COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 48 (2024) e13406

ISSN: 1551-6709 online
DOI: 10.1111/cogs.13406

Evaluative Deflation, Social Expectations, and the Zone of Moral Indifference

Pascale Willemsen,[a] Lucien Baumgartner,[a] Bianca Cepollaro,[b] Kevin Reuter[a]

[a] *Department of Philosophy, University of Zurich*
[b] *Faculty of Philosophy, University Vita-Salute San Raffaele*

Received 6 April 2023; received in revised form 6 November 2023; accepted 9 January 2024

Abstract

Acts that are considered undesirable standardly violate our expectations. In contrast, acts that count as morally desirable can either meet our expectations or exceed them. The zone in which an act can be morally desirable yet not exceed our expectations is what we call the zone of moral indifference, and it has so far been neglected. In this paper, we show that people can use positive terms in a deflated manner to refer to actions in the zone of moral indifference, whereas negative terms cannot be so interpreted.

Keywords: Evaluative language; Thin and thick concepts; Moral judgments; Blame; Praise; Cancellability test; Norms

1. Introduction

In this paper, we show that the way we use morally evaluative terms to talk about other people's character or behavior is systematically asymmetrical. Negative evaluative terms are used to assign blame to a person when they do not meet some moral standard of expectable, common moral decency. The assertive use of *negative* evaluative terms almost exclusively serves the purpose of negatively evaluating moral transgressions and assigning blame for them. In contrast, *positive* evaluative terms have two different uses: (1) a proper evaluative use that is intended to speak positively about a person and to assign praise because our moral expectations have been exceeded; and (2) an evaluatively deflated use that lacks an expression

Correspondence should be sent to Pascale Willemsen, Department of Philosophy, University of Zurich, Zürichbergstrasse 43, 8044 Zürich, Switzerland. E-mail: Pascale.Willemsen@uzh.ch

of positive attitudes and praise. This *evaluatively deflated* use indicates that an action lies in the *zone of moral indifference*, in which we remain morally unimpressed or underwhelmed.[1] Metaethicists and moral psychologists have not paid proper attention to the evaluatively deflated use of positive moral terms and to the zone of moral indifference. However, they are essential to a full-fledged understanding of the practice of holding others responsible. What is more, this phenomenon challenges the frequently made assumption that positive and negative moral language work alike. Such an asymmetry deserves scholarly consideration.[2]

We develop these ideas based on previous empirical evidence as well as three new experimental studies. In Section 2, we provide an overview of the literature indicating that positive and negative moral judgments are asymmetrical—both at the level of psychological processing and the linguistic level. We zoom in on a recently detected effect, namely, the *polarity effect* of evaluative language (Baumgartner, Willemsen, & Reuter, 2022; Väyrynen, 2021; Willemsen & Reuter, 2021). The polarity effect can be described thus: the evaluation of a negative term, such as rude or cruel, is harder to explicitly cancel than the evaluation of a positive term, for example, friendly or compassionate. We argue that the polarity effect can only be properly understood by postulating two uses of positive terms in ordinary discourse.

In Section 3, we present our first preregistered experiment. We demonstrate that positive evaluative language can be used to say something positive about a person, and also in an evaluatively deflated way in which the speaker convincingly claims to intend to remain neutral. We use an adapted version of the cancellability test to determine whether the evaluation can be neutralized, that is, whether the evaluation is defeasible. While only very few people consider the evaluation of a negative term defeasible, a large number of people make this judgment about positive terms.

In Section 4, we provide evidence that the defeasibility of positive terms is best explained by the speaker's social expectations as to how people should act. We demonstrate that participants in our second preregistered study do, in fact, consider positive behavior more expectable compared to negative behavior. At the same time, however, participants indicate that, in general, they consider positive traits, such as honesty or friendliness, something to approve of.

We present our third experiment in Section 5. With the help of vignette studies, we manipulate directly whether an agent's behavior merely meets an expectation or exceeds it. Our results suggest that when an agent's behavior only meets the expectation, and someone refers to this behavior using a positive thick term, this term is interpreted to be used less positively and to assign little to no praise. When the behavior exceeds the expectation, the thick term is considered to be used very positively and in order to assign a significant amount of praise.

2. Evaluative language in moral cognition

2.1. Moral cognition—Symmetrical or asymmetrical?

The last two to three decades have brought about an explosion of publications dedicated to the investigation of the cognitive foundations of morality and the effects that moral or, more generally, normative intuitions have on purportedly non-normative judgments. Researchers have examined the psychological processes leading to blame judgments (e.g., Alicke, 2000; Cushman, 2013; Greene, 2009; Greene & Haidt, 2002; Hauser, Cushman, Young, Kang-Xing Jin, & Mikhail, 2007; Ditto, Pizarro, & Tannenbaum, 2009; Knobe, 2010; Mallon & Nichols, 2011; Willemsen, 2019; Young & Phillips, 2011; Young & Tsoi, 2013), and several models of moral cognition have been proposed (e.g., Fincham & Shultz, 1981; Darley & Shultz, 1990; Greene, Nystrom, Engell, Darley, & Cohen, 2004; Malle, Guglielmo, & Monroe, 2014). In addition to understanding how we make moral judgments, the question of when we are willing to revise those judgments has been addressed. For instance, scholars have investigated the circumstances under which we are willing to excuse others for otherwise blameworthy acts (e.g., Amaya & Doris, 2015; Gray & Wegner, 2011; Hauser et al., 2007; Kneer & Machery, 2019; Turri & Blouw, 2015; Woolfolk, Doris, & Darley, 2006). What is more, substantial research has been conducted on how moral intuitions can be manipulated and biased (e.g., Kern & Chugh, 2009; Petrinovich & O'Neill, 1996; Shenhav & Greene, 2010; Bartels & Medin, 2007; Gino, Shu, & Bazerman, 2010; Powell et al., 2014; Ritov & Baron, 1999; Rai & Holyoak, 2010; Shenhav & Greene, 2010).

Providing a full overview of the diversity and complexity of the relevant psychological literature goes beyond the scope of this paper (for overviews, see, e.g., Doris et al., 2020; Waldmann et al., 2012, and Wiegmann & Sauer, 2021). However, what unites this research is its almost exclusive attention to negative, blame-related phenomena resulting from harmful consequences or norm violations. How can we explain this strong focus on negativity, blame, and blame-related phenomena and norm violations, and also the neglect of positivity, praise, and praise-related phenomena?

One reason for why praise-related phenomena have received less attention may stem from the philosophical assumption that praise is the positive counterpart to blame—"praiseworthiness is methodologically mirrored in blameworthiness" (Eshleman, 2014, p. 217 see also Stout, 2020; Talbert, 2023). By understanding how moral cognition about negative cases works, we can (it is often assumed) infer all we need to know about positive cases as well.

Symmetry: Blame and praise are symmetrical in any philosophically and psychologically relevant respects. Therefore, by understanding blame, we can make proper inferences about the nature of praise.

Additionally, it has further been argued that the practice of blaming others is theoretically more interesting and illuminating. Blame has significant social consequences and, if done unjustifiably, blaming raises several normative problems. This attitude is visible in, for instance, Wallace (1994): "praise does not seem to have the central, defining role that blame and moral sanction occupy in our practice of assigning moral responsibility" (p. 61).

Priority of blame: Since blaming has more severe ethical implications for our social interactions, we should focus our research activities on the examination of blame.

Both these assumptions stand in sharp contrast to and are challenged by empirical research. It has been demonstrated that people are generally more sensitive to bad than to good outcomes (Siegel et al., 2017). The so-called negativity bias is a well-documented psychological effect in which negative events are not only more salient to people and evaluated more intensely, but memory of negative events is often more pervasive (Ito et al., 1998). Also, some factors that seem vital to determining an agent's blameworthiness, such as their causal involvement, alternative behavioral options, and the agent's control (Ohtsubo, 2007; Pizarro et al., 2003), are less recognized and considered in judgments of praise. In turn, the agent's intentions when performing a good deed are evaluated more critically. Even exceptional actions are often not considered praiseworthy if the agent had questionable intentions (Critcher & Dunning, 2011). Guglielmo and Malle (Guglielmo & Malle, 2019) provide additional empirical evidence supporting such a blame–praise asymmetry. They propose that our responses to negative events show more differentiation and variety and that they are expressed in a more fine-grained way compared to positive events. Another asymmetry arises in the context of evaluation adaptations due to new evidence. Negative actions are considered more diagnostic of a person's real moral character, and people quickly change their explicit positive evaluation of a person in light of new, negative information (Gregg et al., 2006; Rydell et al., 2007; Rydell & McConnell, 2006). In contrast, negative judgments remain fairly robust even if new, positive counter-evidence is presented. It has recently been found that this effect even arises for "implicit" judgments, which were previously thought to be quite robust against new information (Cone & Ferguson, 2015). Negative evaluations of a person are, thus, more robust and consequential than positive evaluations.

Anderson, Crockett, and Pizarro (2020) recently argued that more attention needs to be dedicated to praise and positive moral judgments. They argue that while both praise and blame are essential to sustaining social relationships and facilitating social regulation, "praise is relatively more directed towards building, establishing, and maintaining relationships and affiliative alliances" (Anderson et al., 2020, p. 696; for similar, yet philosophical arguments see Eshleman, 2014, and Stout, 2020). Given the importance of such relationships and social alliances, praise and related attitudes have unjustifiably been underrepresented in philosophical and empirical investigations. Even more strikingly, Anderson and colleagues identified a totality of only 22 empirical papers in which positive and negative moral judgments are tested side by side. Of those 22 papers, 20 report asymmetries between positive and negative judgments.

These results, taken together, cast severe doubts on the presumed symmetry between praise and blame, and, consequently, they challenge the adequacy of giving priority to the investigation of blame-related aspects of moral cognition. In fact, it seems both descriptively and explanatorily more adequate to derive the following two asymmetries:

Functional asymmetry: Blame and praise serve different social functions.

Psychological asymmetry: The cognitive processes underlying praise and blame judgments differ in important respects, such that they are sensitive to different factors.

Recently, a new line of research in experimental metaethics and moral psychology started to include praise and praise-related phenomena more systematically by investigating the language that is used to express positive and negative moral judgments. It is this line of research we now turn to and which provides additional reason to believe that blame and praise are systematically asymmetrical.

2.2. *Evaluative language and the polarity effect*

While moral *cognition* has been at the center of moral psychology, empirical research on the *language* that is used to express moral judgments is still in its infancy. Philosophers and, more specifically, metaethicists distinguish between two types of terms and concepts by which we can communicate moral content, namely, *thin* and *thick* ethical terms and concepts (e.g., Eklund, 2011; Roberts, 2013; Väyrynen, 2021; Willemsen & Reuter, 2021).[3] Thin ethical terms are rather abstract ways of expressing a speaker's positive or negative attitude and can, therefore, be used to speak positively or negatively about a person or their actions. Typical examples are "good", "bad", "right", and "wrong" those terms most frequently used in moral psychological research (see Abend, 2013 for an overview). In addition, thick terms also evaluate positively or negatively, but they further communicate descriptive content and are, therefore, more specific. For instance, the terms "cruel" and "dishonest" both imply negative evaluations and are more specific than, for instance, "bad". However, being cruel differs from being dishonest in various descriptive ways. Being cruel is about inflicting physical or emotional harm, whereas being dishonest is about lying or cheating.

Metaethicists aim to understand the nature of evaluative terms and by what linguistic means evaluative content is conveyed. In doing so, they make two critical assumptions. First, they at least implicitly assume that whatever theory of evaluative language is true for negative terms will also be true for positive ones. They presume—again—some kind of symmetry between positive and negative evaluative concepts, such that whatever we learn about terms of one polarity will also hold for the opposite polarity. Second, the standard view assumes that evaluative language semantically entails positive or negative evaluations. Calling an agent's behavior friendly or compassionate means, among other things, evaluating it positively; and calling a person rude or cruel entails a negative evaluation of them.

In a series of papers, both these assumptions have been challenged empirically (Baumgartner et al., 2022; Willemsen & Reuter, 2021). In an attempt to decide whether the evaluation of a thick term is semantically and pragmatically conveyed, Willemsen and Reuter (2020, 2021) employ the cancellability test for conversational implicatures (see, e.g., Grice, 1989, but see Zakkou, 2018 for discussions of the test's limitations). Here are two examples of the experimental stimuli they used:

(1) Negative: Tom is rude, but by that I am not saying something negative about Tom.
(2) Positive: Tom is friendly, but by that I am not saying something positive about Tom.

In both cases, participants were asked whether the speaker contradicts herself. If positive and negative terms work alike and the symmetry assumption holds, whatever contradiction ratings are found for negative statements should also be found for positive statements.

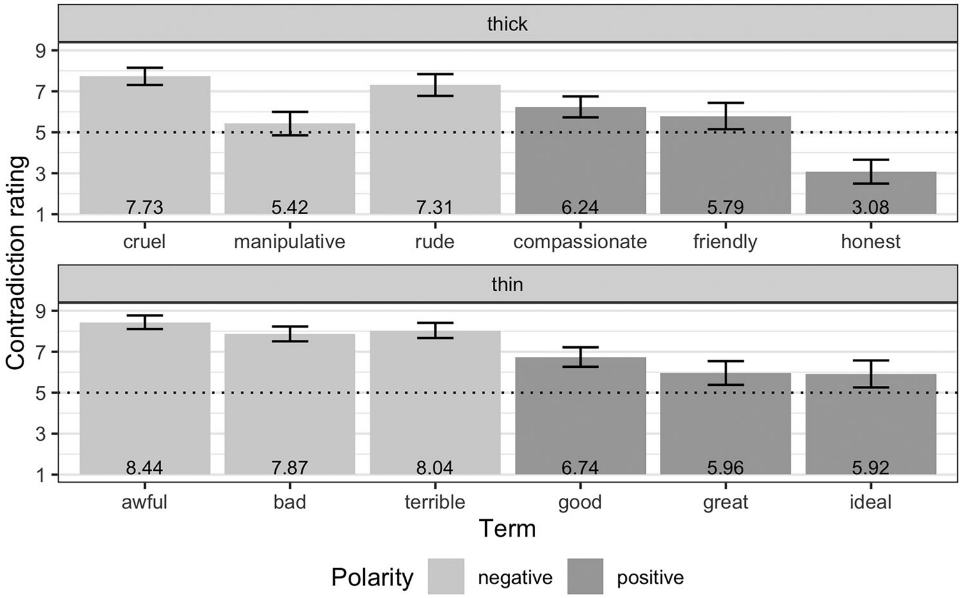


Fig. 1. Cancellability Ratings for a selection of thin and thick terms in Baumgartner et al. (2022).

However, negative evaluations were significantly harder to cancel compared to positive ones. Statements like (1) were judged to be significantly more contradictory than statements like (2). This demonstrated asymmetry, the polarity effect, seriously challenges the symmetry assumption.

Baumgartner, Willemsen, and Reuter (2022) examined the pervasiveness of the polarity effect and in which embeddings it occurs. The effect remains robust even when the scope of application is extended from one individual person to groups of people and even when we make generic statements like “People are rude.” Even more strikingly, the polarity effect seems to occur not only for thick concepts but for thin concepts, as well (see Fig. 1). This finding comes as a surprise. It is usually assumed that thin terms do not do anything beyond evaluating positively or negatively. As a consequence, there should not be much space for a speaker to say, for instance, “What Tom did was good/bad, but by that I am not saying something positive/negative about him.” Saying that what Tom did was good (or bad) *just is* saying something positive (or negative) about him.

How can the polarity effect be explained? The aim of the present paper is to provide a systematic explanation that can account for the polarity effect for thick and thin terms. We suggest that positive behavior is expected—we ought to do good things and have valuable character traits—but that negative behavior is not—we rarely expect anyone to be manipulative, cruel, rude, and so on. While there are *minimal expectations* of how good people need to be, there are *no corresponding expectations* of how bad they need (or are even allowed) to be.

2.3. Social expectations and the zone of moral indifference

The expectations we wish to consider are all *social* expectations in the sense that they relate to human behavior, and they are based on corresponding norms. Among them, we can distinguish two types. *Empirical* (also called *statistical*) *expectations* are based on statistical norms about how people are likely to behave within in a community. Empirical expectations refer to a person's subjective belief that a sufficiently large subset of their community conforms to the corresponding norm (in most relevantly similar situations). For instance, a person might know that in Mediterranean countries like Spain, people tend to eat dinner rather late. Once we learn that a colleague is Spanish and lives in Spain, we infer that they are also likely to eat dinner late. Therefore, we *expect* them to eat late. Violations of empirical expectations usually lead to surprise or w confusion.

In contrast, *normative expectations* are beliefs that a certain norm must or ought to be followed, and they are based on either social or moral norms of what ought to be done. Violations of normative expectations are often met with blame or punishment. Examples of social norms include traffic regulations, dress codes, or table manners.[4]

While distinguishing empirical and normative expectations provides conceptual clarity, in ordinary life, they usually go hand in hand. The fact that a community upholds certain normative expectations (in part by sanctioning violators) contributes to compliance becoming an empirical norm—normative expectations create empirical expectations (e.g., Bicchieri, 2006, 2014, 2017; Wysocki, 2020). For instance, the fact that people ought to drive on the left in England, but on the right in France, and that violations of this rule are heavily sanctioned contribute to the statistical expectation that this is what people are most likely to do. In reverse, the fact that most people of a community empirically expect others to behave in a certain way often suggests some deeper, normative expectation—we tend to infer an "ought" from an "is" (e.g., Bicchieri, 2006; Nichols, 2008; Roberts, Ho, & Gelman, 2019; Roberts, Guo, Ho, & Gelman, 2018). Some researchers even argue that empirical and normative expectations are so intertwined in ordinary life that what is considered "normal" is a function of what we believe to be statistically most likely and what we think ought to be the case (e.g., Bear, Bensinger, Jara-Ettinger, & Knobe, 2018; Bear & Knobe, 2017; Horvath & Nado, 2021; Wysocki, 2020).

Expectations on what *normal* behavior looks like is typically, so we argue, positively biased. We expect others to be honest and we do not expect them to deceive or betray us. We also expect others to reciprocate acts of generosity and to react to another person's distress with at least minimal compassion and decency. These positive expectations are necessary for a group's successful cooperation and long-term stability. We expect the members of a functioning social group to be minimally decent—not morally outstanding and particularly saint-like but ok. One might object that this is not always true. We might expect a pathological liar not to tell the truth or a particularly selfish colleague to try to get a free ride whenever they can. Sometimes we do expect bad things. However, note that for any social group to function properly, conforming to social rules, respecting others, and acting prosocially is the norm, and transgressions are rather the exception. Given our expectations, we can distinguish

Table 1
Relationship between expectations about an agent's actions, the desirability of the action, and whether or not the agent will be considered praise- or blameworthy for the action

Expectations	Desirability	Moral desert
violated	undesirable	blameworthy
met	desirable	
exceeded	desirable	praiseworthy

among three different types of morally relevant actions depending on how they relate to those expectations:

- morally undesirable actions that violate our expectations and for which the agent is *blameworthy*
- morally desirable actions that *meet* our expectations and for which the agent is *neither blameworthy nor praiseworthy*
- morally desirable actions that *exceed* our expectations and for which the agent is *praiseworthy*

It is important to note that expectations are trichotomous (violated, met, exceeded), whereas the desirability of an action (undesirable, desirable) is dichotomous, as is the agent's moral desert (blame, praise). Consequently, there cannot be a 1-to-1 correspondence between expectations and either the desirability of or the moral desert for a moral action (see also Table 1). A morally *undesirable* action almost always violates a moral expectation. The situation is more complicated, however, for morally *desirable* actions. On the one hand, we have desirable actions that exceed our expectations. These actions are praiseworthy, as they go beyond what we expect of others. On the other hand, there are moral actions that meet but do not exceed our expectations. These actions are not particularly praiseworthy, because, after all, they merely satisfy a level of minimal decency.

Moral philosophers have already recognized that acts of common decency seem to occupy, what Calhoun (2004, p. 129) calls, "a shadowy territory between the obligatory and the supererogatory." Some actions do not seem *morally* obligatory but we certainly expect them from one another, such as being grateful and thanking someone for a kindness, doing small favors to friends and family, or accepting an honest apology.[5] The inherent expectation becomes most visible when an agent fails to be grateful, do a small favor, or accept an honest apology, and we react with disappointment, anger, or frustration. However, while these actions are expected and their omission negatively recognized, we normally do not believe that being grateful requires positive recognition (for similar positions see, e.g., Chisholm, 1982; Driver, 1992; Feinberg, 1968; Stout, 2020). Calhoun (2004) (p. 130) states: "Common decency has to do with what can be expected from any minimally wellformed moral agent. To have common decency is to be a good or acceptable moral agent, but just barely." Note that the taxonomy we offer does not correspond to the classical philosophical distinction between *forbidden*, *permissible*, *obligatory*, and *supererogatory* actions (for an overview and discussion see, e.g., Archer, 2016; Heyd, 1982, 2019). Acts within the zone of moral indifference, those acts of common decency, are not merely permissible in the sense that they are neither good or bad to do or not to do. They are also not necessarily *morally obligatory*. Arguably, there is no moral obligation to do small favors. However, there is a strong social expectation that people do, especially to those with whom they are close. Our taxonomy is neutral with respect to the existence of moral obligation and offers a different perspective on moral actions that takes *social expectations* on how decent members of our society should act as its starting point.[6]

With this picture in mind, we can begin to explain the polarity effect of evaluative language. One might think that calling someone "generous" commits us to a positive, praising evaluation of this person, whereas calling someone "selfish" commits us to a negative, blaming evaluation. However, since evaluative language is more complex than this, such an inference would not be warranted. The use of a positive evaluative term wrongly suggests that we can *only* intend to positively evaluate in the sense of praising the agent, while, in fact, we might not intend to do so. An understandable but mistaken inference is made from the superficial features of positive evaluative terms to the underlying communicative intentions of the speaker (and, even further, to their underlying cognitive states). Instead, we argue, positive terms can be and are often used to communicate that a person or their action merely meets our expectations, but is not deserving of praise for doing so. This is what we call the *evaluatively deflated* use of positive terms.

In the next three sections, we provide empirical evidence for the main claims we developed in this section. More specifically, in Section 3, we provide empirical evidence (Study 1) that positive terms can be used in both an evaluative as well as a deflated manner. In Section 4, we present the results of an experiment (Study 2) showing that social expectations seem to play the defining role in determining whether positive terms are used in a full-blown evaluative or deflated sense. In Section 5 (Study 3), we show that manipulating whether an agent merely meets or exceeds an expectation impacts the assessment of the evaluative nature of positive terms.

3. Study 1: Defeasibility

The account we propose claims that positive evaluative terms can be used in an evaluatively deflated way, while negative evaluative terms cannot be used in this way. The results presented in Willemsen and Reuter (2020, 2021) and Baumgartner et al. (2022) can only count as tentative evidence. In these studies, the speaker utters a statement including a thin or thick term and then explicitly takes back the corresponding evaluation.

Why do we believe that this evidence is only tentative? It might be argued that in previous studies, we only demonstrate that the standard evaluation of a thick concept can or cannot be cancelled: The positive evaluation usually communicated by honesty or friendliness can be cancelled, and the negative evaluation usually communicated by selfishness or cruelty cannot. However, this does not prove that positive terms can be used *neutrally*. People sometimes communicate disapproval despite using standardly positive evaluative terms. This can happen when a person thinks that another person's behavior was good but not good *enough* or *too* honest. In such cases, it might be objected, thick terms are not used neutrally but the speaker intends to change the polarity of the statement from positive to negative. Critically, this polarity change is restricted to statements featuring positive terms only. Such a reversal would violate a key assumption of our theory, namely that positive terms can be used in an *evaluatively deflated* sense. Therefore, we need to rule out that the evaluation was reversed instead of cancelled.

In the philosophical literature, the idea that (some) thick concepts can be used nonevaluatively or in a neutral sense is known as the defeasibility thesis see, for example, Väyrynen (2021).[7] If positive terms can be used in a deflated way, the polarity effect would still occur under the defeasibility paradigm in which the speaker explicitly says that they intend to use the term in a fully neutral way. In contrast, if positive terms can change their polarity to a negative meaning, then no such effect should be salient. Thus, if the polarity effect occurs, this would not only suggest that the polarity change explanation is on the wrong track, but it would also indicate that positive terms (in contrast to negative terms) can be used in *two different* manners: A truly evaluative use, on the one hand, and a deflated use, on the other hand. To investigate the potential defeasibility of the evaluative content of evaluative terms, we used a variation of the cancellability test, which makes it explicit that the evaluative term is supposed to be used in a neutral way. We recorded contradiction ratings for what we call defeasibility statements:

> “What Tom did last week was [term, e.g., courageous], but by that I am not saying something positive or negative about his behaviour that day. I mean this in a fully neutral way.”

Study 1 was designed to investigate three hypotheses:[8]

Defeasibility Hypothesis (H1): Contradiction ratings for defeasibility statements with thick concepts are not significantly above the midpoint of the rating scale (= 5).

If we can reject H1, that is, the average contradiction ratings are above the midpoint, thick concepts do not seem to be defeasible, but instead seem to always communicate an evaluative attitude. While H1 tests a a more general hypothesis about thick concepts, we also conduct a more fine-grained analysis taking into account the polarity of thick concepts. Indeed, previous studies have shown that the evaluative component of positive terms like “‘honest’” is more easily cancelled than the evaluative component of negative thick terms like “‘rude.’” Based on these results, we also investigate a second hypothesis:

Polarity Hypothesis (H2): Average contradiction ratings for defeasibility statements featuring positive terms are significantly below contradiction ratings of those featuring negative terms.

Contrary to H1, H2 is not restricted to thick concepts. Rather, we want to investigate whether the polarity effect persists even if we pool together thick and thin concepts. Additionally, previous studies have shown (see Section 2), that while the polarity effect also occurs for thin concepts, the contradiction ratings are higher for thin terms compared to thick terms. We, therefore, also examined whether this difference holds for defeasibility statements:[9]

Thick-Thin Hypothesis (H3): There is no significant difference in the average contradiction ratings for defeasibility statements between thick and thin terms.

3.1. Methods

Our sample consisted of 700 native English speakers, all over 18 years old, with an approval rate of at least 90% on Prolific. The average age in the sample was 34.42 years, and the gender representation was 36.57% male, 62.57% female, and 0.86% nonbinary. Before participants gave their contradiction ratings, there was a training round giving participants instructions on how to understand what a contradiction is, including two test questions.[10] After the training round, participants were randomly assigned to a single defeasibility statement featuring a term from one of the following five categories:[11]

- Thick positive: compassionate, courageous, friendly, generous, honest
- Thick negative: cowardly, cruel, manipulative, rude, selfish
- Descriptive: pragmatic, ordinary, conventional, expected, coordinated
- Thin positive: good, great, ideal
- Thin negative: bad, awful, terrible

After being presented with one of the defeasibility statements, participants were asked: "Does [subject] contradict herself/himself?". Answers were given on a Likert scale from 1 = "definitely not" to 9 = "definitely yes." For each term, we collected 33 data points on average, leading to 165 data points for descriptive concepts, 333 for thick concepts (166 positive, 167 negative), and 201 for thin concepts (100 positive, 101 negative). Five hundred and forty-three participants passed both training test questions. We performed analyses for the whole set of responses as well as for the subset of responses from participants who passed both test questions (as preregistered).

3.2. Results

The average responses for the four evaluative conditions, as well as the descriptive concepts, are depicted in Fig. 2. We performed a general ANOVA with *type* (thick/thin) and *polarity* (positive/negative) as independent variables, and *contradiction ratings* as dependent measure. The independent effects for term *type* ($F(526) = 18.850,\ p < .001, \eta^2 = 0.027$) and *polarity* ($F(526) = 54.908,\ p < .001,\ \eta^2 = 0.093$) were both significant, but their interaction was not ($F(526) = 2.994,\ p = .084, \eta^2 = 0.004$), on a 0.05-alpha level. Contrary to the Defeasibility Hypothesis (H1), a one-sample Wilcoxon test ($V = 28, 898,\ p < .001$) showed that the average defeasibility ratings for thick terms (both positive and negative) were significantly *above* the midpoint of the scale (0.01-alpha level). With $r = .191$ (effect size), the probability to reject H1, that is, $1-\beta$, is 87.48%.[12] Given the effect of *polarity* on the cancellability ratings, we analyzed the mean contradiction ratings separately for positive and negative terms.

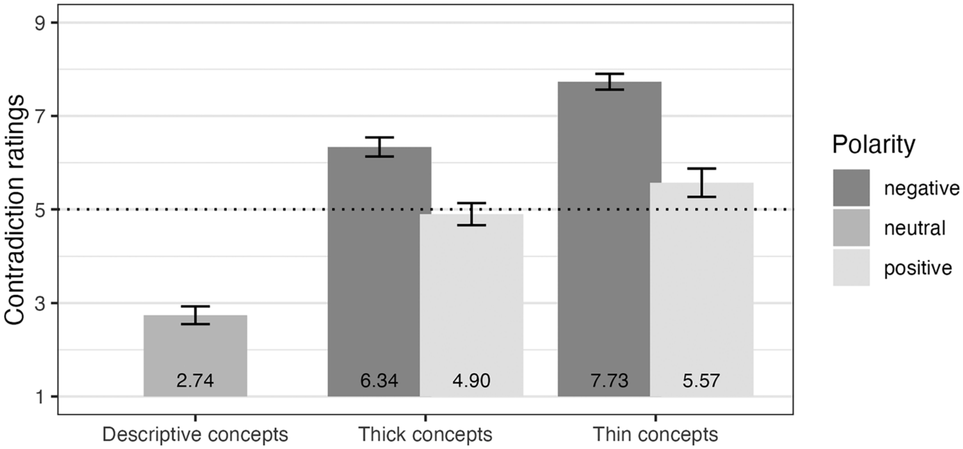


Fig. 2. Defeasibility ratings for thick and thin terms.

While for negative thick terms alone, the average rating was significantly above the midpoint ($V = 8778.5,\ p < .001,\ r = .4399,\ 1 - \beta = 99.95\%$), the mean for positive thick terms was not significantly above the midpoint of 5 ($V = 5723.5,\ p = .764,\ r = .0487,\ 1 - \beta = 4.43\%$).[13]

Regarding H2, a one-sided Welch Two Sample t-test ($t(505.78) = 7.163,\ p < .001$) showed that positive terms indeed have significantly lower contradiction ratings than negative terms, on a 0.01-alpha level. The probability to reject the null hypothesis on 0.01-alpha level (i.e., negative and positive terms have equal ratings) in favor of H2 is 99.99% (Cohen's $d = 0.6194$). Hence, our data support the H2.[14]

Investigating hypothesis H3, we found that thick concepts have a lower average rating ($\mu = 5.61,\ \sigma_\mu = 0.162$) compared to thin concepts ($\mu = 6.66,\ \sigma_\mu = 0.189$). A Welch Two Sample t-test yielded that this difference was significant, $t(453) = -4.187,\ p < .001$, suggesting that H3 (i.e., the null hypothesis) should be rejected.[15] The probability to reject H3 is 93.92% on 0.01-alpha level (Cohen's $d = 0.3693$).

Fig. 3 displays the distribution of the responses from all participants for negative, descriptive, and positive terms. For descriptive and negative terms, the distributions have rather unimodal shapes. However, when it comes to positive terms, the distribution reveals a bimodal character with the endpoints of the scale being most frequently selected.

The bimodal distribution for positive terms is further detailed in Fig. 4. While the mean rating for positive thick concepts was 4.90 and for thin concepts 5.57, only a few participants gave ratings of "4," "5," or "6." Instead, the most frequent answers were "1" and "9," suggesting that participants either considered the statements presented in Study 1 highly contradictory, or not contradictory at all.

3.3. Discussion

The aim of Study 1 was to provide support for a crucial assumption of the theory we propose, namely that positive terms can be used in an evaluatively deflated way. The results from

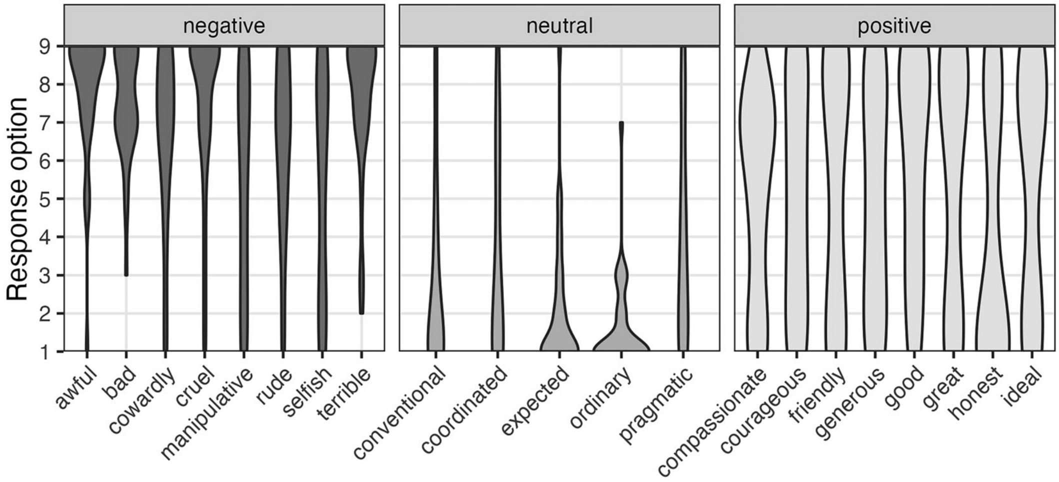


Fig. 3. Violin plots revealing the response distribution of negative, descriptive, and positive terms.

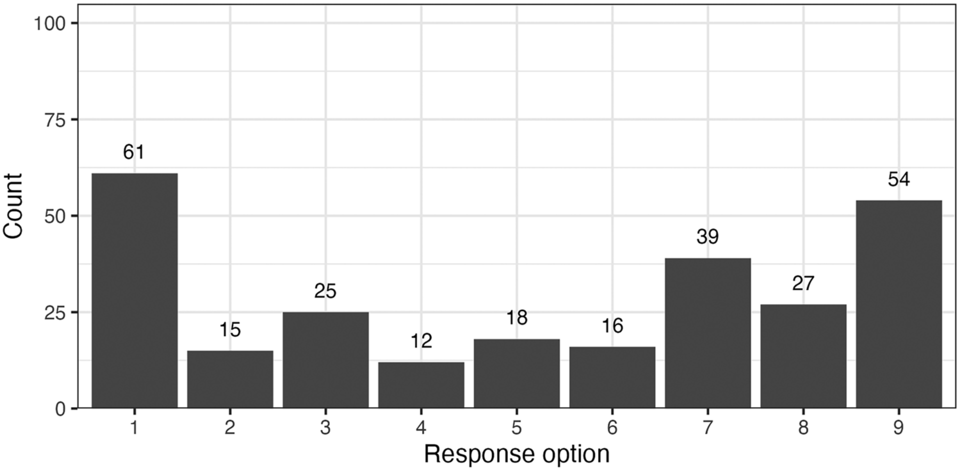


Fig. 4. Distribution of responses for positive terms.

Study 1 indeed suggest that positive terms behave differently compared to negative terms. First, defeasibility ratings are significantly lower for positive terms compared to negative terms. Second, the average defeasibility ratings for positive terms are not significantly above the midpoint of the rating scale. And third, when we look at the distribution of the responses for positive terms, we find a bimodal distribution. From this, we infer that there are two different ways in which positive terms may be used: a full-blown evaluative as well as a deflated use. In contrast, negative evaluative terms can only be used to disapprove, with no deflated use possible. It is this difference that we will explore further in the next section.[16]

4. Study 2: Expectations and approval

Study 1 indicates that positive evaluative terms can have an evaluatively deflated use. We propose to explain this phenomenon by arguing that most people expect others to behave in a minimally decent way, that is, to be minimally honest, generous, friendly, and so on. When a person's behavior fulfills but does not exceed these expectations, we can refer to such behavior by using positive terms in a deflated sense, that is, without intending to say something positive or even praising that person's behavior.

Following this reasoning, we designed a second experiment to investigate whether and to what extent negative forms of behavior are less expected of others and constitute deviations from the norm, while positive behavior is more strongly expected and less of a deviation from the norm. To test this general idea, we made the following empirically testable prediction:[17]

Expectation Hypothesis (H4): There is a significant effect of the polarity of an evaluative term (positive vs. negative) on expectation ratings, that is, the mean absolute distance from the midpoint ("exactly what I expect of him") is significantly higher for negative evaluative concepts compared to positive evaluative concepts.

At the same time, we believe that participants can use evaluative terms in a full-blown evaluative manner, that is, in order to blame or praise a person. Consequently, we would not expect any difference between positive and negative terms when it comes to using these terms in order to approve or disapprove of people. Thus, concerning the dependent variables Approval and Expectation, we made the following prediction:

Double Use Hypothesis (H5): There is a significant interaction between Polarity (good vs. bad) and the dependent variables (Expectation vs. Approval).

4.1. Methods

Here is how the experiment was presented to the participants:

> Please think about what kind of behaviour you expect of people in general.
>
> Please consider the following statement:
>
> "What Tom did was [evaluative term]."
>
> To what degree is Tom's behaviour below your expectations, exactly what you would expect of him, or exceeds your expectations?

Expectation ratings were measured on a 9-point Likert scale anchored at 1 = "strongly below my expectations," 5 = "exactly what I expect of him," and 9 = "strongly exceeding my expectations." Next, participants were asked:

> Next, we would like to know:

> How strongly do you disapprove or approve when people do something [evaluative term]?

Approval ratings were measured on a 9-point Likert scale anchored at 1 = "strongly disapprove," 5 = "neither approve nor disapprove," and 9 = "strongly approve." As stimuli, we used the same evaluative terms as in Study 1, and further added the evaluative terms "virtuous," "vicious."

- Thick positive: compassionate, courageous, friendly, generous, honest, virtuous
- Thick negative: cowardly, cruel, manipulative, rude, selfish, vicious
- Thin positive: good, great, ideal
- Thin negative: bad, awful, terrible

All our hypotheses addressed differences between the categories into which these 18 terms fall, not the specific terms themselves. Participants were randomly assigned to one of the 18 terms (falling into one of four between-subjects conditions).

In order to test our two hypotheses, we needed to transform our data by calculating the difference of the responses from the neutral midpoint. Expectations ratings are given on a 9-point scale, with 5 indicating that the agent acted just as participants expect. We predicted that whenever a negative term is used, that behavior deviates strongly from the neutral midpoint. With a positive term, on average, less of a deviation is signaled. Thus, instead of comparing (say) the absolute Expectation ratings for thick negative terms and thick positive terms, we calculated the differences of people's ratings from the neutral midpoint of 5. In a next step, we calculated the means of those difference ratings. All further analyses were conducted based on these differences, which we call Δ Expectation and Δ Approval.

4.2. Results

We collected responses from 352 participants.[18] After excluding participants who did not finish the survey, did not consent, or failed an attention test, we were left with 340 responses ($\mu_{\text{age}} = 32.28$; 29.42% male, 70.00% female, 0.58% nonbinary).

To test hypothesis H4, we ran a one-sided Wilcoxon test ($W = 19,032,\ p < .001$) with Polarity (positive vs. negative) as the independent variable and Δ Expectation as the dependent variable. The mean absolute distance from the midpoint was significantly higher for negative terms ($\mu = 1.66,\ \sigma_\mu = 0.14$) compared to positive terms ($\mu = 0.91,\ \sigma_\mu = 0.10$), as predicted by H4 (0.05-alpha level). The probability to reject the null hypothesis (i.e., the absolute distance to the midpoint is identical for positive and negative terms) in favor of H4 is 95.71% (Cohen's $d = 0.468$). Hence, H4 is supported by our data. To test H5, we ran a nonparametric two-way ANOVA of our Aligned Rank Transformed (ART) data (Elkin, Kay, Higgins, & Wobbrock, 2021; Wobbrock, Findlater, Gergle, & Higgins, 2011).[19] We used the pooled delta values as the independent variable, and the interaction between Polarity (positive vs. negative) and Question (Expectation vs. Approval) as independent variables. Since Question is a within-subjects variable, we used a mixed effects model for the ANOVA.[20] The reason we used ART is that the assumptions for factorial mixed model ANOVA are violated. The

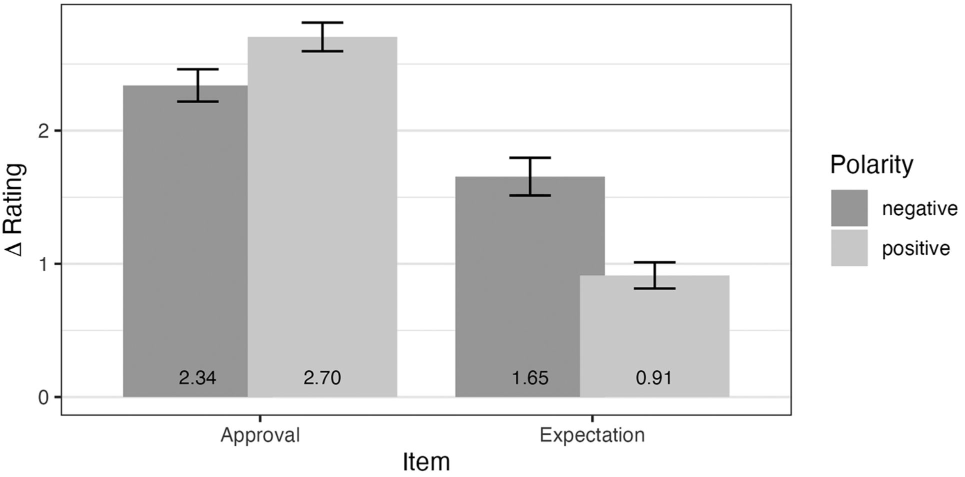


Fig. 5. Means for Δ Expectation and Δ Approval (i.e., the absolute deviation from the midpoint) by Polarity. The Expectation items reflect the Polarity Effect, whereas the Approval items do not.

effects for Question ($F(338) = 166.681,\ p < .001$), Polarity ($F(338) = 7.395,\ p = .006$), as well as the interaction between Question and Polarity ($F(338) = 43.2552,\ p < .001$) were all significant on 0.01-alpha level.[21] The probability to reject the null hypothesis (i.e., there is no interaction effect) is 99.45% (Cohen's $f = 0.281$). Hence, we cannot reject H5, since the interaction term was significant. We also ran pairwise contrasts on the estimated marginal means to investigate the differences between level interactions. We were especially interested in whether people approve of positive behavior to the same extent as they disapprove of negative behavior. This is indeed the case: there is no indication of a Polarity Effect in the Approval items (contrary to the Expectation items tested with H3), as the average difference between positive and negative items is not significant ($t(319) = -2.246,\ p = .025$) on a 0.01-alpha level. Fig. 5 displays the average observed deviation from the midpoint (= 5) for both the Expectation and Approval ratings.

4.3. *Discussion*

The results of Study 1 suggest that positive terms can be used in a deflated way, on the one hand, as well as to communicate full-blown evaluations, on the other hand. In contrast, negative terms only allow for a strongly evaluative use. In Study 2, we tested this account by asking participants two questions, each of which targeted a different use of evaluative terms. The expectation question aimed to elicit a deflated use of evaluative language, whereas the approval question aimed to trigger an evaluative use. Based on the design of the experiment, we predicted different outcomes for both questions depending on the polarity of the term.

People indeed seem to use positive evaluative language in two different ways. We expect others to behave in a minimally decent way—to occupy the zone of moral indifference—which is to say that we expect people to be minimally decent, honest, and friendly, and we do not see these as particularly praiseworthy behaviors. Of course, once we move from the zone of moral indifference into the zone in which people's behavior exceeds our expectations, approval is warranted, and positive thin and thick terms are used to praise such behavior.

5. Study 3

The results of Study 1 suggest that positive evaluative terms can be used in two different ways: a proper evaluative use in which a positive attitude is expressed, and an evaluatively deflated use that primarily serves to describe a certain state of affairs. Study 2 has demonstrated that behavior described with negative terms like "rude" deviates strongly from people's expectations, whereas behavior described with positive terms such as "friendly" is close to what people would expect of them. Consequently, the two different ways of using positive terms might be accounted for by the way in which expectations are met (deflated use) or exceeded (evaluative use).

Study 3 tests whether the different uses of positive terms actually depend on whether everyday expectations are merely met or clearly exceeded.[22] Study 3 focuses only on positive terms, because it tests an explanation for the distinct patterns of use revealed by the two previous studies. As it is unclear whether positive thick concepts explicitly express praise or merely evaluate positively, we tested for both positivity as well as praiseworthiness. We expect the following hypothesis to hold:

Deflated Use Hypothesis (H6): In Expectations Met conditions, Positivity and Praiseworthiness ratings will be significantly lower compared to the corresponding Expectations Exceeded conditions.

The idea behind H6 is that a positive term expresses a genuinely positive evaluation whenever expectations are clearly exceeded. Whenever expectations are merely met, the positive term is significantly less evaluative, that is, more evaluatively deflated.

5.1. Methods

In this experiment, we investigate the use of the terms "friendly" and "compassionate", which were used already in the two previous studies. Participants were assigned to either Term condition (between-subjects independent variable). We presented subjects with two scenarios each, one in which everyday expectations are merely met, and one in which they are clearly exceeded (within-subjects independent variable). Here is the vignette for "friendly":

> Eric and Andrew work for the same company. In order to improve the work climate and to make people feel appreciated, the company expects their employees to show collegial behaviour.
>
> One morning, Eric meets his colleague Andrew in the hallway. Eric says, "Good morning." Andrew clearly heard Eric and realised that Eric was talking to him.
>
> Met: Andrew looks and responds: "Good morning."

> Exceeded: Andrew smiles and responds: “Good morning. I heard that you will give an important presentation later today. Best of luck. I’m sure you’ll do great.”
>
> Another colleague, John, saw Eric and Andrew’s interaction and says to Eric, “What Andrew did was friendly.”

After reading the vignette, the participants were presented with either of these questions (between-subject dependent variable):

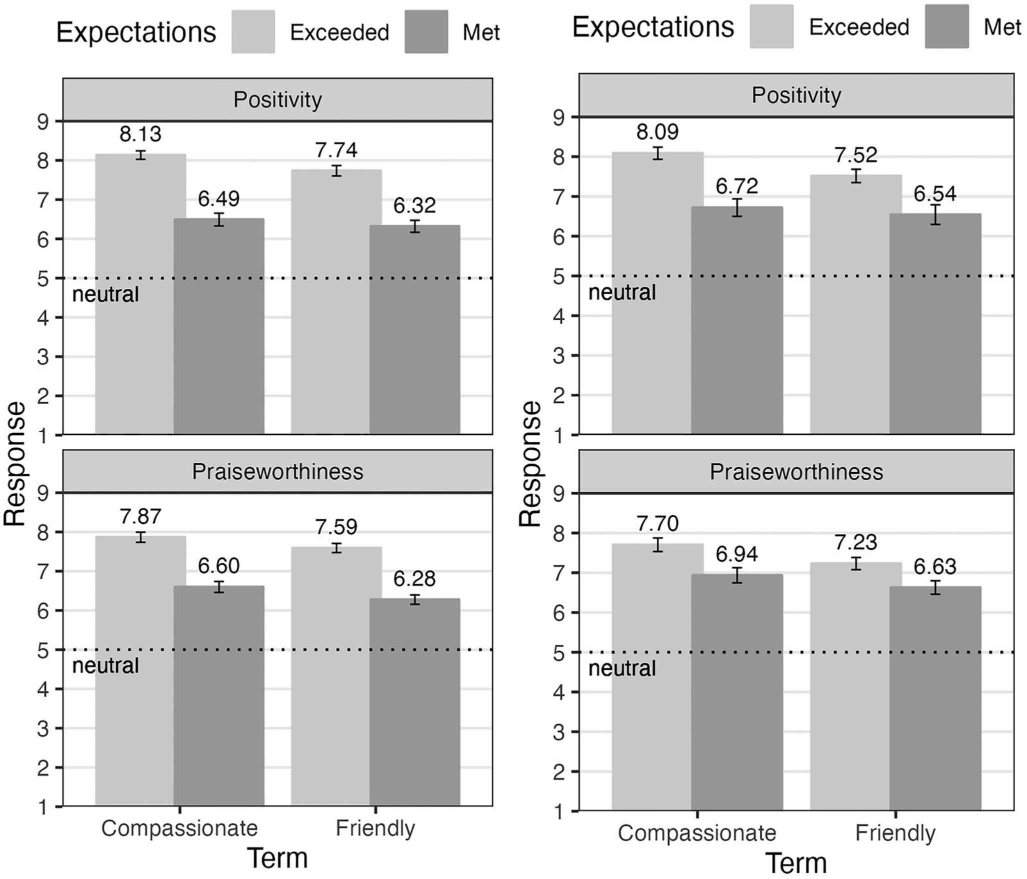
> In your opinion, how is the term “friendly” used by John in regards to Andrew’s behavior?
>
> Positivity: Is it used very negatively, neutrally, or very positively (or something in between)?
> Praiseworthiness: Is it used to express a lot of blame, neutrally, or used to express a lot of praise (or something in between)?

The responses were recorded on a corresponding 9-point Likert scale, anchored at “1 = very negatively [a lot of blame],” “5 = neutrally [neutrally],” and “9 = very positively [a lot of praise].” The two dependent variables (Positivity and Praiseworthiness) reflect different conceptions of the evaluative component of thick concepts.

5.2. Results

Our sample comprises 505 participants, with an equal share of female and male participants. We did not collect additional demographic data for this study.

For H6, we ran two 2×2 repeated-measures ANOVAs to test for the effects of Term (“‘friendly’” vs. “‘compassionate’”) and Expectation (Met vs. Exceeded). The analysis was performed separately for the two dependent variables (Positivity and Praiseworthiness). For Praiseworthiness, there is a significant effect for the within-subject condition Expectation ($F(252) = 195.382,\ p < .001$). The effects for the between-subjects variable Term ($F(252) = 3.84,\ p = .051$) and the interaction of Term and Expectation ($F(252) = 0.065,\ p = .798$) are not significant. The probability to reject the null hypothesis (i.e., there is no difference between Expectation Met and Expectation Exceeded conditions) in favor of H6 is $> 99.9\%$ (Cohen’s $f = 0.8805$). The results look similar for Positivity: The effect for Expectation is significant ($F(249) = 193.674,\ p < .001$), whereas the effects for Term ($F(249) = 2.989,\ p = .0851$) and the interaction ($F(249) = 1.065,\ p = .303$) are not. The probability to reject the null hypothesis in favor of H6, here again, is $> 99.9\%$ (Cohen’s $f = 0.8822$). Hence, we cannot reject H6. Moreover, the odds of choosing the neutral value of 5 were 4.606 times higher in the Expectations Met condition compared to the Expectation Exceeded condition. This, again, indicates a more neutral use of thick terms when expectations are merely met but not exceeded. We also tested whether the effects appear in a pure between-subject design as well. For this, we ran two 2×2 ANOVAs



(a) Participants' responses to both items (N = 1010)

(b) Participants responses to the first item only (N = 505)

Fig. 6. Average expressed Positivity/Praiseworthiness of positive thick Terms, for both Expectation conditions.

with the responses to the first question only. The effect for Expectation on Praiseworthiness ($F(250) = 16.042,\ p < .001$) and on Positivity ($F(247) = 35.399,\ p < .001$) indeed remains significant. Moreover, two one-sided t-tests show that the mean is significantly above the neutral midpoint of 5 for both Praiseworthiness ($t(507) = 29.985,\ p < .001$) and Positivity ($t(501) = 27.791,\ p < .001$). Fig. 6 shows the average responses collected in Study 3. Fig. 6a depicts the pooled responses for both Expectation conditions (Exceeded and Met), whereas Fig. 6b only shows the responses to the first item presented to the participants.

5.3. *Discussion*

With Study 3, we investigated whether the two hypothesized uses of positive thick terms (deflated and evaluative use) are dependent on everyday expectations. Study 3 shows that the evaluation of positive thick terms is indeed less pronounced when characterizing behavior that merely meets social expectations, compared to behavior that exceeds those expectations. Hence, everyday expectations can (partially) explain different uses of terms like “compassionate” and “friendly.”

Various strategies can be employed to elicit distinct expectations, such as (i) altering the magnitude of the agent’s actions, (ii) modifying the frequency of specific behaviors exhibited by the agent, or (iii) adjusting the agent’s social standing or reputation. Each of these manipulations introduces a potentially problematic confound. We decided to trigger differential expectations by manipulating the magnitude of the acts that the agent performed, as we believe this method offers the most straightforward means of manipulation. Admittedly, we cannot rule out that these manipulations themselves are responsible for the difference in the results. Thus, further empirical studies are necessary to robustly establish a more direct connection between expectations and ratings of praise. [23]

6. General discussion

6.1. Summary of the empirical studies

Terms like “friendly”, “good”, “rude”, and “bad” are standardly considered to communicate evaluative content. Previous studies have shown that the strength with which such evaluative content is tied to these terms depends on their polarity. While it is relatively easy to use a positive term and deny the intention to communicate a positive evaluation, it is significantly harder—some would say impossible—to use negative terms in this way. How can we explain this so-called polarity effect of evaluative language?

In this paper, we developed the following account to explain the polarity effect: Acts that are considered undesirable standardly violate our expectations. In contrast, acts that count as morally desirable can either *meet* our expectations or they can *exceed* our expectations. The zone in which an act can be morally desirable, yet not exceed our expectations, is what we call the zone of moral indifference. The polarity effect emerges because people can use positive terms in a deflated manner to refer to actions in the zone of moral indifference, whereas negative terms cannot be so interpreted.

In three studies, we provided empirical evidence for this view. In Study 1, we demonstrated that positive terms can indeed be used in a deflated sense: The evaluation of positive terms is often considered defeasible. The evaluation of negative terms, in contrast, turns out to be nondefeasible: negative terms are exclusively interpreted to be evaluative, at least when used in a literal, assertive way. In Study 2, we investigated this deflated use more closely and examined the role that expectations play. Our data revealed that behavior referred to by negative terms violates our expectations more strongly than behavior referred to with positive terms. At the same time, though, positive behavior is approved of to the same extent that negative behavior is disapproved of. These results provide direct evidence for a robust asymmetry of moral language, and for the existence of a zone of moral indifference. The aim of Study 3 was to test more directly, whether manipulations of descriptions of the way agents merely meet or exceed expectations will have an impact on the assessment of the evaluativity of thick terms. The results of Study 3 reveal—for the thick terms we tested—that if an agent merely meets an expectation, most people will consider statements of the form "[Agent] is friendly/compassionate" to be neutral or close to neutral. In contrast, if an agent exceeds an expectation, most people consider the same statement to express a positive attitude and praise to a much greater extent.

6.2. Evaluative deflation and social norms

Positing a zone of moral indifference—a set of actions to which deflated uses of positive terms refer—is not the only way to explain the polarity effect. In fact, Willemsen and Reuter (2020, 2021) have proposed a different account which is very much in line with the functional asymmetry proposal put forward by Anderson et al. (2020). Accordingly, while openly evaluating a person positively does not usually have grave consequences, negative evaluations will often have a severe impact on a person's reputation: for instance, it can diminish other people's willingness to cooperate with that person. Given the higher stakes of negative evaluation, a person using a negative evaluative term had better make sure they express themselves clearly. A person who uses a term that usually conveys a negative evaluation, but who does not intend to evaluate, fails to convey her intentions accurately and thereby risks harming a person's reputation. While using positive terms in a nonevaluative manner might be equally nonstandard, the damage (if any) is likely to be far less severe.

While this explanation is different from the explanation proposed in this paper, it might help us to understand why negative terms do not have a deflated sense the way positive terms do. Negative terms, we have argued, usually communicate that expectations have been violated and that the agent has fallen below our standard of minimal moral decency. This is, as Willemsen and Reuter argue, a serious accusation that can cause the kind of social damage that philosophers and psychologists have already alluded to. As a consequence, it makes sense that we have established stricter norms for when and how negative evaluative language may be used. For positive terms, no such strict norms are needed.

The account proposed in this paper is, however, superior in providing a more promising explanation for the defeasibility of positive terms, and, more specifically, for the distribution of the responses we received in Study 1. If the social norms explanation were correct, then we would expect people in general to find it more acceptable to cancel the evaluation of positive terms compared to negative terms. Such a shift in acceptability should be a matter of degree only, and hence, we would expect a normal distribution of the responses around a lower average value. However, this is not what we found. Responses for positive terms in Study 1 were bimodally distributed, with very few participants giving ratings around the midpoint of 5 (see Fig. 4). Positing two different uses of positive terms—a deflated nonevaluative use and a fully evaluative use—can account much better for this distribution and explains not only the difference in average responses but also the bimodality of the distribution for positive terms. We, therefore, consider the account we propose in this paper not only explanatorily but also empirically superior.

6.3. Limitations and outlook

We would like to end this paper with a discussion of potential limitations of our research, possible connections to other strands of research in moral psychology, and also point toward some directions for future research.

The studies here presented use samples from English native speakers in Australia, Canada, Ireland, New Zealand, the United Kingdom, and the United States of America. The sample is, therefore, limited in two ways. First, we only have data based on English stimuli presented to English native speakers. It is so-far unclear whether the polarity effect would hold in other languages as well. Second, even our sample of English native speakers is biased toward WEIRD (Western, Educated, Industrialized, Rich and Democratic) native speakers, excluding, for instance, English native speakers in African, Asian, and Caribbean countries. If, as we have argued, the polarity effect is best explained by differences in norms, cultural differences could play an important role. Social groups are defined not only by differences in moral norms, but also by how competing moral norms (say, honesty and friendliness) are weighed against each other, the conditions under which moral norms apply (e.g., when it is appropriate to demonstrate generosity and to whom), and, importantly, what one can say to whom and how. All these factors may affect whether the evaluation of moral terms can be cancelled. A large-scale cross-linguistic and cross-cultural study is yet to be conducted to determine the robustness of the polarity effect.

One might also wonder if an alternative or, perhaps, complementary explanation holds: In our study, we only discussed moral expectations concerning how an agent should or should not behave, and the speaker's evaluation of that behavior. However, participants may not only have moral expectations about the *agent*, but further about how the *speaker* should talk about the agent. Social groups may vary greatly on when and how speakers can evaluate others, including whether taking the evaluation back is possible. The interplay of morality and politeness and authority norms, including cultural differences, is a topic for future investigations.

A further potential limitation of our research concerns the extension of the effect. So far, our main focus lies on evaluative terms in the moral domain. While the Polarity Effect is insensitive to changes regarding the terms or experimental design we used, one might wonder whether it is limited to *morally* evaluative terms. Philosophers further distinguish epistemic terms, for example, "open-minded," "intelligent," "unconvincing," "gullible," and aesthetic terms, for example, "beautiful," "tasty," "asymmetric," or "incoherent." One might also add evaluative terms used to describe more general competences, such as "skilled," "athletic," "untalented," or "clumsy." So far, we do not know whether the Polarity Effect extends to these nonmoral concepts as well, and if there are even cases where the Polarity Effect is reversed, such that *negative* terms have an evaluatively deflated use.

While many details of the picture are still unknown, it seems undeniable that both the Symmetry and the Priority of Blame assumptions are untenable. The Polarity Effect adds an exciting piece of evidence to the literature that blame and praise, or negatively and positively valenced phenomena, lead to asymmetrical judgments, expressed in an asymmetrical way.

Acknowledgements

Pascale Willemsen's research was funded by the Swiss National Science Foundation (SNSF), project number PZ00P1_201737. The research of Lucien Baumgartner, Kevin Reuter, and Pascale Willemsen was also supported by the Swiss National Science Foundation, grant number PCEFP1_181082. Bianca Cepollaro gratefully acknowledges the support of the PRIN project The Mark of Mental (MOM) 2017P9E9N at Vita-Salute San Raffaele University, Milan. Further, we would like to express our gratitude to Severin Frohofer, Walker Siu Hong Au, Ori Friedman, Ethan Landes, Judith Martens, Isidora Stojanovic, Jan Garcia Olier, and three reviewers of this journal for providing feedback on an earlier draft of this manuscript.

Open access funding provided by Universitat Zurich.

Open Research Badges

This article has earned Open Data, Open Materials and Preregistered Research Design badges. Data, materials and the preregistered design and analysis plan are available at https://osf.io/5y8xz/, https://osf.io/u96td/.

Notes

[1] It seems natural to think and well-established in the philosophical and linguistic literature that evaluative language can be of negative or positive polarity. Thus, those evaluative terms whose standard evaluation is negative are usually labeled "negative terms" and those terms with a positive standard evaluation "positive terms." In order to be consistent with the existing literature, we adopt the same labeling throughout this paper. However, whether so-called "positive" terms indeed evaluate positively is the subject of our investigation.

[2] In this paper, we restrict ourselves to *morally* evaluative terms, that is, those used to express moral values. We acknowledge that there are other evaluative terms which express, for example, epistemic or aesthetic value judgments, and discuss a possible extension of our results to other domains in the General Discussion.

[3] It has recently been suggested that we further need to distinguish value-associated concepts and dual character concepts. Value-associated concepts are descriptive concepts that are positively or negatively charged because people tend to associate positive or negative things with them, for instance, "rainy", "moldy", "rich", or "active" (for a discussion, see Reuter, Baumgartner, & Willemsen, 2023). Dual character concepts are distinct, as they have independent descriptive and normative components that are double-dissociable, such as "father", "scientist", and others (Del Pinal & Reuter, 2017; Knobe, Prasada, & Newman, 2013; Reuter, 2019).

[4] When we talk about *expectations*, we refer to mental states of an individual with the corresponding norm as its content. This content can be an empirical, social, or moral norm, or even a mix thereof. Speaking of expectations rather than norms has several

advantages. Norms can exist independently of an individual knowing of or endorsing it. However, only a norm that affects an individual's expectation will, in turn, have an effect on this person's judgments.

[5] Heyd (2019) offers some examples and refers to "small acts of favor, politeness, consideration and tact, which are good though not morally praiseworthy, which can be expected of people even though not strictly demanded."

[6] The empirical evidence on the relationship between supererogatory acts and obligations is rather scarce. Tomasello (2020) yet suggests that the concept of obligation is essential to both children and adults across a variety of cultures. Most studies investigating supererogatory acts have a strong developmental perspective, such as Dahl, Gross, and Siefert (2020), Kahn (1992), Khan, Jaffer-Diaz, Najafizadeh, and Starmans (2023), Marshall, Wynn, and Bloom (2020); Marshall et al. (2022), and Miller, Bersoff, and Harwood (1990).

[7] Note that the deflated use we postulate is weaker than the defeasibility thesis. We do not require a term to be fully neutral to be evaluatively deflated. An evaluatively deflated term might communicate some evaluation or some more general normative content, but this evaluation will be significantly less pronounced.

[8] The experimental design, predictions, and statistical models were pre-registered with the Open Science Framework: https://osf.io/n2uqh

[9] We preregistered an additional null hypothesis in regards to differences in contradiction ratings for thick and descriptive concepts. As this hypothesis does not pertain to the questions at issue, we limit ourselves to H1, H2, and H3. Note also that we labeled the hypotheses in a different order compared to the preregistration.

[10] The whole survey as well as the description of the training round can be found here: https://osf.io/5y8xz/

[11] We used the same thick terms as in Willemsen & Reuter (2021). The selection criteria were preregistered (https://osf.io/avbq3). The terms that were categorized as "descriptive" have sentiment values (from sentiWords dictionary) between –0.1 and 0.1, that is, very close to the neutral midpoint of 0.

[12] The power analysis (i.e., $1-\beta$) is a measure of how likely you are to correctly reject a false null hypothesis given a certain effect size, sample size, and alpha level. The p-value from the analysis of variance, on the other hand, indicates the likelihood of the observed data under the assumption that the null hypothesis is true. Throughout this paper, a significance level of 0.01 is employed for the power analyses.

[13] We also performed the same analyses for the subset of responses from participants who passed both training questions. The results were highly similar to the results of the full set of responses: (a) average defeasibility ratings for all thick terms were significantly above the neutral midpoint ($V = 17,892,\ p < .004$); (b) the ratings for negative thick terms only were significantly above the midpoint ($V = 5893.5,\ p < .001$); and (c) the ratings for positive thick terms only were not significantly above the midpoint ($V = 3223.5,\ p = .825$).

14 An analysis of responses from participants who passed both training questions confirmed the significant difference between negative $\mu = 6.82$ and positive terms $\mu = 5.12$: $t(386.04) = 6.244,\ p < .001$.

15 We also conducted the same test with only those data points from participants who passed both training questions, yielding a significant difference between thick ($\mu = 5.57$) and thin terms ($\mu = 6.73$): $t(339.16) = -4.14,\ p < .001$.

16 A reviewer for this journal raised a concern regarding our methodology. They pointed out that if positive terms, such as "friendly," are more frequently followed by the word "but" compared to negative terms like "rude," then our results might be attributed solely to the difference in occurrences of such phrases in everyday language. Participants might infer that because negative terms are followed less frequently by a positive, relativizing phrase, our stimulus sentences are more contradictory. To address this objection, we conducted an analysis of the relative frequencies with which our chosen terms are paired with "but" using the publicly accessible NOW corpus. For instance, the term "generous" appears a total of 280,489 times in the NOW corpus, and the phrase "generous but" occurs 701 times. This results in a percentage of 0.25%. We computed the average percentage for all positive terms, which is 0.32%, while the average percentage for negative terms is 0.38%. In essence, this suggests that the observed results are unlikely to be explained by difference in frequency of the phrase "[term] but."

17 The experimental design, predictions, and statistical models were preregistered with the Open Science Framework (https://osf.io/f8pxh).

18 In this experiment, we did not collect data for descriptive terms, which allowed us to reduce the number of participants compared to Experiment 1. A power analysis prior to conducting the experiment revealed that to test H5, we need at least 328 participants to detect a small effect of 0.1.

19 For the ART, we used the `ARTools`-package (Kay et al., 2021) in R (4.1.0).

20 The mixed effects model was computed using the `lmer4`-package (Bates et al., 2021) in R (4.1.0).

21 We performed an additional 2×2 ANOVA with Polarity and a factor for thick–thin concepts, to test whether the latter affects the results. We only find a very weakly significant effect for the thick–thin factor ($F(336) = 4.646,\ p = .032$).

22 The experimental design, predictions, vignettes, and statistical models were preregistered with the Open Science Framework (https://osf.io/zspnf).

23 We thank a reviewer for this journal for discussions on this point.

References

Abend, G. (2013). What the science of morality doesn't say about morality. *Philosophy of the Social Sciences*, *43*(2), 157–200.

Alicke, M. D. (2000). Culpable control and the psychology of blame. *Psychological Bulletin*, *126*(4), 556–574.

Amaya, S., & Doris, J. M. (2015). No excuses: Performance mistakes in morality. In J. Clausen, & N. Levy (Eds.), *Handbook of neuroethics* (pp. 253–272). Netherlands: Springer.

Anderson, R. A., Crockett, M. J., & Pizarro, D. A. (2020). A theory of moral praise. *Trends in Cognitive Sciences*, *24*(9), 694–703.

Archer, A. (2016). Are acts of supererogation always praiseworthy? *Theoria*, *82*, 238–255.

Baumgartner, L., Willemsen, P., & Reuter, K. (2022). The polarity effect of evaluative language. *Philosophical Psychology*. https://doi.org/10.1080/09515089.2022.2123311

Bartels, D. M., & Medin, D. L. (2007). Are morally motivated decision makers insensitive to the consequences of their choices? *Psychological Science*, *18*(1), 24–28.

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., Dai, B., Scheipl, F., Grothendieck, G., Green, P., Fox, J., Bauer, A., & Krivitsky, P. N. (2021). lme4: Linear mixed-effects models using ‘Eigen’ and S4. *R package version 1.1-27.1*. https://CRAN.R-project.org/package=lme4

Bear, A., Bensinger, S., Jara-Ettinger, J., & Knobe, J. (2018). What comes to mind? A mix of what’s likely and what’s good. In T. T. Rogers, M. Rau, X. Zhu, & C. W. Kalish (Eds.), *Proceedings of the Annual Meeting of the Cognitive Science Society* (pp. 130–135). Cognitive Science Society.

Bear, A., & Knobe, J. (2017). Normality: Part descriptive, part prescriptive. *Cognition, 167*, 25–37.

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

Bicchieri, C. (2014). Norms, conventions and the power of expectations. In N. Cartwright (Ed.), *Philosophy of social science* (pp. 208-231). Oxford: Oxford University Press.

Bicchieri, C. (2017). *Norms in the wild. How to diagnose, measure, and change social norms*. Oxford: Oxford University Press.

Calhoun, C. (2004). Common decency. In C. Calhoun (Ed.), *Setting the moral compass: Essays by women philosophers* (pp. 128–144). Oxford: Oxford University Press.

Chisholm, R. M. (1963). Supererogation and offence: A conceptual scheme for ethics. *Ration*, *5*, 1–14.

Cushman, F. (2013). Action, outcome, and value: A dual-system framework for morality. *Personality and Social Psychology Review*, *17*(3), 273–292.

Dahl, A., Gross, R. L., & Siefert, C. (2020). Young children’s judgments and reasoning about prosocial acts: Impermissible, suberogatory, obligatory, or supererogatory? *Cognitive Development*, 55, Article 100908.

Darley, J. M., & Shultz, T. R. (1990). Moral rules: Their content and acquisition. *Annual Review of Psychology*, *41*, 525–556.

Del Pinal, G., & Reuter, K. (2017). Dual character concepts in social cognition: Commitments and the normative dimension of conceptual representation. *Cognitive Science*, *41*, 477–501.

Ditto, P. H., Pizarro, D. A., & Tannenbaum, D. (2009). Motivated moral reasoning. *Psychology of Learning and Motivation*, *50*, 307–338.

Doris, J., Stich, S., Phillips, J., & Walmsley, L. (2020). Moral psychology: Empirical approaches. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*.

Driver, J. (1992). The suberogatory. *Australasian Journal of Philosophy*, *70*, 286–295.

Eklund, M. (2011). What are thick concepts? *Canadian Journal of Philosophy*, *41*(1), 25–49.

Elkin, L. A., Kay, M., Higgins, J. J., & Wobbrock, J. O. (2021). An aligned rank transform procedure for multifactor contrast tests. In J. Nichols, R. Kumar, & M. Nebeling (Eds.), *UIST ’21: The 34th Annual ACM Symposium on User Interface Software and Technology* (pp. 754–768). Association for Computing Machinery.

Eshleman, A. (2014). Worthy of praise. Responsibility and better-than-minimally decent agency. In *Oxford Studies in Agency and Responsibility*, Volume 2. Oxford University Press.

Feinberg, J. (1968). Supererogation and rules. *Ethics*. *71*(4), 276–288. https://doi.org/10.1086/291362

Fincham, F. D., & Shultz, T. R. (1981). Intervening causation and the mitigation of responsibility for harm. *British Journal of Social Psychology*, *20*(2), 113–120.

Gino, F., Shu, L. L., & Bazerman, M. H. (2010). Nameless+harmless=blameless: When seemingly irrelevant factors influence judgment of (un)ethical behavior. *Organizational Behavior and Human Decision Processes*, *111*(2), 93–101.

Gray, K., & Wegner, D. M. (2011). To escape blame, don’t be a hero–Be a victim. *Journal of Experimental Social Psychology*, *47*(2), 516–519.

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in coral judgment. *Neuron*, *44*(2), 389–400.

Grice, H. P. (1989). Logic and conversation. In H. P. Grice (Ed.), *Studies in the way of words* (pp. 22–40). Harvard University Press.

Guglielmo, S., & Malle, B. F. (2019). Asymmetric morality: Blame is more differentiated and more extreme than praise. *PLOS ONE*, *14*(3), e0213544.

Hauser, M., Cushman, F., Young, L., Kang-Xing Jin, R., & Mikhail, J. (2007). A dissociation between moral judgments and justifications. *Mind & Language*, *22*(1), 1–21.

Heyd, D. (1982). *Supererogation: Its status in ethical theory*. Cambridge: Cambridge University Press.

Heyd, D. (2019). Supererogation. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*.

Horvath, J., & Nado, J. (2021). Knowledge and normality. *Synthese*, *198*(12), 11673–11694.

Kahn, P. H., Jr. (1992). Children's obligatory and discretionary moral judgments. *Child Development*, *63*(2), 416–430.

Khan, U., Jaffer-Diaz, M., Najafizadeh, A., & Starmans, C. (2023). Going above and beyond? Early reasoning about which moral acts are best. *Cognition*, 105444. https://doi.org/10.1016/j.cognition.2023.105444

Kay, M., Elkin, L. A., Higgins, J. J., & Wobbrock, J. O. (2021). ARTool: Aligned Rank Transform. *R package version 1.1-27.1*. https://cran.r-project.org/web/packages/ARTool/index.html.

Kern, M. C., & Chugh, D. (2009). Bounded ethicality: The perils of loss framing. *Psychological Science*, *20*(3), 378–384.

Kneer, M., & Machery, E. (2019). No luck for moral luck. *Cognition, 182*, 331–348.

Knobe, J., Prasada, S., & Newman, G. E. (2013). Dual character concepts and the normative dimension of conceptual representation. *Cognition*, *127*(2), 242–257.

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A theory of blame. *Psychological Inquiry*, *25*(2), 147–186.

Mallon, R., & Nichols, S. (2011). Dual processes and moral rules. *Emotion Review*, *3*(3), 284–285.

Marshall, J., Gollwitzer, A., Mermin-Bunnell, K., Shinomiya, M., Retelsdorf, J., & Bloom, P. (2022). How development and culture shape intuitions about prosocial obligations. *Journal of Experimental Psychology. General*, *151*(8), 1866–1882.

Marshall, J., Wynn, K., & Bloom, P. (2020). Do children and adults take social relationship into account when evaluating people's actions? *Child Development*, *91*(5), e1082–e1100.

Miller, J. G., Bersoff, D. M., & Harwood, R. L. (1990). Perceptions of social responsibilities in India and in the United States: Moral imperatives or personal decisions? *Journal of Personality and Social Psychology*, *58*(1), 33–47.

Nichols, S. (2008). Sentimentalism naturalised. In W. Sinnott-Armstrong (Ed.), *Moral psychology. The cognitive science of morality: Intuition and diversity* (pp. 255–274). Cambridge, MA; London: MIT Press.

Petrinovich, L., & O'Neill, P. (1996). Influence of wording and framing effects on moral intuitions. *Ethology and Sociobiology*, *17*(3), 145–171.

Rai, T. S., & Holyoak, K. J. (2010). Moral principles or consumer preferences? Alternative framings of the trolley problem. *Cognitive Science*, *34*(2), 311–321.

Reuter, K. (2019). Dual character concepts. *Philosophy Compass*, *14*(1), e12557.

Reuter, K., Baumgartner, L., & Willemsen, P. (2023). Tracing thick and thin concepts through corpora. *Language & Cognition*, 1–20.

Ritov, I., & Baron, J. (1999). Protected values and omission bias. *Organizational Behavior and Human Decision Processes*, *79*(2), 79–94.

Roberts, D. (2013). Thick concepts. *Philosophy Compass*, *8*(8), 677–688.

Roberts, S. O., Guo. C., Ho, A. K., & Gelman, S. A. (2018). Children's descriptive-to-prescriptive tendency replicates (and varies) cross-culturally: Evidence from China. *Journal of Experimental Child Psychology*, *165*, 148–160.

Roberts, S. O., Ho, A. K., & Gelman, S. A. (2019). The role of group norms in evaluating uncommon and negative behaviors. *Journal of Experimental Psychology: General*, *148*(2), 374–387.

Shenhav, A., & Greene, J. D. (2010). Moral judgments recruit domain-general valuation mechanisms to integrate representations of probability and magnitude. *Neuron*, *67*(4), 667–677.

Shenhav, A., Rand, D. G., & Greene, J. D. (2016). The path of least resistance: Intertemporal choice and its relationship to choices, preferences, and beliefs. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2724547

Stout, N. (2020). On the significance of praise. *American Philosophical Quarterly*, *57*(3), 215–226.

Talbert, M. (2023). Moral responsibility. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy* (Fall 2023 Edition). Metaphysics Research Lab, Stanford University.

Tomasello, M. (2020). The moral psychology of obligation. *Behavioral and Brain Sciences, 43*, Article e56.

Turri, J., & Blouw, P. (2015). Excuse validation: A study in rule-breaking. *Philosophical Studies*, *172*(3), 615–634.

Väyrynen, P. (2021). Thick ethical concepts. In N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*.

Waldmann, M. R., Nagel, J., & Wiegmann, A. (2012). Moral judgment. In K. J. Holyoak & R. G. Morrison (Eds.), *The Oxford handbook of thinking and reasoning* (pp. 364–389). New York: Oxford University Press.

Wallace, J. (1994). *Responsibility and the moral sentiments*. Harvard University Press.

Wiegmann, A. & Sauer, H. (2021). The psychology and rationality of moral judgment. In G. Spohn & M. Knauf (Eds.), *The handbook of rationality* (pp. 962–974). MIT Press.

Willemsen, P., & Reuter, K. (2021). Separating the evaluative from the descriptive: An empirical study of thick concepts. *Thought: A Journal of Philosophy*, *10*(2), 135–146.

Wobbrock, J. O., Findlater, L., Gergle, D., & Higgins, J. J. (2011). The aligned rank transform for nonparametric factorial analyses using only ANOVA procedures. In D. Tan, G. Fitzpatrick, C. Gutwin, B. Begole, & W. A. Kellogg (Eds.), *CHI '11: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 143–146). Association for Computing Machinery.

Woolfolk, R. L., Doris, J. M., & Darley, J. M. (2006). Identification, situationalconstraint, and social cognition: Studies in the attribution of moral responsibility. *Cognition*, *100*(2), 283–301.

Wysocki, T. (2020). Normality: A two-faced concept. *Review of Philosophy and Psychology*, *11*(4), 689–716.

Young, L., & Phillips, J. (2011). The paradox of moral focus. *Cognition*, *119*(2), 166–178.

Young, L., & Tsoi, L. (2013). When mental states matter, when they don't, and what that means for morality. *Social and Personality Psychology Compass*, *7*(8), 585–604.

Zakkou, J. (2018). The cancellability test for conversational implicatures. *Philosophy Compass*, *13*(12), e12552.