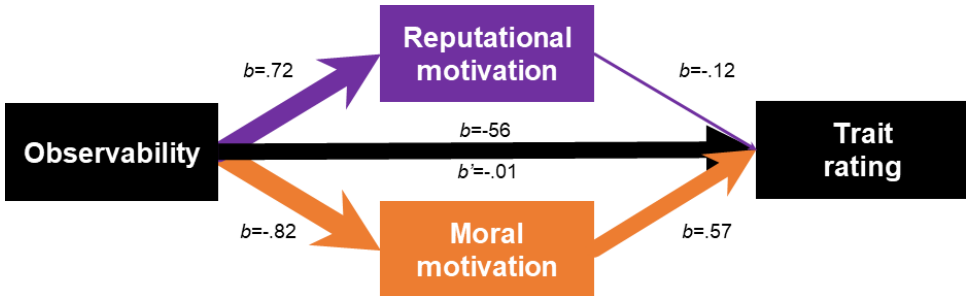

Fig 3. Virtue discounting (of generosity) is explained by observers' inferences that actors are motivated by moral motivation and reputational benefit. The center arrow (black) represents the correlation of the observability and trait ratings (of generosity) with (b) and without (b') motivation attribution factor scores as covariates. From left-to-right, the first set of arrows represents the correlation of motivation attribution factor scores with observability, and the second set of arrows represents the correlation of trait ratings (of generosity) with motivational attribution factor scores (*N*=603). Line thickness represents correlation strength (all variables standardized for this analysis).

First, public displays of generosity are significantly associated with both moral motivation (*b*=-.64, 95% CI [-.80, -.49], *p*<.001) and reputational motivation (*b*=.74, 95% CI [.59, .90], *p*<.001) factor scores. Next, we find that trait generosity ratings are significantly associated with both moral motivation (*b*=.59, 95% CI [.54, .64], *p*<.001) and reputational motivation (*b*=-.12, 95% CI [-.17, -.07], *p*<.001) factor scores. We note that these associations are significantly different in magnitude (i.e. the absolute values of their confidence intervals do not overlap), such that observers' ratings of actors' generosity are more strongly (by about a factor of 5) associated with motivational attributions of moral motivation (positively) compared to their reputational motivation (negatively). Finally, we turn to the mediation results. The total effect of public display on trait ratings of generosity is significant (*b*=-.54, 95% CI [-.68, -.41], *p*<.001), but the direct effect is not (*b'*=-.07, 95% CI [-.18, .03], *p*=.172), implying full mediation (86.2% of the total effect). Calculating indirect effects as percent of total effect mediated, moral motivation motivational attributions account for 70.0% of this mediation while reputational motivation motivational attributions account for 16.7%. Again (and following from previous results), this implies that observers' perceptions of actors' virtue is more strongly influenced (by about a factor of 5) by the degree to which observers infer that actors are motivated by moral motivation than reputational motivation.

In sum, in Study 2 we conceptually replicate previous work demonstrating virtue discounting for generosity. We do not, however, find virtue discounting for impartiality, thus providing the first evidence of *differential virtue discounting*. Further, we provide evidence that observers' motivational attributions may be a mechanism of this effect. Specifically, we see that a 2-factor structure explains most (91.9%) of the variance across 6 motivation attribution items, and that these factors are interpretable as *moral motivation* and *reputational motivation*. Also, these motivational inference factors clearly map onto our proposed mediator of virtue signaling:

actors' selfish motivation; *reputational motivation* is an instantiation of "selfish" motivation whereas *moral motivation* is a type of "selfless' motivation. We note that because we employ varimax (orthogonal) rotation on our factors, they represent somewhat independent dimensions (i.e., they are not two ends of the same spectrum), although they are significantly but weakly negatively correlated (r=-.10, p<.001). Finally, consistent with our proposed mechanism of differential virtue discounting, we show that these factor scores fully mediate the effect of observability on trait virtue judgments (for generosity), and that this mediation is accounted for by moral motivation to a greater extent than by reputational motivation.

To demonstrate independent replications and avert file drawer concerns, we include all additional data collected on these questions in the SI and provide additional analyses as robustness checks and demonstrations of boundary conditions (inclusive of manuscript, total N=6,130). We emphasize one finding consistent with our differential virtue discounting hypothesis, from a meta-analysis of all studies with the 2 (virtue: generosity vs. impartiality) x 2 (observability: public vs. private) factorial design (N=4,804, see SI Analysis 4), that shows greater discounting for generosity than impartiality.

Study 3: Discounting of generosity is explained by motivational inferences

The purpose of Study 3 was to provide stronger causal evidence of our proposed mechanism of differential virtue discounting—that observers' perceptions of actors' virtue will be explained by observers' inferences about actors' motivations—using an experimental mediation design (Imai et al., 2013; Pirlott & MacKinnon, 2016). Rather than rely on participants' inferences about actors' motivations (as well as causal inference through correlational mediation), here we manipulate our proposed mechanism by stipulating actors' motivations explicitly (for previous work implementing this design strategy, see e.g. Kraft-Todd et al., 2018).

To explain the experimental mediation design in greater detail: it involves comparing the effect of the independent variable on the dependent variable to this same effect when the proposed mediator is also manipulated. Here, this means comparing the effect of observability on trait ratings to the effect of observability on trait ratings when actors' motivation (our proposed mechanism) is also manipulated. In other words, we compare the effect of observability in "Study 3a", where we only manipulate observability (public vs. private) to the effect of observability in "Study 3b", where we cross our observability manipulation (public vs. private) with stipulated actor motivation (reputational motivation vs. moral motivation). The logic of this design is that if the effect of the independent variable on the dependent variable is reduced when the proposed mediator is also manipulated (relative to when only the independent variable is manipulated), a valid causal inference can be drawn that the independent variable affects the dependent variable *through* the proposed mediating variable.

Methods

We randomly assigned N=716 participants (43.7% female, average age=38.6 years) to one of 6 between-subjects conditions in a 2 (observability: public vs. private) x 3 (stipulated motivation: reputational motivation vs. moral motivation vs. none) design. The procedure was identical to Study 2, except that we presented additional information following the original stimuli in the stipulated motivation conditions (i.e. all save the "none" condition). In the

[reputational motivation/ moral motivation] conditions, participants read, "She [was/was NOT] motivated to do these things for the following reasons: she wants others to think that she is generous, and she thinks she will benefit from others perceiving her as generous. She [was NOT/was] motivated to do these things for the following reasons: she wants to be generous, she wants to benefit others, she thought it was the right thing to do, she wants others to be generous and she was trying to lead by example." Again, because we did not find evidence of virtue discounting for impartiality (i.e. an effect of our observability manipulation on perceptions of virtue) in Study 2, we investigate only generosity here. Also, because we are manipulating actors' motivations here, we do not include the secondary dependent measures (motivational inferences) administered in Study 2.

This design enables us to test for experimental mediation of the effect of observability on virtue judgments (for generosity) by actor motivation by employing the following logic. We call the two "no stipulated motivation" conditions "Study 3a" and the remaining four conditions "Study 3b". In Study 3a, we expect to replicate the effect of observability on virtue judgments (for generosity) that we find in Study 2 (note that, prior to the administration of secondary dependent measures, these Studies are identical; see SI for complete experimental instructions). Accordingly, our hypothesis here is that public generosity will be perceived as less virtuous than private generosity. In Study 3b, we cross our observability manipulation with a manipulation of our proposed mechanism for the effect of observability: actor motivations (consistent with the findings of our correlational mediation in Study 2). Note that the language we use for our manipulation of actor motivation is the same as we used in the secondary dependent measures employed in Study 2. Our hypothesis is that the coefficient on the effect of observability on virtue judgments in Study 3b will be substantially lower than that in Study 3a. This comparison is made possible because we randomly assigned participants to conditions across Studies 3a and 3b. Thus, though 3a and 3b were run as a single experiment, we believe the experimental mediation method we employ is conceptually clearer when described as two "studies" for which participants were simultaneously recruited and across which random assignment is maintained.

Results

First, we conduct a multivariate regression analysis to test for evidence of *virtue discounting* in the two conditions comprising Study 3a, investigating whether public displays of generosity are perceived as less virtuous than private displays of generosity (replicating the result of Study 2; compare green bars in Figures 2 and 4). As predicted—and replicating the effect we see in Study 2—we find a significant effect of observability on trait ratings of generosity ($F(1,230)= 20.16$, $p<.001$, $d=.59$), such that participants rate public displays ($m=77.96$, 95% CI [75.29, 80.63]) as demonstrating virtue less than private displays ($m=86.19$, 95% CI [83.73, 88.65]).

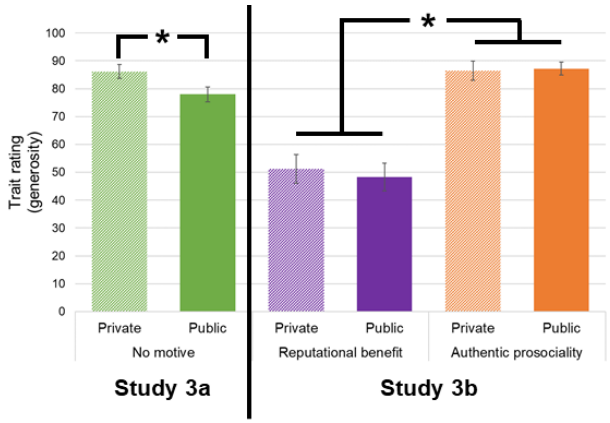

Fig 4. Actor motivation explains the effect of observability on perceived virtue (generosity). Shown are means (with 95% CIs) of trait ratings (0-100 unmarked slider), as a function of whether the actor is said to engage in a set of 3 behaviors that are generous in private (lines) or in public (solid) and (**Study 3a**) whether we do not stipulate actor motivation (green) or (**Study 3b**) that the actor is motivated by reputational motivation (purple) or that the actor is motivated by moral motivation (orange). Crucial (i.e., bearing on hypotheses) significant contrasts denoted with (*). From left to right: private no motive *N*=116, public no motive *N*=116, private reputational motivation *N*=109, public reputational motivation *N*=114, private moral motivation *N*=112, public moral motivation *N*=115. Note here that observability does not have an effect on trait ratings of generosity when actors' motivations are stipulated (i.e. orange and purple bars; in contrast to green bars here and in Figure 2).

Then, we conduct a multivariate regression analysis in the four conditions comprising Study 3b to test for experimental mediation of the effect of observability on virtue judgments (for generosity) by actor motivation, i.e. whether the effect of observability on perceived generosity is reduced when we stipulate actor motivation. As predicted, we find a significant effect of actor motivation on perceived generosity ($F(1,446)=317.84$, $p<.001$, $d=1.69$), such that participants rate actors who we describe as being motivated by reputational motivation ($m=79.17$, 95% CI [76.66, 81.68]) to demonstrate generosity significantly less than actors who we describe as being motivated by moral motivation ($m=84.98$, 95% CI [82.79, 87.18]). Here, we do not detect a significant effect of observability on trait ratings ($F(1,446)=.25$, $p=.621$, $d=.05$), nor a significant interaction between observability and actor motivation on trait ratings ($F(1,446)=.80$, $p=.371$, $d=.08$). Crucial to the test of experimental mediation, we note that the coefficient on observability is 110% smaller in Study 3b than in Study 3a (it is 10% as large, but in the opposite direction, although the 95% CI includes zero), providing causal evidence that actor motivation mediates the effect of observability on virtue judgments (for generosity).

In sum, in Study 3a we first replicate our result from Study 2 demonstrating *virtue discounting* for generosity. Crucially, we provide causal evidence supporting our proposed mechanism for virtue discounting in Study 3b, showing that actor motivation explains the effect of observability on perceptions of trait virtue (for generosity), consistent with our correlational mediation result from Study 2. Specifically, when observers think that actors are prosocially- versus reputationally-motivated, they perceive actors as *more* virtuous (generous) regardless of whether actors engaged in generous behavior in public or in private. Similarly, when observers think that actors are motivated by reputational motivation and not moral motivation, they perceive actors as *less* virtuous (generous) regardless of whether actors engaged in generous behavior in public or in private.

General Discussion

In the Studies presented here (N=2,413), we show that: the virtues of generosity and impartiality—understood as an aggregation of 50 participant-generated behaviors—are perceived differently across many (6 of 9 measured) dimensions; generosity is discounted but impartiality is not when actors' behavior is observable (i.e. *differential virtue discounting*); and discounting of generosity is explained by observers inferring that actors have selfish motivations for their public displays of virtue (with suggestive evidence that lower *moral* motivations explain this effect to a greater extent than higher *reputational* motivations). In sum, our findings contribute to a broad program of research that asks 3 related questions: 1) How can virtues be distinguished? 2) Are different virtues discounted to different extents (i.e. as with accusations of "virtue signaling" or "performative virtue")? 3) What are the mechanisms of virtue discounting across virtues? Below, we summarize how our results bear on each of these questions and represent contributions to the literature, and then we discuss limitations of our work as well as clear directions for future work that ours suggests.

Additionally, our investigation speaks to two important paradoxes for moral psychology. First, people want others to know about their virtuous behavior, but if they engage in virtuous behavior so that others can see it, observers may discount actors' virtue (the "virtue signaling paradox"). Second, people can engage in costly virtuous behavior to provide a more honest signal of their moral character, but if they engage in costly virtuous behavior so that others can see it, observers may be even more likely to discount actors' virtue (the "costly signaling of virtue paradox"). We also discuss how our findings (and potential future work) might shed light on these.

Summary and contributions to the literature

How can virtues be distinguished? In short, our work expands the toolkit of methods available to differentiate virtues (Peterson & Seligman, 2004), as well as providing additional evidence distinguishing generosity and impartiality in particular (e.g. Shaw & Olson, 2012). As we review in the Introduction, prior work has argued that generosity and impartiality can be distinguished conceptually (Shaw, 2013) and with regard to their evolutionary functions (Shaw, 2016), and empirical work has shown that they can be distinctly operationalized in behavioral paradigms (e.g. Shaw et al., 2018). To this literature, we add a novel approach, aggregating participants' ratings across 9 dimensions of an independent sample's suggestions for exemplary behaviors of each virtue (for previous work demonstrating this approach, see e.g. Kraft-Todd & Rand, 2019). We find overall, that participants rated generosity and impartiality differently on 6 of 9 dimensions. Specifically, participants rated generous behaviors more highly than impartial behaviors on the dimensions of: cost to actors, benefit to recipients, potential for anonymity (in the actor/recipient relationship), and potential for ulterior motives. Participants rated impartial behaviors more highly than generous behaviors on the dimensions of: injunctive normativity (the degree to which they think others think people should do these behaviors) and how prototypical the behavior is as a demonstration of the virtue. We believe that our method of assessing virtue via concrete examples avoids complicating factors such as assuming common knowledge of their definition, participants reading comprehension and memory for definitions we provide, etc. Further, by gathering independent ratings of our example behaviors on various dimensions of interest, we gain a finer-grained understanding of differences in how the virtues are perceived, as well as insight into potential mechanisms for other observed differences between virtues.

Are different virtues discounted to different extents (i.e. as with accusations of "virtue signaling" or "performative virtue")? Consistent with much prior work (e.g. De Freitas et al., 2019), we find that public generosity is rated less favorably than private generosity. We are unaware of similar investigations for the virtue of impartiality, and so our investigation represents a novel extension of this literature. We find that while observability affects perceptions of generosity, it does not affect perceptions of impartiality (i.e. *differential virtue discounting*). Therefore, the "virtue signaling paradox" would seem to apply to generosity, but may not apply to impartiality, because our results suggest that observable impartiality is not discounted. Further, we provide additional evidence (two independent replications, N=1,200; see SI Analysis 1) that it is specifically the disclosure of the information that virtuous behavior is performed in public that leads to discounting—again, in generosity but not impartiality—compared to a "baseline" condition where no information is provided regarding observability (i.e. actors' virtue in the public condition is rated significantly lower than in the private and baseline conditions, and the latter are not different from each other, for generosity, whereas the observability manipulation again has no effect for actors demonstrating impartiality). We provide further evidence (six independent replications, N=2,396; see SI Analysis 6) that both generosity and impartiality are discounted—although generosity is discounted to a greater extent—when our observability manipulation is conflated with a manipulation of actors' motivations (in these experiments, the experimental conditions are public + reputational motivation vs. private + moral motivation).

What are the mechanisms of virtue discounting across virtues? As we discuss in the Introduction, much previous work demonstrates that observers' perceptions of actors' ulterior motives lead to devaluing generous behavior (e.g. Newman & Cain, 2014). We replicate this finding using both correlational (Study 2) and experimental mediation (Study 3) designs, and extend it by shedding light on the structure of the underlying motivational inferences. Specifically, our exploratory factor analysis reveals two relatively independent (r=-.1, p<.001) factors of motivational inferences among our six motivational inference items: *reputational motivation* and *moral motivation*. Interestingly, our correlational mediation design (Study 2) provides suggestive evidence that decreased attributions of *moral motivation* explain more of the virtue discounting effect than increased attributions of *reputational motivation*. For actors confronting the "virtue signaling paradox", this might suggest that their efforts would be better spent highlighting their moral motivations to observers rather than minimizing the potential for observers to infer reputational motivations. More broadly, our results suggest that actors confronting this paradox may be better able to demonstrate their virtue without observers discounting it if actors make the motivational context clearer (i.e. honestly represent their motivations). We also provide a more fine-grained correlational mediation analysis—i.e. using each of the six motivational inference items rather than the two factor scores as mediators—shedding additional light on this question (see SI Analysis 3).

Theoretical implications

In the Introduction, we motivated our differential discounting hypothesis (i.e. that generosity would be discounted more than impartiality) with the intuition that, faced with an exemplar of generosity and impartiality, observers might be more eager to interact with the former because they might stand to benefit more. This logic derives from the theory of *partner choice* (Barclay, 2013; Noë & Hammerstein, 1994), which describes that individuals compete to display desirable traits so they are chosen as interaction partners for potentially mutually

beneficial cooperative interactions. Accordingly, one might hypothesize that generosity is a more attractive trait than impartiality among potential interaction partners. Alternatively, one might seek to explain these results with the theory of *indirect reciprocity* (Nowak & Sigmund, 2005), which holds that we gain a good reputation by following social norms and cooperating with others who do the same (Ohtsuki & Iwasa, 2006). Given our findings, one might correspondingly hypothesize that there is a stronger social norm for generosity than impartiality. It would not be surprising if the attractiveness of interaction partner traits and the strength of social norms were correlated, and thus both theories might help explain differential virtue discounting. Further, it is likely that there are situational moderators of this association (and their correlation), including the relational obligations of actors, recipients, and observers (McManus et al., 2020). The cultural context, across which social norms vary, would also likely play a large role by either account. For example, different cultures are likely to value different virtues to different degrees, affecting both the traits that are seen as indicative of good interaction partners (as per partner choice) as well as the norms dictating reputation (as per indirect reciprocity). Thus, variation in the virtues that are valued across cultures are likely to dictate *which* virtues are subject to the "virtue signaling paradox." Partner choice and indirect reciprocity are neither mutually exclusive nor exhaustive, and future work may dissociate their contributions to differential virtue discounting across generosity and impartiality, as well as between other virtues.

In Study 1, we named one factor of participants' ratings of our 9 dimensions "signaling value" in reference to the evolutionary theory of *costly signaling* (Zahavi, 1975). This theory has been extensively applied to human cooperative behavior (Barclay & Willer, 2007; Bliege Bird et al., 2001; Gintis et al., 2001; Jordan, Hoffman, Nowak, et al., 2016) and is at the heart of the "costly signaling of virtue" paradox. The problem this theory addresses (particularly with regard to human behavior) is: how can observers make inferences about a trait that is difficult to directly observe (i.e. actors' virtue)? In brief, the crux of the solution is that if there are observable phenomena that are sufficiently costly such that high-quality (i.e. virtuous) individuals can produce them but low-quality (i.e. non-virtuous) individuals cannot (technically, creating a "separating equilibrium"), these observable phenomena can be an honest signal of the individual's (difficult to observe) traits. We might therefore conclude that people can convince others of their virtue if they engage in costly virtuous behavior observable by others. As we show, however, generous behavior is *both* costlier to actors *and* discounted more than impartiality. It may help shed light on this instantiation of the "costly signaling of virtue" paradox to consider its relation to the human capacity for theory of mind, i.e. the ability to reason about others' mental states (Premack & Woodruff, 1978; Saxe et al., 2004). Theory of mind includes individuals' reasoning about others' motivations, which in turn affects our moral judgments (Young et al., 2007). Applying theory of mind iteratively (e.g. Shum et al., 2019)—which has been shown to impact moral judgment and prosocial behavior (Kleiman-Weiner et al., 2017)—leads to complicated and strategic considerations of virtue display, as the "costly signaling of virtue" paradox implies. Consider the mental inference arms race in this context: suppose an actor could know that engaging in costly virtuous behavior will cause observers to think that their motivation is selfless; actors could then strategically engage in such behavior to reap reputational rewards. Further, observers might anticipate that actors could engage in this strategy, and therefore discount costly signals when they are observable (Barclay & Willer, 2007). In turn, actors might anticipate that observers will engage in such discounting, and seek to avoid the appearance of considering the costliness of virtuous behavior in order to counteract it (Jordan, Hoffman, Nowak, et al., 2016).

Limitations and future directions

Despite its merits, our investigation does not provide comprehensive answers to the three major questions we have so far considered. We discuss the limitations to our work and possible contributions of future work again organized by the questions: 1) How can virtues be distinguished? 2) Are different virtues discounted to different extents (i.e. as with accusations of "virtue signaling" or "performative virtue")? and 3) What are the mechanisms of virtue discounting across virtues?

How can virtues be distinguished? A necessary first step to the primary focus of the present work—documenting *differential virtue discounting*—was to provide evidence that the virtues (generosity and impartiality) under investigation are distinct. We believe the 9 dimensions we employed as dependent measures in Study 1 grant valuable and novel insight in this regard, though we do not claim that the dimensions we measure are exhaustive. Future work might consider a broader array of underlying dimensions that are capable of distinguishing virtues. We also want to clarify that the differences we observe across these dimensions represent mean ratings across all 50 participant-generated examples of each virtue; it is not the case that this pattern holds when investigating any pair of examples. In other words, although (for example), generous behaviors on the whole were rated as costlier to actors than impartial behaviors, there are numerous examples of specific impartiality behaviors that were rated as costlier to actors than specific generous behaviors. Also, it could be the case that our method of aggregating ratings across specific behaviors to make general claims about the virtue misses complexities in understanding how each virtue is broadly perceived. For example, with regard to our "consistency across situations" dimension, it could be that participants would respond qualitatively differently to the "impartiality" construct on its own (e.g. rating it higher) compared with aggregated ratings across impartial behaviors (e.g. "drawing names from a hat for a project at work", one behavior in our stimulus set). Further, among the 9 dimensions we measure, it could be the case that further refinement is warranted. For example, for both the "cost to actors" and "benefit to recipients" ratings, we asked for ratings "(in terms of money, time, effort, etc.)." Yet, it might be reasonable to think that, e.g. "money" and "effort" are meaningfully different aspects of these dimensions worthy of differentiating (Soman, 2001; Whillans et al., 2017). Additionally, although descriptive and injunctive normativity are two of the dimensions we employ, future research might also explore a relevant but subtly different concept to further delineate differences among virtues: the violation of expectations. That is, virtues might also be distinguished by exemplary behaviors (in certain situations, for certain actors, etc.) that are sufficiently normative that *not* doing them violates a norm. There might also be broader differences in how virtues are perceived; in particular, the extent to which they are considered on a continuous or binary scale. Such differences have been observed in research on moral foundations theory (Graham et al., 2011), for example showing that people often think of "purity" in a more dichotomous manner than "harm" (Rottman & Young, 2019).

Also, we do not intend our approach—or the differences we observe in aggregate ratings of example behaviors—to imply that generosity and impartiality (nor other virtues) are neatly dissociable. There are likely countless examples of these virtues being instantiated, and even in tension, in the same behavior: consider nepotism, a class of behaviors that could be interpreted as simultaneously generous and partial. Finally, previous work has attempted to distinguish virtues by other means (Peterson & Seligman, 2004), while other lines of research have similarly attempted to differentiate the space of related abstract concepts such as morality (e.g. into

"domains"; Graham et al., 2011). Future work specifically interested in dissociating virtues might weigh the method we demonstrate here with other relevant methods such as these, and apply them to studying virtues beyond generosity and impartiality (such as trustworthiness, compassion, etc.).

Are different virtues discounted to different extents (i.e. as with accusations of "virtue signaling" or "performative virtue")? Our intention in the work presented here was to understand, broadly, how participants respond to public vs. private displays of generosity and impartiality. Our example behaviors were therefore devoid of situational context, which is often an important input for our moral judgment (e.g. as discussed in work on impartiality; Shaw, 2016). For example, people might reasonably perceive "giving a waiter a large tip" (one behavior in our stimuli) differently depending on whether the meal was a date, according to local tipping norms, whether the waiter gave a free dessert etc. Future work might consider the interaction of observability with key aspects of situational context to gain a richer and more ecologically valid of these phenomena. For example, the relational context of the actor and the perceiver is likely to affect ratings like the ones we employ. Our studies use perceivers who are 3rd-party observers, though we would not be surprised if different patterns of results were found in a study of perceivers as 2nd-party recipients (Gummerum & Chu, 2014). To further understand real-life heterogeneity in these phenomena, future work might also explore participant-level individual differences in virtue discounting. That is, people may vary in the extent to which they discount some virtues more than others when displayed in public.

What are the mechanisms of virtue discounting across virtues? Although we provide evidence that observers' motivational inferences are one proximate mechanism of virtue discounting, we do not claim that this is the *only* mechanism of virtue discounting. For example, it could be the case that differences in perceptions of the underlying dimensions that differentiate virtues (see Study 1) provide a "secondary mechanism" that explains differences in observers' motivational inferences. We provide evidence (two independent replications, N=806; see SI Analysis 7) that when sets of example behaviors are matched on these dimensions across generosity and impartiality, neither is discounted. This result suggests that the 4 dimensions across which our generosity and impartiality stimuli differed (cost to actors, benefit to recipients, prototypicality as a demonstration of the virtue, and injunctively normativity; see Study 2) may contribute to observers' motivational inferences, that in turn affect virtue discounting. A more comprehensive exploration of these dimensions could both contribute to our understanding of differences among virtues as well as providing greater mechanistic detail to our account of virtue discounting.

Further, although we included six motivational inference questions (see Study 2) that we thought were most relevant to this context, we do not claim that this is a comprehensive investigation of motivation (and we note that, for concision, we present the correlational mediation analysis in Study 2 using motivation factor scores, though we also replicate this analysis using each of our motivational inference items as mediators; see SI Analysis 3). Motivation (like virtue) is a multidimensional construct (e.g. Reiss & Havercamp, 1998), and so future work might more broadly explore the space of motivations to better understand how observers' inferences of actors' motivations affect virtue discounting. In particular, we anticipated that "norm-signaling" ("…because she wants others to be [generous/impartial], and she is trying to lead by example?") might be an important motivation for actors to engage in public virtue. We were surprised to find that observability did not affect participants' perceptions of this motivation, though this null result may have been due to idiosyncrasies of our stimuli or

item wording (e.g. our operationalization was not intuitively comprehensible). In popular discussion of real-world phenomena relevant to "virtue signaling" (e.g. expressions of support for social justice issues on social media), this "norm-signaling" idea is often evoked in defense of public displays of virtue. Therefore, we believe this is an important construct for future research to explore. One important construct that we neglected to include among our motivation items regards "self-signaling" (Bodner & Prelec, 2003). Our motivational inference questions concerned the effect actors wanted to have on observers, but the literature on self-signaling demonstrates that often actors are motivated by the effect they have on *themselves*, e.g. for affirming their identity (Mazar et al., 2008). This motivation may be particularly relevant to the class of behaviors on social media pejoratively described as "virtue signaling", though it is likely to also impact such in-person behaviors that we explore here. We believe that related future work—especially regarding displays of virtue on social media—should incorporate this construct.

More broadly, our approach to studying these phenomena—in keeping with colloquial use of "virtue signaling"—was to investigate observers' perceptions of actors' behavior. It is also interesting to consider whether actors are, in fact, virtue signaling, i.e., whether they are motivated to engage in public virtue in order to demonstrate their virtue to observers (as the theoretical accounts discussed above imply). Shedding light on these phenomena from both observers' as well as actors' perspectives would therefore help us gain a more comprehensive understanding of relevant psychological processes.

Conclusion

Oscar Wilde beautifully articulates the virtue signaling paradox: "The nicest feeling in the world is to do a good deed anonymously—and have somebody find out." We all want to be seen as virtuous. As we have shown here—and consistent with prior work—publicly displaying our virtue (here: generosity) may result in observers believing that we are selfishly motivated. Yet, we show that not all virtues (here: impartiality) are subject to such virtue discounting. Further, our results suggest that the key to untangling this paradox may be for others to believe that beneath our public virtue, our intentions are pure (i.e. our motivations are not selfish). Perhaps anonymity in charitable giving (and other domains of virtue) could become obsolete if we can find a way to make our motivations as observable as our virtue.

Context

There is extensive evidence that people care about their reputations (Emler, 1990), engage in strategic self-presentation to manage others' impressions of them (Jones & Pittman, 1982), and are more likely to behave prosocially when observed (Kraft-Todd et al., 2015). In the past few years, increasing attention has been paid to phenomenon of "virtue signaling" (Bartholomew, 2015), i.e. conspicuous, public displays of admirable moral behavior, particularly on social media. This is a new instantiation of a paradox long debated in philosophical and spiritual traditions (Maimonides, 1170): if others are to think well of you, they must be able to observe your virtuous behavior; yet, if others think you have made your virtuous behavior observable to this end, they will not in fact think well of you. Is this a paradox for all virtues? We address this question by comparing the virtue most often entangled by this paradox, *generosity* (e.g. De Freitas et al., 2019), with one recently differentiated from it in the literature, *impartiality* (Shaw, 2013). We introduce terminology—such as "virtue discounting"—to organize relevant literature, demonstrate a novel method for differentiating virtues, and provide evidence that a key mechanism of this phenomenon is observers inferring actors' selfish motivations.

Acknowledgments

This research was made possible by funding by the John Templeton Foundation, The Virtue Project at Boston College, and NSF award #1627157. We would like to thank the Morality Lab at Boston College and the reviewers for their feedback.

Author Contributions

All authors developed the study concept and contributed to the study design. Data collection and Study were performed by G. T. Kraft-Todd. G. T. Kraft-Todd performed the interpretation under the supervision of L. Young. G. T. Kraft-Todd drafted the manuscript, and M. Kleiman-Weiner and L. Young provided critical revisions. All authors approved the final version of the manuscript for submission.

Declaration of Conflicting Interests

The authors declare that there were no conflicts of interest with regard to the authorship or the publication of this article.

References

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. *Nicomachean Ethics*. Hackett Publishing Company, Inc.

Bai, F. (2017). Beyond Dominance and Competence: A Moral Virtue Theory of Status Attainment. *Personality and Social Psychology Review*, *21*(3), 203–227. https://doi.org/10.1177/1088868316649297

Barclay, P. (2013). Strategies for cooperation in biological markets, especially for humans. *Evolution and Human Behavior*, *34*(3), 164–175.

Barclay, P., & Willer, R. (2007). Partner choice creates competitive altruism in humans. *Proceedings of the Royal Society B: Biological Sciences*, *274*(1610), 749–753. https://doi.org/10.1098/rspb.2006.0209

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

Bereczkei, T., Birkas, B., & Kerekes, Z. (2010). Altruism towards strangers in need: Costly signaling in an industrial society. *Evolution and Human Behavior*, *31*(2), 95–103. psyh. https://doi.org/10.1016/j.evolhumbehav.2009.07.004

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

Bliege Bird, R. L., Smith, E. A., & Bird, D. W. (2001). The hunting handicap: Costly signaling in human foraging strategies. *Behavioral Ecology and Sociobiology*, *50*(1), 9–19.

Bodner, R., & Prelec, D. (2003). Self-signaling and diagnostic utility in everyday decision making. *The Psychology of Economic Decisions*, *1*, 105–126.

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

Emler, N. (1990). A social psychology of reputation. *European Review of Social Psychology*, *1*(1), 171–193.

Fehr, E., Bernhard, H., & Rockenbach, B. (2008). Egalitarianism in young children. *Nature*, *454*(7208), 1079–1083. https://doi.org/10.1038/nature07155

Frederick, S., Loewenstein, G., & O'Donoghue, T. (2002). Time discounting and time preference: A critical review. *Journal of Economic Literature*, *40*(2), 351–401.

Gintis, H., Smith, E. A., & Bowles, S. (2001). Costly signaling and cooperation. *Journal of Theoretical Biology*, *213*(1), 103–119.

Grace, D., & Griffin, D. (2006). Exploring conspicuousness in the context of donation behaviour. *International Journal of Nonprofit and Voluntary Sector Marketing*, *11*, 147–154.

Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., & Ditto, P. H. (2011). Mapping the moral domain. *Journal of Personality and Social Psychology*, *101*(2), 366.

Griskevicius, V., Ackerman, J. M., Cantú, S. M., Delton, A. W., Robertson, T. E., Simpson, J. A., Thompson, M. E., & Tybur, J. M. (2013). When the Economy Falters, Do People Spend or Save? Responses to Resource Scarcity Depend on Childhood Environments. *Psychological Science*, *24*(2), 197–205. https://doi.org/10.1177/0956797612451471

Gummerum, M., & Chu, M. T. (2014). Outcomes and intentions in children's, adolescents', and adults' second- and third-party punishment behavior. *Cognition*, *133*(1), 97–103. https://doi.org/10.1016/j.cognition.2014.06.001

Gurven, M., Allen-Arave, W., Hill, K., & Hurtado, M. (2000). "It's a Wonderful Life": Signaling generosity among the Ache of Paraguay. *Evolution and Human Behavior*, *21*(4), 263–282. https://doi.org/10.1016/S1090-5138(00)00032-5

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

Henrich, J., & Gil-White, F. J. (2001). The evolution of prestige: Freely conferred deference as a mechanism for enhancing the benefits of cultural transmission. *Evolution and Human Behavior*, *22*(3), 165–196. https://doi.org/10.1016/S1090-5138(00)00071-4

Imai, K., Tingley, D., & Yamamoto, T. (2013). Experimental designs for identifying causal mechanisms. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *176*(1), 5–51. bth. https://doi.org/10.1111/j.1467-985X.2012.01032.x

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

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

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., & 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

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

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.

Marmot, M. G., Shipley, M. J., & Rose, G. (1984). Inequalities in death—Specific explanations of a general pattern? *The Lancet*, *323*(8384), 1003–1006.

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.

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

Milinski, M., Semmann, D., & Krambeck, H.-J. (2002). Reputation helps solve the 'tragedy of the commons.' *Nature*, *415*(6870), 424–426.

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

Noë, R., & Hammerstein, P. (1994). Biological markets: Supply and demand determine the effect of partner choice in cooperation, mutualism and mating. *Behavioral Ecology and Sociobiology*, *35*(1), 1–11.

Nowak, M. A., & Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature*, *437*(7063), 1291–1298. https://doi.org/10.1038/nature04131

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

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

Pirlott, A. G., & MacKinnon, D. P. (2016). Design approaches to experimental mediation. *Journal of Experimental Social Psychology*, *66*, 29–38. https://doi.org/10.1016/j.jesp.2015.09.012

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behav Brain Sci*, *1*(04), 515–526.

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

Rottman, J., & Young, L. (2019). Specks of Dirt and Tons of Pain: Dosage Distinguishes Impurity From Harm. *Psychological Science*, *30*(8), 1151–1160. https://doi.org/10.1177/0956797619855382

Saxe, R., Carey, S., & Kanwisher, N. (2004). Understanding other minds: Linking developmental psychology and functional neuroimaging. *Annual Review of Psychology*, *55*, 87–124. aph. https://doi.org/10.1146/annurev.psych.55.090902.142044

Shamosh, N. A., DeYoung, C. G., Green, A. E., Reis, D. L., Johnson, M. R., Conway, A. R. A., Engle, R. W., Braver, T. S., & Gray, J. R. (2008). Individual Differences in Delay Discounting: Relation to Intelligence, Working Memory, and Anterior Prefrontal Cortex. *Psychological Science*, *19*(9), 904–911. https://doi.org/10.1111/j.1467-9280.2008.02175.x

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., Choshen-Hillel, S., & Caruso, E. M. (2018). Being biased against friends to appear unbiased. *Journal of Experimental Social Psychology*, *78*, 104–115. https://doi.org/10.1016/j.jesp.2018.05.009

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

Shum, M., Kleiman-Weiner, M., Littman, M. L., & Tenenbaum, J. B. (2019). Theory of Minds: Understanding Behavior in Groups through Inverse Planning. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(01), 6163–6170. https://doi.org/10.1609/aaai.v33i01.33016163

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

Soman, D. (2001). The mental accounting of sunk time costs: Why time is not like money. *Journal of Behavioral Decision Making*, *14*(3), 169–185. https://doi.org/10.1002/bdm.370

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

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

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

Whillans, A. V., Dunn, E. W., Smeets, P., Bekkers, R., & Norton, M. I. (2017). Buying time promotes happiness. *Proceedings of the National Academy of Sciences*, *114*(32), 8523–8527. https://doi.org/10.1073/pnas.1706541114

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

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

Young, L., Cushman, F., Hauser, M., & Saxe, R. (2007). The neural basis of the interaction between theory of mind and moral judgment. *Proceedings of the National Academy of Sciences*, *104*(20), 8235–8240. https://doi.org/10.1073/pnas.0701408104

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