Routledge
Taylor & Francis Group

The Journal of Positive Psychology

Dedicated to furthering research and promoting good practice

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/rpos20

Assessing and dissociating virtues from the 'bottom up': A case study of generosity vs. fairness

Gordon T. Kraft-Todd, Max Kleiman-Weiner & Liane Young

To cite this article: Gordon T. Kraft-Todd, Max Kleiman-Weiner & Liane Young (2022): Assessing and dissociating virtues from the 'bottom up': A case study of generosity vs. fairness, The Journal of Positive Psychology, DOI: 10.1080/17439760.2022.2154254

To link to this article: https://doi.org/10.1080/17439760.2022.2154254

THE JOURNAL OF POSITIVE PSYCHOLOGY
https://doi.org/10.1080/17439760.2022.2154254

Routledge
Taylor & Francis Group

RESEARCH ARTICLE

Assessing and dissociating virtues from the 'bottom up': A case study of generosity vs. fairness

Gordon T. Kraft-Todd [a], Max Kleiman-Weiner [b] and Liane Young [a]

[a]Department of Psychology and Neuroscience, Boston College, Chestnut Hill, MA, USA; [b]Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA

ABSTRACT

The empirical study of virtue is plagued by imprecise definitions and assessment. Here we propose a three-stage, data-driven ('bottom-up') method to differentiate lay perceptions of virtues. Employing two virtues – *generosity* (as cooperation) and *fairness* (as impartiality) – as a case study, we present findings utilizing data from three studies (total $N = 2{,}667$). First, natural language processing of free-response data indicated that participants used different 'topics' (i.e. clusters of words) to describe behaviours representing generosity (topics: 'charity' and 'kindness') and fairness ('equality'). Second, participants in a survey experiment rated behaviours expressing generosity and fairness differently across 6 out of 9 underlying features measured. Third, participants perceive that actors in vignette-based experiments engaging in behaviours expressing generosity versus fairness were motivated differently on 5 out of 6 motivations measured. Our findings support the distinction of the virtues of generosity (as cooperation) and fairness (as impartiality) and indicate the utility of our bottom-up method for assessing and distinguishing virtues.

ARTICLE HISTORY
Received 3 May 2022
Accepted 21 September 2022

KEYWORDS
Virtue; moral; generosity; cooperation; fairness; impartiality; structural topic modelling (STM); bottom-up; data-driven

Introduction

Virtue, among the most ancient and ubiquitous topics of recorded history (as summarized, e.g., by Dahlsgaard et al., 2005), has only become the focus of serious empirical inquiry in the past few decades (see SOM Section 1 for a brief overview of virtue taxonomies proposed by ancient spiritual and philosophical traditions as well as modern empirical approaches). Following recent empirical approaches to virtue (Fowers et al., 2021), we define a *virtue* as follows: a quality of individuals valued by their culture and expressed through a stable pattern of properly motivated behaviour. Conceptions of *virtue* and lists of *virtues* vary over time and across cultures (see a more detailed discussion of this in MacIntyre, 1981, Chapter 16), and given that humans cognitively represent many concepts taxonomically (Osherson et al., 1990), it might seem appropriate to investigate the taxonomic structure of virtue. Yet, reliable empirical dissociations among virtues in existing taxonomies remain rare, evidenced most prominently by the failure of numerous re-analyses and replication attempts (summarized in McGrath, 2014) to recover the structure of virtues proposed by the highly cited 'Values in Action' model (Peterson & Seligman, 2004).

The prospect of reliably dissociating virtues may be hindered by relying exclusively on the predominant 'top-down', or *a priori*, approach to studying virtue (for extended discussion on this and other methodological issues, see Inbar, 2018). Progress may be facilitated by additionally approaching the study of virtue from the 'bottom up', i.e. leveraging data-driven methods facilitated by new technologies, for example, for inexpensive online data collection (Arechar et al., 2017) and the application of machine learning algorithms to large corpuses of text (Roberts et al., 2014). The utility of such approaches has been indicated in many psychological domains (e.g., emotion; Cowen & Keltner, 2017), including recently in the study of virtue (Gulliford et al., 2021). We emphasize that (at least) two aspects of virtue – as it is represented in the minds of the general public – can be clarified from incorporating such a bottom-up approach: the definition of distinct virtues (the focus of Gulliford et al., 2021), and the assessment of distinct virtues for empirical research (our primary focus in the present work). Here we propose a novel, bottom-up method for assessing and distinguishing virtues, employing *generosity* (as cooperation) and *fairness* (as impartiality) as an exploratory case study among virtues.

CONTACT Gordon T. Kraft-Todd gordon.kraft-todd@bc.edu Department of Psychology and Neuroscience, Boston College , McGuinn 300, 140 Commonwealth Ave, Chestnut Hill, MA 02467

We ascribe the notion that lay virtue concepts are likely to be 'fuzzy' (Zadeh, 1965), i.e. that there will be at least some overlap in how people perceive virtues. We do not propose a clear threshold of differences between virtue concepts that definitely designate them 'distinct.' Instead, we are interested in documenting the degree of 'fuzziness', i.e. to what extent different virtues elicit differentiable judgments and behaviours. We believe this is an important first step in virtue research because it would be impractical to study (purportedly) different virtues without evidence that there is some dimension on which they are perceived differently. We do not intend this method to be definitive, nor our investigation to be comprehensive. Rather, we intend our exploratory case study to suggest the utility of our social psychologically motivated and multi-method approach, and hope that it serves as a foundation for future virtue research.

Generosity and fairness: a case study of dissociable virtues

We believe that generosity and fairness are strong candidates to develop a method for dissociating virtues. They are among the earliest and most commonly cited virtues (e.g., Fowers, 2014), and they also emerge as predominant themes within quantitative analyses of virtues (McGrath, 2015). Second, generosity and impartiality are the focus of a recent line of work that provides both theoretical rationale and empirical evidence suggesting their dissociation, echoing distinctions made by virtue ethicists (e.g., Schneewind, 1990). Before reviewing these arguments and this evidence, we clarify the scope of our investigation within our framework of virtues as 'fuzzy' concepts.

The virtue of *generosity* is typically understood in a broad sense, as being concerned with the benefit of one's actions to others (Swanton, 2003). Generosity (or as it is sometimes translated, 'liberality') has been recognized as a virtue since some of the earliest writings on virtue (Aristotle, 1999). Yet, it is often discussed in terms similar to the virtue of 'benevolence' (see SOM Section 1), which can be understood as an umbrella concept encapsulating generosity as well as other virtues that are 'allocentric' (i.e. concerned with 'intelligent caring about people'; e.g., compassion; Gulliford & Roberts, 2018). We concentrate on the subdomain of *generosity* described by game theorists' notion of *cooperation* (Rand & Nowak, 2013), defining generosity (as cooperation) in our studies as, 'trait willingness to confer benefits to others at cost to oneself.' We acknowledge that not all *cooperation* is *generous*, as when reciprocity drives cooperation (Trivers, 1971), and also that not all *generosity* is *cooperation*; as with *non*-costly other-benefit (Rand & Kraft-Todd, 2014) or other-benefitting attitudes (Gulliford & Roberts, 2018).

Our treatment of *fairness* is similarly narrow, occupying an analogous subset of the conceptual space traditionally covered by the idea of *justice*. Since some of the earliest writings on the subject, the virtue of *justice* (Reeve, 2004) is often thought about in terms of whether it is a virtue of individuals (Hursthouse, 1999; as with our focus here) or societies (e.g., Rawls, 1971), as well as whether it regards justice in decision-making methods ('procedural justice') or in the distribution of resources ('distributive justice'; Tyler, 1994). Regarding justice as a virtue of individuals, it is frequently referred to as *fairness* (Peterson & Seligman, 2004), which is also the term predominantly used in empirical research on individual decision-making and behavior (e.g., McAuliffe et al., 2017). We follow this convention in our use of the term *fairness*, and further, we focus on *impartiality* – a 'procedural' aspect of this virtue – defining fairness (as impartiality) in our studies as, 'trait desire to treat others equally and without bias'. Our treatment of fairness is also narrow, for example, because we do not consider the need of recipients (i.e., charities) or obligations between interaction partners (e.g. reciprocity; Niemi & Young, 2017).

Our motivation for comparing generosity and fairness arises from recent work suggesting their dissociation. Some theories hold that these virtues are functionally linked, e.g., fairness (like generosity) enables individuals to form cooperative partnerships (Baumard et al., 2013). Contrary to this view, Shaw (2016) proposes that the function of fairness is to avoid punishment from an individual's cooperative partnerships, i.e. obscuring favouritism of some partnerships over others, engendering punishment from less favoured partners. This theoretical argument echoes earlier distinctions made by virtue scholars, similarly distinguishing generosity and fairness (respectively) as 'imperfect' vs. 'perfect' duties (per Grotius; Schneewind, 1990) and 'natural' vs. 'artificial' virtues (Hume, 1902). These theoretical distinctions between fairness and generosity have further been substantiated by a recent line of empirical work.

Consistent with recent arguments expressed by virtue scholars that virtues, in general, can come into conflict (Darnell et al., 2022, 2019; Kristjánsson & Fowers, 2022), a series of recent studies show how generosity (as cooperation) and fairness (as impartiality), in particular, can come into conflict. For example, in a resource distribution paradigm where participants must divide a pot with an odd number of resources (Kleiman-Weiner et al., 2017; Shaw & Olson, 2012), rather than distributing the entire pot (the cooperative choice), they choose to destroy a resource so that the pot is divided equally

(the impartial choice). Further, people engage in behaviour that is actually biased (i.e. partial) – 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). Acknowledging this limited scope of our investigation, it might be more precisely described as follows: 'a case study of generosity (as cooperation) 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*.

Toward a method for dissociating virtues

We propose a 3-stage, cumulative method for dissociating virtues based on a 'person-centred' social psychological understanding of virtue. First, we solicit free response text of behaviours demonstrating target virtues, and examine differences in participants' responses using structural topic modelling (STM; Roberts et al., 2014), a natural language processing algorithm. Second, we recruit a separate group of participants to rate these participant-generated example behaviours of each virtue on several features that impact moral judgment (e.g., 'costliness to the actor'), and examine differences in feature ratings across virtues. Third, we construct vignette-based experiments (again using participant-generated example behaviours of each virtue as stimuli), and examine differences in participants' ratings of hypothetical actors' motivations (e.g., 'to benefit others') across virtues.

Our method is motivated by a 'person-centred' social psychological understanding of virtue. In contrast to 'act-centred' investigations predominant in moral psychology research (Uhlmann et al., 2015), which focus on judgments of *behaviour* as the basic unit of analysis, 'person-centred' moral judgment instead focus on judgments of individuals' *character* (Pizarro & Tannenbaum, 2012). Our method is also motivated by work in social perception (Tamir & Thornton, 2018), which focuses on how people infer individuals' traits from their behaviours, and how these inferences influence social prediction. Because individuals' *motivations* provide essential information for predicting their future behaviour (Glasman & Albarracín, 2006) – and because 'proper motivation' has long been considered a requirement for virtue (Aristotle, 1999) – our method investigates judgments of both individuals' behaviour and their motivations to yield an understanding of the social perception of virtue. This emphasis is consistent with a recent, comprehensive conceptual framework (STRIVE-4) for the empirical study of virtue (Cokelet & Fowers, 2019) which similarly identifies behaviour and motivation as two 'major components' of virtue.

General methods

We recruited non-representative convenience samples (total N = 2,667; 38.6% female, average age = 37.0 years; note that we did not collect demographic information in Study 1) using the crowdsourcing tool Cloud Research and Amazon Mechanical Turk ('mTurk'; Arechar et al., 2017) and administered experiments using Qualtrics survey software. We excluded duplicate mTurk worker IDs and IP addresses to prevent analysing multiple observations per participant (as well as participants who dropped out prior to assignment to condition). Informed consent was obtained from all participants. In each experiment, we randomly assigned participants to condition, including our manipulation of virtue (generosity vs. impartiality; see SOM Section 7 for complete experimental instructions). Following the stimuli, we presented participants with dependent measures that varied by study (see respective study methods). All continuous dependent measures were answered on 100-point unmarked slider scales with extreme anchors labelled. We conducted analyses using STATA (16.1) and R software (4.1.2). We obtained effect sizes (Cohen's D) through the use of an online calculator (Lenhard & Lenhard, 2016). The data that support the findings of these studies are openly available on the Open Science Framework at https://osf.io/bw4fd/?view_only=2f590d72850e40a0a5bd4fffa31e05d3.

Study 1: participants' natural language distinguishes generous and impartial behaviours

Here we present an exploratory analysis of participants' natural language use in the description of behaviours expressing generosity (as cooperation) and fairness (as impartiality) We present data on participants' spontaneous descriptions of generous and impartial behaviours, and then analyse their language using structural topic modelling (STM; Roberts et al., 2014) to investigate whether the language used to describe these virtues is dissociable by the topics that emerge (see SOM Section 2 for expanded rationale of this method).

Methods

We recruited N = 114 participants and prompted them to generate example behaviours of each virtue using

free-response text boxes (average number of behaviours: generosity = 4.18; impartiality = 3.30; total behaviours = 853; see SOM Section 8). We analyse these data using the STM package in R (4.1.2; Roberts et al., 2019).

Results

We found that a 4-topic solution – comprised by *charity, kindness, equality*, and *helping* (our labels) – is preferable based on accepted criteria for model selection (Silge, 2018; see SOM Section 3). We then estimate topic proportions in participants' free-response (i.e. documents) across virtue condition. We found two topics were more frequent among examples of generous behaviours, which we characterize as *charity* (Topic 1; including terms such as 'charity', 'donate', 'homeless'; *coeff* = .33, $p < .001$; see Figure 1) and *kindness* (Topic 2; including terms such as 'care', 'stranger', 'pay'; *coeff* = .18, $p < .001$). Interpreting these coefficients, participant responses in the *generosity* compared to the *impartiality* condition were 33% and 18% more likely to use words from topics *charity* and *kindness* (respectively). We found one topic that was more frequent among examples of impartial behaviours, which we characterize as *equality* (Topic 3; including terms such as 'equal', 'coin', 'judge'; *coeff* = .47, $p < .001$). Interpreting this coefficient, participant responses in the *impartiality* compared to the *generosity* condition were 47% more likely to use words from the topic *equality*. Finally, we found one topic that was roughly equivalent among examples of both generous and impartial behaviours, which we characterize as *helping* (Topic 4; including terms such as 'help', 'volunteer', 'elder'; *coeff* = .04, $p = .032$). Interpreting this coefficient, participant responses in the *impartiality* compared to the *generosity* condition were 4% more likely to use words from the topic *helping*.

Discussion

These findings suggest that generosity (as cooperation) and fairness (as impartiality) can be distinguished by the behaviours participants spontaneously generate as examples. The evidence comes from the association of at least one topic with each virtue to a greater extent than the other (which we replicate across alternative models, see SOM Section 3). Further, the existence of a common topic – *helping* – across virtues supports the claim that virtues are graded concepts with 'fuzzy' boundaries. Having distinguished these virtues via participants' self-generated behaviours, we next measure whether these behaviours are merely semantically differentiable, or whether they also differ in terms of underlying features that affect moral judgment.

Study 2: generous and impartial behaviours are perceived differently on underlying features that affect moral judgment

Here we explore whether examples of generous and impartial behaviours – generated by participants (see Study 1) – are perceived differently across nine (non-exhaustive) underlying features that affect moral

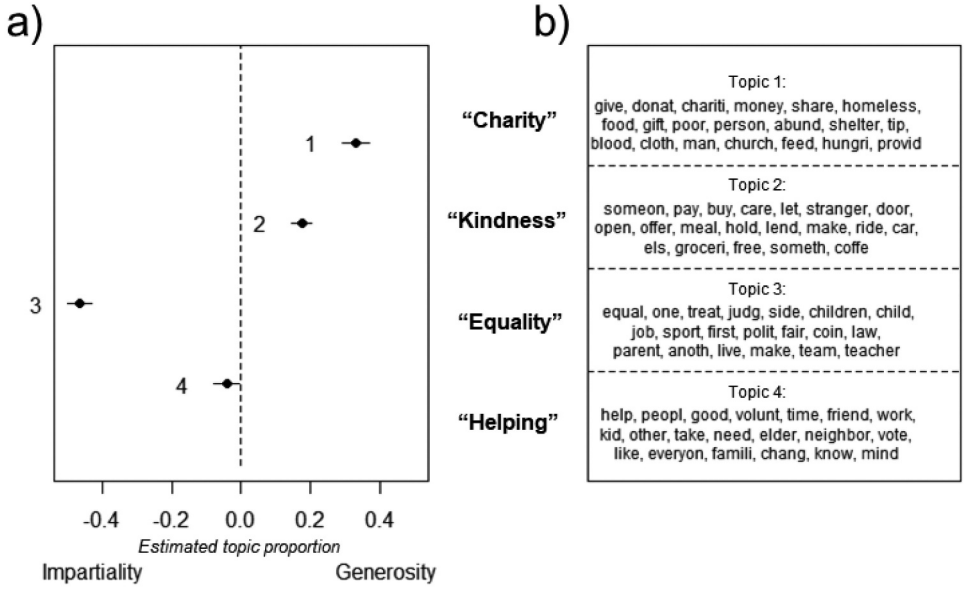

Figure 1. Participants use different topics (clusters of words) when providing examples of generous vs. impartial behaviours. Shown are a) the difference in topic prevalence for each topic depending on condition (with a 4-topic model); and b) the 20 most frequent word stems in each topic; *N*=114. These results are robust to changing the number of topics; see SOM Section 3.

Table 1. Feature rating labels and stimuli text. Shown are the nine feature rating items administered in Study 2 grouped by factor loadings from EFA. Bolded words indicate abbreviations used in text.

Factor	Feature	Item wording
Normativity	**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?"
Inauthentic altruism	**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?"
	Potential for **ulterior motives**	"How likely is it that someone engaging in this behavior does so for ulterior motives?"
	Benefit to the recipient	"In your opinion, how much benefit (in terms of money, time, effort, etc.) does the recipient of this behavior receive?"
	Potential for **anonymity**	"How possible is it for someone engaging in this behavior to be anonymous to the recipient(s) of this behavior?"
Virtue diagnosticity	**Prototypicality** as demonstration of the virtue	"How much does this behavior exemplify the virtue of [generosity/impartiality]?"
	Moral goodness	"In your opinion, how morally good is it to do this behavior?"
	Extent to which the behavior indicates the actor's **consistency** across situations	"How likely is it that someone engaging in this behavior acts similarly in other situations?"

judgment (see, Table 1, and also SOM Section 2 for the expanded rationale of this method).

Methods

To construct our stimuli, we began with the corpus of participant-generated examples of generous and impartial behaviours (total behaviours = 853; see Study 1 Methods and SOM Section 8). We minimally edited these to preserve semantic content (see SOM Section 4), yielding a list of 50 unique behaviours for both generosity and impartiality (see, SOM Table 2 and 3, respectively).

We recruited a sample of $N = 496$ and presented them with a randomly selected subset of 10 behaviours (presented in randomized order) from the 50 generated for the respective virtue (see SOM Section 7 for complete experimental instructions). Thus, each behaviour was rated by an average of $m = 47$ participants. Participants rated each behaviour on nine features (see, Table 1).

We use a generalized structural equation model (Hayes, 2013) to fit a multivariate, multilevel mixed-effects model to compare each feature rating (as the dependent measure) across virtues, with target behaviour nested within virtue condition, and estimate covariance for each pair of ratings. Our multivariate approach accounting for covariance among feature ratings allows us to account for the correlation among these measures for each participant. Our nesting of behaviour within virtue condition allows us to account for the variance explained by each behaviour in our estimation of the effect of virtue condition. We also conduct an exploratory factor analysis of feature ratings with iterated principal factors and oblique rotation. Our iterated re-estimation of the communalities enables us to explain a greater portion of the variance among feature ratings with fewer factors. We use oblique rotation because of the correlation among feature ratings (average absolute value $r = .29$, see SOM Section 5).

Results

Across all behaviours (for ratings of each behaviour, see SOM Section 6), we find that generous behaviours are rated higher than impartial behaviours on the features of (in decreasing order of mean difference): *cost to actors* (*coeff* = 10.60, $d = .37$, $p < .001$; see, Figure 2a), *benefit to recipients* (*coeff* = 7.08, $d = .24$, $p < .001$), potential for *ulterior motives* (*coeff* = 5.00, $d = .15$, $p < .001$), and potential for *anonymity* (*coeff* = 3.97, $d = .07$, $p = .017$). Conversely, we find that impartial behaviours are rated higher than generous behaviours on the features of *prototypical* as a demonstration of the virtue (*coeff* = 2.81, $d = .12$, $p < .001$) and *injunctive normativity* (*coeff* = 2.66,

Table 2. Motivational inference item labels and stimuli text. Shown are the six motivational inference rating items administered in Study 3 grouped by factor loadings from EFA.

Factor	Motivation item	Item wording*
Principled	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?
	Norm-signalling	"..because she wants others to be [generous/impartial], and she is trying to lead by example?"
Reputation-signalling	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?"

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

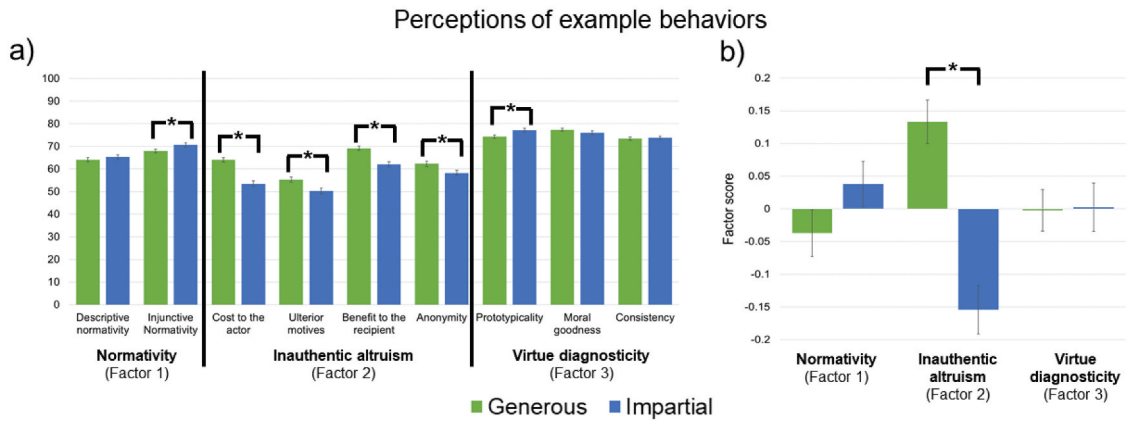

Figure 2. Participant-generated examples of generous and impartial behaviours are perceived differently across many underlying features. Shown are means (with 95% CIs) of ratings (0–100) across example behaviours of generosity (green) and impartiality (blue) generated by an independent group of participants. Ratings shown by **a)** item and **b)** factor scores. Significant contrasts denoted with (*); $N = 496$.

$d = .06$, $p = .028$). We find that generous and impartial behaviours are not perceived differently on the features of *moral goodness* (*coeff* = 1.34, $d = .04$, $p = .148$), *descriptive normativity* (*coeff* = −1.11, $d = .03$, $p = .370$), and the extent to which the behaviour is indicative of the actor's *consistency* across situations (*coeff* = −.41, $d = .02$, $p = .545$).

To better understand the correlation structure among ratings, we next examine our factor analytic results. The analysis yielded three factors explaining 95.7% of the variance. Factor 1 explained 57.3% of the variance and we labelled it 'normativity' due to high loadings (>.4) by the items: *descriptive normativity* and *injunctive normativity*. Factor 2 explained 26.4% of the variance and the items with high loadings (>.4) were *cost to the actor*, potential for *ulterior motives, benefit to the recipient*, and potential for *anonymity*. We labelled Factor 2 'inauthentic altruism' because cost to the actor is often discussed when costly signalling theory is applied to human behaviour (Jordan et al., 2016), and because ulterior motives undermine actors' intended signalling. Factor 3 explained 11.9% of the variance and we labelled it 'virtue diagnosticity' due to high loadings (>.4) by the items: *prototypicality* as demonstration of virtue, *moral goodness*, and the extent to which the behaviour is indicative of the actor's *consistency* across situations. These factors share low-to-moderate correlations (*normativity* and *inauthentic altruism*, $r = .39$; *normativity* and *virtue diagnosticity*, $r = .38$; *inauthentic altruism* and *virtue diagnosticity*, $r = .16$).

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 behaviour nested within virtue condition, and estimating covariance for each pair of factor scores (this analysis strategy was motivated by the same logic as the item-level analysis described above). Across all behaviours, we find that generous compared to impartial behaviours are perceived as higher on the *inauthentic altruism* factor (*coeff* = .27, $d = .31$, $p < .001$; see, Figure 2b). Consistent with the item-level analysis, generous and impartial behaviours are not perceived differently on the *normativity* factor (*coeff* = −.07, $d = .03$, $p = .261$) or the *virtue diagnosticity* factor (*coeff* = −.002, $d = .001$, $p = .961$).

Discussion

Generous and impartial behaviours were statistically distinguished on 6 of the 9 measured features. Although we do not intend this to be an exhaustive exploration of the feature space, it is notable that we find significant differences across the majority of features employed. Also, across all participant-generated behaviours, generosity (as cooperation) and fairness (as 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. Further, consistent with the ideas that virtue concepts are 'fuzzy' and include some overlap, these virtues were not perceived differently on the two broad factors of *normativity* and *virtue diagnosticity*.

Our exploratory factor analysis results may also be instructive for more precisely describing the difference between these two virtues. Namely, that generosity (as cooperation) may be characterized by more *inauthentic altruism* than fairness (as impartiality). This result is consistent with previous research showing, for example, that actors express generosity to attract cooperation

partners (Barclay & Willer, 2007). This *inauthentic altruism* factor also captures an 'act-person dissociation' (Uhlmann et al., 2015) that is a source of tension and underlies accusations of 'virtue signalling.' On the one hand, it includes items measuring altruism (i.e., cost to the actor and benefit to the recipient) that contribute to perceptions that such an act is morally good. On the other hand, this factor also includes items that could speak to observers' perceptions that people engaging in these acts do so inauthentically; potential for ulterior motives speaks directly to this idea, while potential for anonymity implies that actors could have concealed their altruism even if they so desired (as often prescribed in considerations of charitable giving; De Freitas et al., 2019).

Thus far we have been primarily concerned with act-centred judgments of virtuous behaviours. Having distinguished (participant-generated) generous and impartial behaviours in terms of participants' perceptions on underlying features here – as well as natural language use (Study 1) – we now turn to person-centred judgments of actors' motivations in Study 3.

Study 3: generous and impartial actors are perceived to have different motivations

In Study 2, we focused on how participants perceive virtuous *acts*. Here, we explore participants' perceptions of virtuous *actors*, and specifically whether participants perceive virtuous (generous vs. impartial) actors to have different motivations. We constructed hypothetical vignette scenarios in which we described actors who publicly engaged in a set of behaviours (generated by participants, see Study 1) and asked participants to rate the extent to which they perceived that actors had six (non-exhaustive) motivations of interest (see, Table 2, and also SOM Section 2 for expanded rationale for this method). To create a more general impression of the actors as demonstrating each virtue in participants' minds, we created scenarios describing actors engaging in a set of three behaviours (rather than a single behaviour, as in the previous analysis). The goal of our design strategy was to avoid any idiosyncrasies of the specific behaviour used to express the hypothetical actors' virtue; by describing an actor who engages in a *set* of behaviours, we aim to capture participants' perceptions of actors' *virtue* with higher fidelity.

Methods

To maximize our power in this analysis, we use data from seven experiments in which we administered six motivational inference items (this is an exploratory analysis; primary analyses, including preregistered analyses, and methodological details not pertinent to the present analysis available here: https://osf.io/sud3m/?view_only=380a169770b9474f93d2b5b73adc7410). We recruited a sample of N = 2,057 who indicated their perceptions of the actor's motivation on six dependent measures (see, Table 2) presented in randomized order.

We conduct a multivariate regression analysis to test for evidence that motivational inferences differentiate generosity and impartiality (controlling for experiment). Our multivariate approach allows us to account for the correlation among these measures for each participant and we control for experiment to account for variance in these measures explained by idiosyncrasies across designs and samples. We also conduct an exploratory factor analysis of feature ratings with iterated principal factors and oblique rotation. Our iterated re-estimation of the communalities enables us to explain a greater portion of the variance among feature ratings with fewer factors. We use oblique rotation because of the correlation among feature ratings (average absolute value r = .34, see SOM Section 5).

Results

We find that participants infer that generous compared to impartial actors are more motivated by *other-benefit* (*coeff* = 4.25, d = .20, p < .001; see, Figure 3a) and *self-benefit* (*coeff* = 2.62, d = .10, p = .025). Conversely, we find that participants infer that impartial compared to generous actors are more motivated by *norm signalling* (*coeff* = 5.25, d = .24, p < .001), *moral rule* (*coeff* = 4.52, d = .24, p < .001), *virtue identification* (*coeff* = 2.57, d = .13, p = .003), and marginally *self-presentation* (*coeff* = 2.23, d = .09, p = .054).

To better understand the correlation structure among motivational inference measures, we conduct an exploratory factor analysis with iterated principal factors and oblique rotation. The analysis yielded two factors explaining 96.0% of the variance. Factor 1 explained 65.2% of the variance and the items with high loadings (>.4) were as follows: *moral rule, virtue identification, other-benefit* and *norm-signalling*. We labelled Factor 1 'principled' because these motivations pertain either to actors' moral beliefs/identity (i.e. moral rule and virtue identification), prosociality (other-benefit) or both (norm-signalling). Factor 2 explained 30.8% of the variance and the items with high loadings (>.4) were as follows: *self-presentation* and *self-benefit*. We labelled Factor 2 'reputation-signalling' (Kodipady et al., 2021) because the *reputation* construct can describe their combination. These factors have a small-to-moderate and negative correlation (r = –.21).

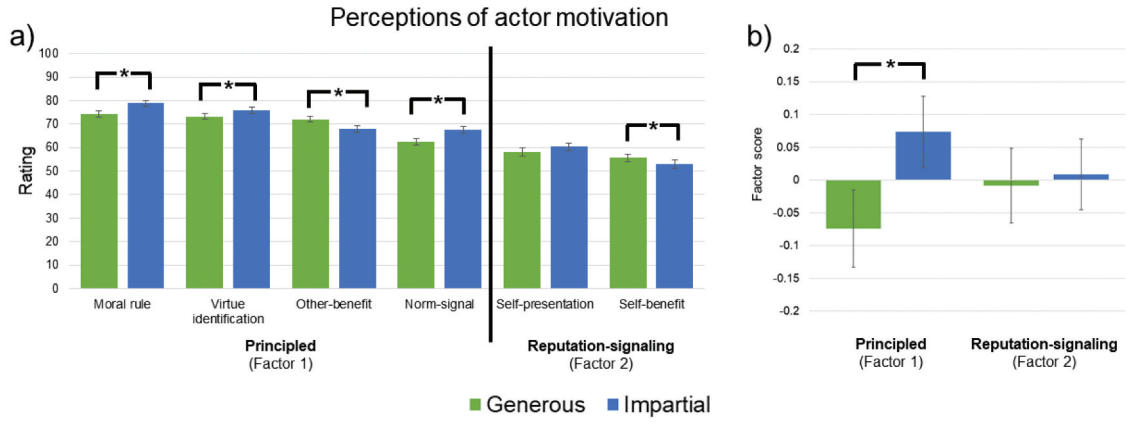

Figure 3. Participants' motivational inferences differentiate generous and impartial actors. Shown are means (with 95% CIs) of ratings (0–100 unmarked slider) of generosity (green) and impartiality (blue), collapsed across observability manipulation and Study. Ratings shown by **a)** item and **b)** factor scores. Significant contrasts denoted with (*); N = 2,057.

We then conduct a multivariate regression analysis to test for evidence that motivational inferences factor scores (as the dependent measures) differentiate generosity and impartiality (controlling for experiment; this analysis strategy is motivated by the same logic as the item-level analysis described above). Consistent with the item-level analysis, we find that participants infer that impartial compared to generous actors have more *principled* motivation (*coeff* = .15, d = .17, p < .001; see, Figure 3b), while participants do not infer different levels of *reputation-signalling* motivation for impartial compared to generous actors (*coeff* = .02, d = .03, p = .566).

Discussion

We find significant differences in perceptions of hypothetical virtuous actors' motivations across most (5 of 6) items measured, echoing our results in Study 2. Our exploratory factor analysis results are particularly provocative, in that the low correlation (r = −.21) between the (obliquely rotated) factors approximately representing *principled* and *reputation-signalling* motivations may function – at least in observers' perceptions – semi-orthogonally, rather than as two ends of a unidimensional construct. This finding sheds light on the phenomenon of 'virtue signalling', suggesting that observers may simultaneously infer the presence of *both* motivations, and suggests that higher perceptions of one may not necessarily entail lower perceptions of the other. Finally, we note that inferences of *moral rule* motivation and *virtue identification* motivation are highly correlated (r = .65), and further, they are more highly correlated than any other two motivations. Consistent with the findings that people self-identify more strongly with moral traits than other mental capacities (Strohminger & Nichols, 2014), these results imply that observers also strongly associate 'morality' and 'self' in their perceptions of others.

General discussion

Across three analyses utilizing data from three studies (total N = 2,667) we provide evidence that *generosity* (as cooperation) and *fairness* (as impartiality) are dissociable via 1) the semantic content of behaviours participants spontaneously generate as examples of each virtue; 2) ratings of the underlying features of participant-generated acts; and 3) inferences of the motivations driving hypothetical actors who engage in those acts. Our approach and results have several implications for the study of virtue, which we discuss through the points of intersection across our analyses. We also discuss limitations to our general approach.

We begin by examining differences between generosity and impartiality across our factor analytic results (Studies 2 and 3). To reiterate, in Study 2, we observed that generous (compared to impartial) *acts* are perceived to have higher *inauthentic altruism* (a factor of feature ratings comprising cost to the actor, potential for ulterior motives, benefit to the recipient and potential for anonymity). In Study 3, we observed that publicly generous (compared to impartial) *actors* are perceived to have lower *principled* motivation (a factor of motivational inferences comprising moral rule, virtue identification, other-benefit and norm-signalling). Interestingly, however, in Study 3 we also find that participants do not rate generous and impartial actors differently with regard to their *reputation-signalling* motivation (a factor of motivational inferences comprising self-benefit and self-presentation). Most previous work demonstrating

'virtue signalling' in the context of generosity (e.g., Lin-Healy & Small, 2012), focuses on the role of ulterior (i.e., selfish) motives in depreciating the value of these acts. This pattern of results suggests, instead, that lower *principled* motivation (rather than higher *reputation-signalling* motivation) may be the mechanism that drives these results. Additionally, this suggests that observers may perceive generosity to have greater motivational ambiguity than impartiality. There are several potential explanations for this pattern that future work might explore; for example, compared to impartial actors, generous actors may more easily conceal their motives, have more variation in their motives; and/or have more inscrutable motives.

Next, we consider parallels among item-level analyses of virtuous *acts* (Study 2) and publicly virtuous *actors* (Study 3). First, compared to generous *acts*, impartial acts are perceived to be more injunctively normative (Study 2); and compared to publicly generous *actors*, publicly impartial actors are perceived to be more motivated by moral rules and by norm-signalling (Study 3). In the social norms literature, *injunctive normativity* and *moral rules* are types of social rules that share 'prescriptive social expectations' (Bicchieri, 2006); that is, they are both concerned with what people think others should do. Given this similarity across these constructs, consistent findings across our act-based (Study 2) and person-based (Study 3) approaches might be unsurprising. Second, compared to impartial *acts*, generous acts are perceived as being costlier to actors (Study 2), although compared to publicly impartial *actors*, publicly generous actors are perceived to be more motivated to benefit themselves (Study 3). This pattern of results is consistent with the 'partner choice' explanation for public generosity, which holds that individuals may express generosity in public in order to gain cooperative interaction partners (Barclay & Willer, 2007). Finally, we find convergent evidence for the other-benefiting nature of generosity across all three analyses. Charitable giving is perhaps the paradigmatic example of a 'purely' altruistic prosocial behaviour, and we observe that participants use words related to the topic of *charity* more for generosity than impartiality (Study 1). Further, compared to impartial *acts*, generous acts are perceived as being more beneficial to recipients (Study 2); and also, compared to publicly impartial *actors*, publicly generous actors are perceived to be more motivated to benefit others (Study 3).

Reflecting on the commonalities between generosity (as cooperation) and fairness (as impartiality) across our analyses supports our conceptualization of the 'fuzzy' boundaries between virtues. In Study 1, *helping* emerged as a topic (i.e., cluster of words) among participant-generated behavioural examples of both virtues. In Study 2, acts demonstrating these virtues are not perceived differently on the underlying features of *normativity* and *virtue diagnosticity*. In Study 3, hypothetical actors demonstrating each virtue publicly are not perceived to be differently motivated by *reputation-signalling*. In other words, despite the numerous differences between these virtues across our analyses to which we have dedicated the majority of our discussion, generosity (as cooperation) and fairness (as impartiality) also have shared features. First, they are both exemplified by helping behaviours that are equivalently normative and diagnostic of the respective virtue (although whether the latter is due to inherent qualities of the virtues themselves compared to the process people use to generate acts of a particular virtue remains an interesting avenue for future research). Second, observers infer that actors publicly demonstrating each virtue are motivated to benefit their reputations to a similar extent. Although our investigation is limited to two virtues (and further, a 'narrow' definition of both), these commonalities are suggestive of 'core features of virtue' that future work might uncover by expanding the number of virtues explored.

Our investigation has several limitations that future work might address to better understand perceptions of virtue. Although Study 1 represents the first demonstration, to our knowledge, of the use of natural language processing algorithms (specifically, structural topic modelling) to the study of virtue, we note that the data employed for our analysis (853 participant-generated behaviours) were quite sparse for this type of analysis (average = 5.02 words per behaviour) and that future work might elicit longer responses to enhance the validity of findings using this method. Interestingly, we note that 12.5% of the top 20 words in our 4-topic solution (i.e. 10 of 80) designated social roles (e.g., 'parent'), and given work showing the importance of social roles in moral judgment (McManus et al., 2020), future work might explore the relevance of distinct social roles to distinct virtues. Finally, despite the novel insights made possible by the use of natural language processing algorithms in the study of virtue, future work might simultaneously employ this computational method alongside traditional qualitative methods for coding text to achieve richer understandings of how the lay public talks about virtue.

Regarding Study 2, we contend that the nine features we measured will also have relevance for the perception of other virtues (see, SOM Table 1), but we do not claim that these nine features are exhaustive. For example, much previous research suggests the importance of 'intentionality' as a feature that impacts moral judgment

(e.g., Cushman & Young, 2011; relevant to the fourth core component of virtue in the STRIVE-4 model; Fowers et al., 2021). Future work might more comprehensively employ this and other features known to impact judgment of virtue, as well as applying this method (and these dimensions) to perceptions of other virtues. Consistent with this limitation, 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 similarly measure a greater range of motivations for virtue.

Finally, we note three limitations to our bottom-up approach that future work might consider. Although we have framed our investigation as bottom up, we acknowledge that it is not maximally so because we presupposed the virtue concepts of generosity (as cooperation) and fairness (as impartiality), and provided participants with definitions of these concepts in our stimuli. A fully bottom-up approach to the study of virtue would not assume the definition (or existence) of *any* concepts, and would attempt to discern these too from the bottom up. For example, such an investigation might elicit examples of 'virtuous' behaviour with minimal instruction, and subsequently differentiate them by the method we employ here. Exciting recent work using 'prototype analysis' (Gulliford et al., 2021) represents a promising step in this direction. Although we have emphasized the dearth of bottom-up approaches to the study of virtue and have advocated for increased research activity from this perspective, we believe that there is much to be gained by simultaneously pursuing the science of virtue from both this approach in combination with traditional top-down approaches motivated by *a priori* theorizing.

Next, we are limited in the generalizability of our findings to the virtues of generosity and fairness because we used 'narrow' definitions of these virtues in our stimuli. For example, applying our method to the subdomain of generosity expressed by *non*-costly other-benefitting behaviour (e.g., Rand & Kraft-Todd, 2014), we might expect a reversal of the pattern of results we observe in Analysis 2 regarding the factor of *inauthentic altruism* feature ratings, because such behaviours are (definitionally) less costly to actors and also likely have less potential for anonymity than the behaviours elicited by our 'cooperation' definition of generosity. Similarly, applying our method to the subdomain of fairness expressed by behaviours carrying expectations of reciprocity (e.g., Niemi & Young, 2017), we might expect a reversal of the pattern of results we observe in Analysis 3 regarding the factor of *principled* motivation because reciprocity is a sufficiently common and widespread expectation in social interactions that justifications *qua* individuals' moral rules or norm-signalling motivations seem overdetermined. To better understand the virtues of *generosity* and *fairness*, generally, future work might therefore expand the assessment of these virtues beyond the narrow case study of generosity (as cooperation) and fairness (as impartiality) that we present here.

A crucial direction for comprehensive bottom-up approaches to virtue is to document cross-cultural heterogeneity in virtue concepts. The modern study of virtue is heavily influenced by the spiritual and philosophical traditions that began contemplating the concept. Although we might forgive the universality among humans assumed by early writers in these traditions because human cultures in their time were less interconnected than they are today, we should resist recapitulating such overgeneralization. Ultimately, the value of a trait is contingent upon the social norms of the culture under investigation; one culture's virtue may be another's vice (and yet another's amoral trait). Evidence for the need for such cross-cultural study is provided, for example, by the many failed attempts to replicate the VIA structure of virtues in non-US samples (summarized by McGrath, 2014).

Conclusion

We provide evidence indicating the utility of a novel, three-stage, 'bottom-up' method for dissociating virtues using *generosity* (as cooperation) and *fairness* (as impartiality) as a case study among virtues. We suggest that these virtues can be differentiated by 1) the natural language people use to describe behaviours demonstrating these virtues; 2) ratings across several features of these behaviours; and 3) motivational inferences of actors demonstrating these virtues. We hope that such basic virtue science will be translated into interventions promoting virtue in the real world.

Acknowledgments

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

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This research was made possible by funding by the John Templeton Foundation. 62221, The Virtue Project at the Morality Lab at Boston College, and National Science Foundation (NSF) award #1627157.;

ORCID

Gordon T. Kraft-Todd http://orcid.org/0000-0003-1220-9269
Max Kleiman-Weiner http://orcid.org/0000-0002-6067-3659
Liane Young http://orcid.org/0000-0001-5178-1853

Data availability statement

The data described in this article are openly available in the Open Science Framework at https://doi.org/10.1080/17439760.2022.2154254.

Open Scholarship

This article has earned the Center for Open Science badges for Open Data and Open Materials through Open Practices Disclosure. The data and materials are openly accessible at https://doi.org/10.1080/17439760.2022.2154254.

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*, *3*(1), 1–11. https://doi.org/10.1007/s40881-017-0035-0

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

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

Baumard, N., André, J.-B., & Sperber, D. (2013). A mutualistic approach to morality: The evolution of fairness by partner choice. *Behavioral and Brain Sciences*, *36*(1), 59–78. https://doi.org/10.1017/S0140525X11002202

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

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

Cowen, A. S., & Keltner, D. (2017). Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the National Academy of Sciences*, *114*(38), E7900–E7909. https://doi.org/10.1073/pnas.1702247114

Cushman, F., & Young, L. (2011). Patterns of moral judgment derive from nonmoral psychological representations. *Cognitive Science*, *35*(6), 1052–1075. https://doi.org/10.1111/j.1551-6709.2010.01167.x

Dahlsgaard, K., Peterson, C., & Seligman, M. E. P. (2005). Shared virtue: The convergence of valued human strengths across culture and history. *Review of General Psychology*, *9*(3), 203–213. https://doi.org/10.1037/1089-2680.9.3.203

Darnell, C., Fowers, B. J., & Kristjánsson, K. (2022). A multifunction approach to assessing Aristotelian phronesis (practical wisdom). *Personality and Individual Differences*, *196*, 111684. https://doi.org/10.1016/j.paid.2022.111684

Darnell, C., Gulliford, L., Kristjánsson, K., & Paris, P. (2019). Phronesis and the knowledge-action gap in moral psychology and moral education: A new synthesis? *Human Development*, *62*(3), 101–129. https://doi.org/10.1159/000496136

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

Fowers, B. J. (2014). Toward programmatic research on virtue assessment: Challenges and prospects. *Theory and Research in Education*, *12*(3), 309–328. https://doi.org/10.1177/1477878514546064

Fowers, B. J., Carroll, J. S., Leonhardt, N. D., & Cokelet, B. (2021). The emerging science of virtue. *Perspectives on Psychological Science*, *16*(1), 118–147. https://doi.org/10.1177/1745691620924473

Glasman, L. R., & Albarracín, D. (2006). Forming attitudes that predict future behavior: A meta-analysis of the attitude–behavior relation. *Psychological Bulletin*, *132*(5), 778–822. doi:10.1037/0033-2909.132.5.778.

Gulliford, L., Morgan, B., & Jordan, K. (2021). A prototype analysis of virtue. *The Journal of Positive Psychology*, *16*(4), 536–550. https://doi.org/10.1080/17439760.2020.1765004

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

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

Hume, D. (1902). *Enquiries concerning human understanding and concerning the principles of morals*. Clarendon Press.

Hursthouse, R. (1999). *On virtue ethics*. Oxford University Press.

Inbar, Y. (2018). Applied moral psychology. In Graham, J., Gray, K. (eds.), *Atlas of moral psychology* (pp. 537–543). The Guilford Press.

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

Kleiman-Weiner, M., Shaw, A., & Tenenbaum, J. B. (2017). Constructing social preferences from anticipated judgments: When impartial inequity is fair and why? 676–681. https://cogsci.mindmodeling.org/2017/papers/0137/paper0137.pdf.

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

Kristjánsson, K., & Fowers, B. J. (2022). Phronesis as moral decathlon: Contesting the redundancy thesis about phronesis. *Philosophical Psychology*, 1–20. https://doi.org/10.1080/09515089.2022.2055537

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. https://doi.org/10.1016/j.obhdp.2011.11.006

MacIntyre, A. (1981). *After virtue: A study in moral theory*. University of Notre Dame Press.

McAuliffe, K., Blake, P. R., Steinbeis, N., & Warneken, F. (2017). The developmental foundations of human fairness. *Nature*

Human Behaviour, *1*(2), 0042. https://doi.org/10.1038/s41562-016-0042

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

McGrath, R. E. (2015). Integrating psychological and cultural perspectives on virtue: The hierarchical structure of character strengths. *The Journal of Positive Psychology*, *10*(5), 407–424. https://doi.org/10.1080/17439760.2014.994222

McManus, R. M., Kleiman-Weiner, M., & Young, L. (2020). What we owe to family: The impact of special obligations on moral judgment. *Psychological Science*, 0956797619900321. https://doi.org/10.1177/0956797619900321

Narvaez, D., & Snow, N., Eds. (2019). Introduction to self, motivation and virtue studies. In *Journal of Moral Education*, *48*(1), 1– 6.

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

Osherson, D. N., Smith, E. E., Wilkie, O., López, A., & Shafir, E. (1990). Category-based induction. *Psychological Review*, *97* (2), 185–200. https://doi.org/10.1037/0033-295X.97.2.185.

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*. Washington, DC: APA PsycBooks. https://doi.org/10.1037/13091-005

Rand, D. G., & Kraft-Todd, G. T. (2014). Reflection does not undermine self-interested prosociality In *Frontiers in behavioral neuroscience* (pp. 300. Vol. 8). PMC. https://doi.org/10.3389/fnbeh.2014.00300

Rand, D. G., & Nowak, M. A. (2013). Human cooperation. *Trends in Cognitive Sciences*, *17*(8), 413–425. https://doi.org/10.1016/j.tics.2013.06.003

Rawls, J. (1971). *A theory of justice*. Belknap Press of Harvard University Press.

Reeve, C. D. (2004). *Plato: Republic*. Hackett Publishing Company, Inc.

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 .

Roberts, M. E., Stewart, B. M., & Tingley, D. (2019). stm: An R Package for structural topic models. *Journal of Statistical Software*, *91*(2), 1–40. https://doi.org/10.18637/jss.v091.i02

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., & Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science*, *58*(4), 1064–1082. https://doi.org/10.1111/ajps.12103

Schneewind, J. B. (1990). The Misfortunes of Virtue. *Ethics*, *101* (1), 42–63. https://doi.org/10.1086/293259

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. https://doi.org/10.1037/a0025907

Silge, J. (2018, September 8). *Training, evaluating, and interpreting topic models*. *Blog: Machine Learning, Text Analysis, and More*. https://juliasilge.com/blog/evaluating-stm/

Strohminger, N., & Nichols, S. (2014). The essential moral self. *Cognition*, *131*(1), 159–171. https://doi.org/10.1016/j.cognition.2013.12.005.

Swanton, C. (2003). *Virtue ethics: A pluralistic view*. Oxford University Press.

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

Trivers, R. (1971). The Evolution of Reciprocal Altruism. *The Quarterly Review of Biology*, *46*(1), 35–57. https://doi.org/10.1086/406755

Tyler, T. (1994). Psychological models of the justice motive: Antecedents of distributive and procedural justice. *Journal of Personality and Social Psychology*, *67*(5), 850–863. https://doi.org/10.1037/0022-3514.67.5.850

Uhlmann, E. L., Pizarro, D., & Diermeier, D. (2015). A Person-centered approach to moral judgment. *Perspectives on Psychological Science*, *10*(1), 72–81. https://doi.org/10.1177/1745691614556679

Zadeh, L. (1965). Fuzzy Sets. *Information and Control*, *8*(3), 338–353. https://doi.org/10.1016/S0019-9958(65)90241-X