

## The polarity effect of evaluative language

Lucien Baumgartner, Pascale Willemsen & Kevin Reuter

To cite this article: Lucien Baumgartner, Pascale Willemsen & Kevin Reuter (2022): The polarity effect of evaluative language, *Philosophical Psychology*, DOI: [10.1080/09515089.2022.2123311](https://doi.org/10.1080/09515089.2022.2123311)

To link to this article: <https://doi.org/10.1080/09515089.2022.2123311>



© 2022 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.



Published online: 27 Sep 2022.



Submit your article to this journal [↗](#)



Article views: 519




View related articles [↗](#)



View Crossmark data [↗](#)

# The polarity effect of evaluative language

Lucien Baumgartner , Pascale Willemsen  and Kevin Reuter 

Department of Philosophy, University of Zurich, Zurich, Switzerland

## ABSTRACT

Recent research on thick terms like “rude” and “friendly” has revealed a polarity effect, according to which the evaluative content of positive thick terms like “friendly” and “courageous” can be more easily canceled than the evaluative content of negative terms like “rude” and “selfish”. In this paper, we study the polarity effect in greater detail. We first demonstrate that the polarity effect is insensitive to manipulations of embeddings (Study 1). Second, we show that the effect occurs not only for thick terms but also for thin terms such as “good” or “bad” (Study 2). We conclude that the polarity effect indicates a pervasive asymmetry between positive and negative evaluative terms.

## ARTICLE HISTORY

Received 12 May 2022  
Accepted 3 September 2022

## KEYWORDS

Polarity effect; thick terms; thin terms; evaluative language; moral judgments; praise & blame

## 1. Introduction

The terms we use to make evaluative judgments fall into at least two main classes (e.g., Eklund, 2011; Väyrynen, 2013). First, thin terms like “great” and “awful” evaluate, i.e., praise or blame, a person or state of affairs without providing any, or at least very little, descriptive information as to what it is that the person or state of affairs is considered praise- or blameworthy for. Second, thick terms like “generous” and “honest” also evaluate but additionally communicate the descriptive features in virtue of which someone or something is evaluated. For instance, by saying that Sally is generous and by calling her honest, we evaluate her behavior positively. However, being generous is clearly different from being honest – generosity is concerned with sharing things with others, honesty is about telling the truth. While “generous” and “honest” share the same evaluative component, they differ in the descriptive features that are the basis for the positive evaluation.<sup>1</sup>

The two different kinds of content of thick concepts have been the focus of philosophical theorizing with potentially wide-ranging consequences. Philosophers regularly assume a clear, categorical distinction between facts and values (Hume, 1739; Moore, 1903; Putnam, 2002). While descriptive statements (e.g., “Nadal won 22 Grand Slams.”) are usually intended to state facts and are, thus, either true or false, evaluative claims (e.g., “Nadal is

the best tennis player of his generation.”) are often thought not to be truth-apt. The existence of thick concepts has been argued to challenge this fact-value dichotomy. However, this challenge only gets off the ground if the content of thick terms cannot be reduced to descriptive content on the one hand, and evaluative content on the other. Thus, if the evaluative and descriptive content of a statement like “Nadal made a compassionate speech after the match” cannot be disentangled, then it seems at least possible for evaluative features to be genuine features of the world that are truth-evaluable. Various *non-reductionist positions* along these lines have been advocated during the last few decades, see, e.g., Williams (1985), Dancy (1996), Putnam (2002), and Roberts (2013).

Among those scholars who argue for a *reductionist position*, a further question arises: how do thick concepts communicate their evaluative content. While some philosophers, e.g., Elstein and Hurka (2009), Hare (1952), or Kyle (2020), believe thick concepts to *semantically* encode their evaluative content, others believe that evaluative content is merely *pragmatically conveyed* by the use of thick concepts (Blackburn, 1992; Cepollaro & Stojanovic, 2016; Cepollaro, 2020; Hare, 1963; Väyrynen, 2021, Willemsen et al., 2023). Arguments for either position usually rely on or are supported by linguistic intuitions. For example, if conversational implicatures only communicated evaluative contents, then statements like “What Tom did was cruel, but by that I am not saying something negative about him” should be felicitous. Semanticists and some pragmatists<sup>2</sup> hold different intuitions, claiming that such sentences are in fact contradictory.

Willemsen and Reuter (2020, 2021) tested these two opposing intuitions by using the cancellability test for conversational implicatures (see Grice, 1989; Sullivan, 2017; Zakkou, 2018). Here are some examples of the experimental stimuli that were used, distinguishing between attributions of thick terms to people (Character) and attributions of thick terms to behavior (Behavior):

- (1) **Negative Character:** Amy is rude, but by that I am not saying something negative about Amy.
- (2) **Negative Behavior:** Amy’s behavior last week was rude, but by that I am not saying something negative about Amy’s behavior that day.
- (3) **Positive Character:** Tom is friendly, but by that I am not saying something positive about Tom.
- (4) **Positive Behavior:** Tom’s behavior last week was friendly, but by that I am not saying something positive about Tom’s behavior that day.

Participants were then asked whether the speaker, Sally, contradicts herself. The most crucial finding goes beyond the initial research question and reveals a systematic difference between positive and negative terms.

Negative evaluations were significantly harder to cancel compared to positive ones ( $\Delta \approx 1.0$  on a 9-point Likert scale), irrespective of whether the thick terms were assigned to the character or the behavior. More specifically, statements like (1) and (2) were judged to be significantly more contradictory than statements like (3) and (4). This asymmetry, called **Polarity Effect**, was previously unknown and provides a challenge to the idea that positive and negative thick terms can be treated alike (see also Väyrynen (2021) and Zakkou (2021)).

So far, the polarity effect has only been recorded for thick terms. One might wonder, though, whether the effect is in fact a more global effect that also holds for other evaluative terms, specifically thin terms like “good” and “bad”. It seems plausible to assume that the effect only occurs for thick concepts but disappears for thin ones. Thin concepts are said to be merely evaluative, with their main function being to express approval or disapproval. What does remain if we cancel this sole content of a thin concept? The term should be empty and no longer express anything. While thin terms are often characterized as lacking descriptive content, not all philosophers agree that thin concepts are merely evaluative. In fact, many scholars believe that thin and thick concepts are not categorically but gradually different (e.g., Chappell, 2013; Scheffler, 1987; Smith, 2013; Väyrynen, 2013). For instance, it has been argued that a concept as thin as “ought” communicates descriptive content, as “ought” is widely agreed to imply “can” (Väyrynen, 2021). Following this line of reasoning, Chappell (2013, p. 182) argues that “there are no thin concepts. Or almost none” (see also Smith, 2013). We wish not to take a stance in this debate. However, we believe that even if we grant that thin concepts express some non-evaluative content, we still consider it plausible that the evaluative content plays a much more significant role for the thin concepts’ semantic meaning. Hare expressed a similar idea by suggesting that the difference between thin and thick concepts was that the evaluation is “more firmly attached” to thin concepts than to thick concepts (Hare, 1963, p. 24–25).

If our reasoning above is on the right track, the polarity effect should not pertain to thin concepts as well. Another reason to think that the polarity effect occurs for thick concepts only, is that thick concepts are often descriptively very rich and contain disjunctive features (Wiggins, 1993), which may lead to unexpected effects in experimental settings like the cancellability task. An example: One person can be called courageous for trying a dangerous trick on a snowboard, while another demonstrates courage by standing up to the class bully, or simply by being themselves and not caring about other people’s opinion. Courage comes in many forms that often cannot be properly reduced to one shared core feature. If this picture is correct, then the evaluation of a thick concept is less central to the

semantic content – it is simply one of many things that make up the concept. For thin concepts, however, the evaluation is highly central.

This line of reasoning can still not explain why positive and negative terms behave differently when the evaluation is canceled, but it provides a suggestion of where to search for the root of the effect. If the polarity effect were a phenomenon restricted to thick terms only, then a promising explanation of the effect, let's call it **thick concept explanation**, would dig into the intricacies of thick terms. Here are three ways one could cash out the thick concept explanation. First, how the evaluative content combines with the descriptive content might be different for positive and negative terms. Whereas such an explanation would be very surprising, it is at least theoretically possible that the content of negative thick concepts cannot be disentangled into evaluative and non-evaluative parts (see non-reductionist accounts like Dancy (1996) and Putnam (2002)), whereas, for positive thick concepts, such a reductionist account is possible (for reductionist accounts see e.g., Gibbard (1992), Elstein and Hurka (2009)). Alternatively, one might follow Kyle's recent suggestion (Kyle, 2020) that thick terms are expanded contents of thin terms. Based on this, it could be suggested that the expansion works differently for negative and positive terms. Second, one might suspect the existence of systematic differences in descriptive richness between positive and negative terms. More specifically, positive thick concepts might be argued to be descriptively richer, see e.g., the courage example above, making the evaluative content less central. For negative terms, the evaluation would be more central and consequently harder to cancel. For the time being, this hypothesis cannot be ruled out, although we have little reason to believe in such systematic differences. Third, one might hold that there is a difference in the way positive and negative terms semantically or pragmatically entail evaluative content. Thus, whereas evaluative content is semantically entailed in the case of negative terms, evaluative content is pragmatically implicated for some positive terms at least.

If the polarity effect were to also hold for thin terms, then an explanation that focuses on the descriptive aspects of thick concepts would not take us very far. Thus, in case the polarity effect is a more pervasive evaluative language effect, then the following claim should hold:

**Pervasive Linguistic Asymmetry:** A negative evaluation is, *ceteris paribus*, harder to explicitly cancel compared to a positive evaluation.

Consequently, a more encompassing explanation would be required. Willemsen and Reuter (2021) suggest an explanation of the polarity effect

that is grounded in different social norms, let's call it **social norms explanation**, that may guide our behavior. They state:

Uttering a positive thick term without the intention to commit to a positive evaluation seems relatively harmless. Being misunderstood in cases of negative thick terms has a potentially greater impact. If mistaken, a speaker communicates a negative evaluation they initially did not want to commit to. Since negative evaluations harm others by diminishing their social status and reputation, people are less willing to accept the cancellation of a negative evaluation. (p. 8)

Such an explanation would be consistent with a growing body of empirical evidence that has shown moral valence to affect non-moral judgments, e.g., of knowledge (Beebe & Buckwalter, 2010) and causation (Sytsma, Bluhm, Willemsen, & Reuter, 2019; for an overview see Willemsen & Kirfel, 2019). Additionally, the philosophical and linguistic literature is rife with results in which social norms seem to have an asymmetrical influence on praise and blame (Guglielmo & Malle, 2019). Recently, Anderson et al. (2020) argued that while both praise and blame are essential to sustaining social relationships and facilitating social regulation, blaming one another comes with significant social costs, both on the part of the blaming and the blamed party. Being blamed can have serious consequences, such as loss of reputation and social alliances, social exclusion, or punishment. Consequently, the *wrongful* attribution of blame that is unjustifiably causing a person to suffer these negative consequences is itself an act of severe social impact.<sup>3</sup>

So far we lack evidence of the effect's robustness across embeddings and whether or not it is a thick concept or an evaluative language effect. In this paper, we demonstrate that the polarity effect is not only robust but extends to thin ethical concepts as well, allowing for the conclusion that the polarity effect is indicative of a pervasive linguistic asymmetry. In the empirical part of the paper, we do two things: First, in Study 1, we provide a clearer understanding of the polarity effect by investigating how far-reaching it is, viz. in what embeddings it occurs (Section 2.1). In Section 2.2, we provide empirical evidence (Study 2) that the polarity effect holds more globally for both thick as well as thin terms.

## 2. The extent and character of the polarity effect

### 2.1. Study 1: Investigating the polarity effect in different embeddings

In this study, we investigate the scope of the polarity effect. It might be argued that the previously recorded effect only holds when a thick term is attributed to an individual person ("Amy is rude.")— hereafter, Individual Statement condition — but not in other embeddings, e.g., generic generalizations ("People are rude."). If that were the case, then the polarity effect

would have a more narrow application and would be moderated by the subject term.

Two main hypotheses guided the design of our study.<sup>4</sup> First, we predicted to replicate the polarity effect recorded in previous studies:

**Polarity Hypothesis (H1):** Contradiction ratings in the Individual Statement condition are significantly higher for negative thick terms compared to positive thick terms.

Second, we expected an inverse relationship between the scope of predication and the assertive commitment: the more general an evaluative statement, the smaller the commitment to the evaluation. Generic statements (“people are rude”) are notoriously easy to take back, due to their inherent scope ambiguity (e.g., Sterken, 2017; Thakral, 2018). Similarly, limited scope statements (e.g., “some people are rude”) do not commit the speaker to the evaluation on a personal level. Individual statements (e.g., “Amy is rude”), in contrast, have higher immediate social costs and thus are most likely to follow social norms. Hence, we hypothesized an embedding effect:

**Embedding Hypothesis (H2):** The polarity effect is significantly reduced in limited scope statements and for generic generalizations.

### 2.1.1. Methods

932 participants were recruited via Prolific and completed an online survey implemented in Qualtrics. All participants were required to be at least 18 years old, English native speakers (or bilingual), and to have an approval rate of at least 95%.

The remaining 872 participants had an average age of 38.47 years, and the gender distribution in the sample was 55.96% male, 43.81% female, and 0.23% non-binary. The 6 positive and 6 negative thick terms we tested were:<sup>5</sup>

- **Positive:** compassionate, courageous, friendly, generous, honest, virtuous
- **Negative:** cowardly, cruel, manipulative, rude, selfish, vicious

Here are three exemplary statements we used (including the question that was asked subsequently), illustrating each variant with a different thick term:

Please imagine that [Sally/Tom] said the following sentence:

*Individual statement.* “[Amy/Steve] is rude, but by that I am not saying something negative about [her/him].”

*Limited scope statement.* Some people are friendly, but by that I am not saying something positive about them.

*Generic statement.* “People are selfish, but by that I am not saying something negative about them.”

Does [speaker] contradict [herself/himself]?

Contradiction ratings were recorded on a 9-point Likert scale ranging from 1 = “definitely not” to 9 = “definitely yes”. Before participants gave their responses to the test sentences, they were given instructions on how to understand what a contradiction is (see preregistration material). The stimuli included proper names, both for the speaker (Sally/Tom) and the target of the predication in the individual person statement (Amy/Steve), which is a possible source of unexpected gender effects. Hence, the gender of the speaker was randomized evenly in order to control for possible gender effects.<sup>6</sup> Each participant was randomly assigned to one of the 72 stimuli (3 (embeddings)  $\times$  6 (concepts)  $\times$  2 (polarity)  $\times$  2 (gender of the speaker)).

### 2.1.2. Results

For the individual statements, the observed mean of positive thick concepts (6.39) was indeed lower compared to negative thick concepts (6.97). As the contradiction ratings significantly deviate from a normal distribution, we used non-parametric alternatives to test our hypotheses. According to a one-sided unpaired two-samples Wilcoxon test ( $W = 9348.5$ ,  $p = 0.013$ ), positive thick concepts have significantly lower average contradiction ratings than negative thick concepts (on 0.05-alpha level). Thus, canceling negative thick concepts was assessed to be more contradictory than canceling positive thick concepts. Hence, we cannot reject H1.

Our second hypothesis was that the difference between negative thick terms and positive thick terms will be largest for individual statements. However, the differences in the estimated marginal means do not support this hypothesis, as shown in Table 1.<sup>7</sup> In fact, the difference for individual statements is the smallest (−0.60). All differences are significant on 0.05-alpha level. Hence, our hypothesis has to be rejected. Lastly, none of the

**Table 1.** Pairwise contrasts (positive – negative) of estimated marginal means by embedding. For individual statements, the difference in average contradiction ratings was 0.60.

Embedding	$\Delta$ Estimate	SE	t-ratio	p-value
Individual	−0.60	0.30	−2.00	0.047
Limited	−0.65	0.30	−2.14	0.033
Generic	−0.78	0.31	−2.56	0.011



control variables (gender of the speaker, age, and gender of the participant) had any significant effect.

### 2.1.3. Discussion

In Study 1, we replicated the polarity effect for statements in which a thick term is attributed to an individual. Furthermore, the scope of this effect is not limited to statements of the form “[Subject] is [thick term]”. Significant differences were found across all three embeddings, providing support for the claim that the polarity effect is rather pervasive. This suggests that the effect does not depend on the linguistic construction used.

## 2.2. Study 2: Extending the polarity effect to thin concepts

In previous studies as well as in Study 1 above, it was found that the polarity of a thick term has an effect on contradiction ratings using the cancellability paradigm. In this experiment, we investigated whether the polarity effect shows up for both thin and thick concepts, which would indicate that the effect is more widespread and holds for evaluative concepts more generally rather than for thick concepts only. We therefore examined whether negative and positive thin terms behave differently from thick terms with respect to canceling their evaluative content. We thus formulated the following hypotheses:<sup>8</sup>

**Main Effect Hypothesis (H3):** There is a significant effect of Polarity (Positive vs. Negative) on contradiction ratings, such that the ratings are higher for negative terms compared to positive terms.

**Interaction Hypothesis (H4):** There is no significant two-way interaction of Concept class (Thin vs. Thick) and Polarity (Positive vs. Negative).

**Thin Concept Hypothesis (H5):** Contradiction ratings are significantly higher for negative *thin* terms compared to positive *thin* terms.

**Thick Concept Hypothesis (H6):** Contradiction ratings are significantly higher for negative *thick* terms compared to positive *thick* terms.

### 2.2.1. Methods

325 participants were recruited via Prolific and completed our online survey implemented in Qualtrics. The same inclusion criteria and instructions were used as in Study 1. The final sample included 303 participants (34.65% male, 63.37% female, 1.98% non-binary) with an average age of 36.69 years.

As stimuli, we used three positive and three negative thick concepts as well as three positive and three negative thin concepts:<sup>9</sup>

- Thin concepts:
  - Positive: good, great, ideal
  - Negative: bad, awful, terrible<sup>10</sup>
- Thick concepts:
  - Positive: friendly, honest, compassionate
  - Negative: rude, manipulative, cruel

After two test questions, participants were presented with the following prompt:<sup>11</sup>

Please imagine that Sally said the following sentence:

“What [person] did last week was [thin/thick term], but by that I am not saying something [positive/negative] about [her/his] behavior that day.”

Does Sally contradict herself?

The participants answered on a 9-point Likert scale anchored at 1 = “definitely not” and 9 = “definitely yes”. Since the gender of the speaker did not have any significant effect in Study 1, we did not add it as a control variable in Study 2. Instead, we varied the gender of the person Sally is speaking about, but without duplicating the number of vignettes. Accordingly, participants were randomly assigned to one of the 12 vignettes (3 (terms) × 2 (concept classes) × 2 (polarity)).

### 2.2.2. Results

In Study 2, we found the main Polarity Effect again: according to a one-sided unpaired two-samples Wilcoxon test ( $W = 15,712$ ,  $p < 0.001$ ), negative terms have significantly higher contradiction ratings than positive terms (across concept classes), thus supporting H3. Furthermore, the differences of differences based on Aligned Rank Transform (ART) non-parametric ANOVA (t-ratio (299) = 1.284,  $p = 0.2002$ ) showed that there is no significant two-way interaction of concept class and polarity, which is in line with our predictions in H4. The Polarity Effect was also found for thin concepts (H5) and thick concepts (H6) respectively: a one-sided unpaired two-samples Wilcoxon test ( $W = 4021.5$ ,  $p < 0.001$ ) showed that negative thin concepts have significantly higher contradiction ratings than positive thin concepts; the same was found for thick concepts ( $W = 3918.5$ ,  $p < 0.001$ ). In summary, none of our hypotheses can be rejected.

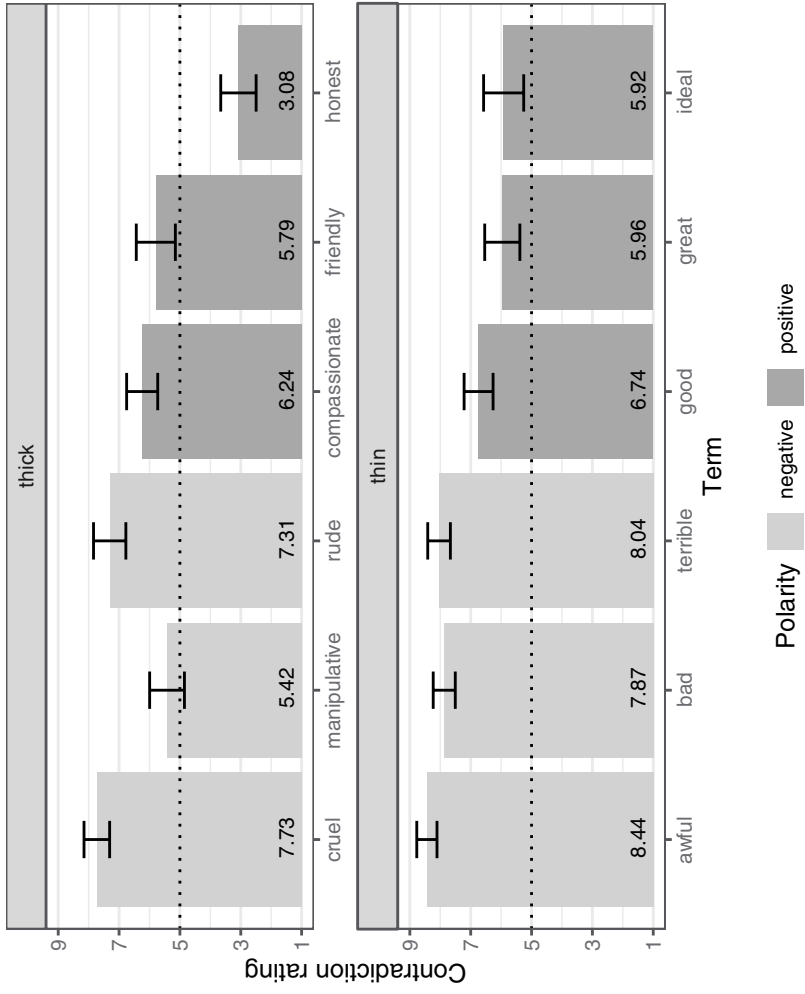


Figure 1. Average contradiction ratings per concept. The error bars display the standard error around the means.

In general, thick terms (5.93) have lower average contradiction ratings than thin concepts (7.15). [Figure 1](#) depicts the means and standard error per term, which reveals two outliers, namely the thick terms “manipulative” (5.42) and “honest” (3.08). We thus ran additional tests to check for concept class differences for positive and negative terms respectively, with and without outliers. A two-sided unpaired two-samples Wilcoxon test ( $W = 2243.5$ ,  $p = 0.016$ ) showed that there are significant differences between positive thin and positive thick concepts; the same is true for negative thin and thick concepts ( $W = 2032.5$ ,  $p < 0.001$ ). These differences are no longer significant for positive thin and thick, if we drop the the outlier “honest“ ( $W = 1770$ ,  $p = 0.555$ ), nor for negative thin and thick concepts after dropping “manipulative“ ( $W = 1667.5$ ,  $p = 0.187$ ).

### 2.2.3. Discussion

The results of Study 2 paint a clear picture, according to which the polarity effect does not hold for thick terms only, but is a more widespread effect that applies to evaluative concepts more generally. Our results suggest that the Polarity Effect between positive and negative terms is a unified phenomenon for thin and thick concepts.

## 3. General discussion

### 3.1. Summary of the results

The purpose of the empirical part of the paper was twofold. First, we aimed to replicate the polarity effect, thereby testing the extent to which the effect holds in different embeddings. Second, we aimed to investigate whether the polarity effect is a narrow *thick concept* effect or holds more widely for a larger set of evaluative terms, including thin terms. Regarding the first aim, we successfully replicated the polarity effect for individual subjects. Furthermore, and against our predictions, the effect popped up in all three embeddings we tested, i.e., not only when thick terms are ascribed to persons, but also when being attributed to a group of people, as well as in generic statements. From this, we can conclude that the polarity effect is not (at least not strongly) dependent on the scope of predication in which the thick term appears. Instead, the polarity effect indicates a pervasive linguistic asymmetry between positive and negative evaluative terms.

In order to pursue our second aim, we tested not only a batch of thick terms but also six thin terms. The results of Study 2 reveal that statements including positive thin terms are also less contradictory than negative thin terms, mirroring the effect for thick terms. While we cannot rule out that the outcome of Study 2 is the result of two independent effects, the similar

results for thick and thin terms in Study 2 do indicate that the same cause is driving the effect in both cases.

### 3.2. Interpretation and discussion of the results

Two accounts were stated in the introduction that may account for the polarity effect of thick terms. First, given that thick terms have both evaluative and descriptive content, we hypothesized that the connection between descriptive and evaluative content might be stronger for negative thick terms than for positive thick terms. The greater entanglement for negative thick terms might be accounted for by the differences in the descriptive content between negative and positive thick terms. Alternatively, one might explain the polarity effect by hypothesizing that negative thick terms semantically entail their evaluative content, whereas positive thick terms pragmatically convey their evaluation.

Second, as suggested by Willemsen and Reuter (2021), certain social roles might be in place that govern the use of positive and negative terms. If a person publicly attributes a negative aspect to another person, she needs to be able to justify the blameworthy aspect more strongly than when attributing a positive aspect. Consequently, the use of negative terms comes with greater social costs, because they can do serious harm and need to have a more solid grounding. If this social norm hypothesis were true, then the polarity effect might as well show up for thin terms. Thus, a positive result would provide some evidence in favor of the social norm explanation.

The results of Study 2 suggest the **thick concept explanation** to be false. If such an explanation were to hold, we would not expect the polarity effect to show up for thin terms. In other words, a positive result for thin terms strongly indicates the falsity of the thick concept explanation. Instead, the data provide some evidence that social norms might be key to understanding the polarity effect. The **social norm explanation** is also in line with recent results that show that people are less inclined to permit the use of negative thick terms when these are not intended to be used to blame a person (Willemsen & Reuter, 2020).

Against our interpretation, one might object that statements of the form “What Amy did last week was rude, but by that I am saying anything negative about her behavior that day.” are not apt to test the social norm hypothesis. Why is that? Well, if the speaker felicitously cancels a negative predication, the resulting expression itself is not negative, even though it contains a negative term. Consequently, the negative thick or thin term would be rather harmless, contrary to the social norms hypothesis. However, it is important to note that our stimuli consisted of a predicative main clause, e.g., “What Amy did last week was rude,” and an anaphoric relative clause, e.g., “but by that that I am not saying anything negative about

her behavior that day.” In fact, the relative clause is a comment on the main clause. The question of whether the speaker contradicts him- or herself thus is about the validity of the comment in relation to the main clause. Even if the expression itself might not be negative, that is not what participants were asked to consider. Rather, participants were asked to assess the use of a thick or thin *term* by virtue of a relative clause.

Looking more closely at our data, one might wonder why the mean value for “honest” was significantly lower than for all other positive items. Interestingly, this experiment is not the first in which “honest” was an outlier (see Willemsen & Reuter, 2020, 2021; and Willemsen, Baumgartner, Cepollaro, & Reuter, ms). We believe that two possible factors drive this effect. First, honesty is one of the virtues that can easily become a vice. Some truths are just tough to bear and often conflict with other norms of politeness, respect, and so on. Thus, calling someone honest does not necessarily involve a positive evaluation. These considerations might have affected participants’ interpretations of the stimulus, making the positive evaluation particularly easy to cancel.<sup>12</sup> Second, many uses of “honest” do not seem to be communicating high praise for an agent, but rather that the agent has merely met a certain minimum standard. We can and should expect others to be honest.

### 3.3. *Alternative accounts*

In a recent paper, Willemsen et al. (ms) provide an alternative explanation for the polarity effect that considers the relevance of social expectations for the interpretation of evaluative language. Let’s call this explanation the **evaluative deflation explanation**. They argue that acts that count as morally desirable and are referred to by the use of positive terms, such as being compassionate, can either meet our expectations or they can exceed our expectations. The results of a series of studies indicate that people can use positive terms in two ways: first, a proper evaluative way in which speakers intend to praise the agent and, second, in an evaluatively deflated manner to refer to actions that only meet our expectations.

Applying this account to the example of “honest” above, we can easily see why people might interpret “honest” in an evaluatively deflated way. For communication, in particular, and cooperation, more generally, to work, people need to be honest.<sup>13</sup> Thus, following Willemsen et al.’s (2022) suggestion, we should expect that when people call a person’s behavior honest, they often do not want to praise the agent for having exceeded our moral standards. Rather, all they wish to communicate is that the agent meets a certain standard necessary for people to cooperate.

Before we conclude, we would like to discuss a further alternative explanation, call this one **politeness explanation**.<sup>14</sup> The polarity effect we

empirically investigated and discussed in this paper is not the only asymmetry found for positive and negative terms. It has been observed that people often prefer to use negated positives like “not tall” or “not friendly” instead of bare negatives like “short” and “rude”. A classic interpretation of this effect highlights the role of the importance of being polite. More specifically, uttering negated positive terms is part of a politeness strategy used to avoid straight-out, face-threatening negative terms (Brown & Levinson, 1987, 1987; Gotzner & Mazzarella, 2021; Horn, 1989; Mazzarella & Gotzner, 2021). Although saying that a person is rude is semantically interpreted to be equivalent to saying that this person is not friendly, the pragmatic effects are quite different: the use of “rude” is threatening the reputation of a person more severely than stating that the person is “not friendly”.

People’s inclination to use positive terms and avoidance of negative terms might also provide an explanation for the polarity effect we recorded. When a person states that someone is friendly or courageous, other people can easily interpret the use of “friendly” and “courageous” as *politeness* talk, especially if that person continues by claiming “but by that I am not saying anything positive about this person.” In other words, our participants might have considered the cancellability statements for positive terms to be less contradictory because they believed the speaker only used the positive term for social reasons. Note that such an interpretation is not possible for negative terms: We usually do not use negative terms unless we really like to communicate something negative about a person or their behavior.

The politeness explanation is related to the social norm explanation presented in this paper in that both accounts consider social norms to be crucial for a comprehensive explanation of the polarity effect. However, while the social norm explanation focuses more strongly on why cancellability statements featuring negative terms are considered to be contradictory, the politeness account provides a more direct explanation for why cancellability statements featuring positive terms are considered less contradictory. Of course, it might well be that it is not a single factor that drives the polarity effect. Future research will hopefully provide more evidence for or against any of the discussed accounts.

## Notes

1. More recently, researchers have identified another class of evaluative concepts, the so-called dual character concepts (Del Pinal & Reuter, 2017; Knobe et al., 2013; Reuter et al., 2020; Reuter, 2019). Given that dual character concepts have two independent dimensions for categorization, we will not empirically investigate this class of concepts in this paper. Also, some philosophers suggest that pejoratives and slurs constitute independent classes of evaluative concepts (for a discussion, see Cepollaro,

- 2020). However, both pejoratives and slurs only communicate negative evaluations and do not have positive counterparts. This paper aims to investigate whether positive and negative evaluations of terms within the same class behave differently. Therefore, we omit pejoratives and slurs.
2. Pragmatists need not advocate that the evaluative content is communicated via conversational implicatures. Instead, they may hold that the evaluative content is presupposed or conventionally implicated.
  3. See also Willemsen, Baumgartner, Cepollaro, & Reuter, ms, for discussion.
  4. The experimental design, predictions, and statistical models were pre-registered with the Open Science Framework. The data file with all the responses can be downloaded here: <https://osf.io/fn84r><https://osf.io/fn84r>.
  5. We selected the same 12 thick terms that were used in Willemsen and Reuter (2021). Among other reasons for their selection (see <https://osf.io/xew6d>), these adjectives are frequently used in ordinary language.
  6. In the individual statements, Sally only talks about Amy and Tom only about Steve (i.e., gender is held constant across speaker and subject term).
  7. The estimation is based on a two-way ANOVA of the interaction of polarity and embedding, with the gender of the speaker (male/female), as well as age (continuous) and gender of the respondent (male/female/non-binary) as controls.
  8. The experimental design, predictions, and statistical models were pre-registered with the Open Science Framework. The data file with all the responses can be downloaded here: <https://osf.io/fn84r><https://osf.io/fn84r>.
  9. Whether or not the six thin concepts indeed encode no descriptive content at all, is a matter of debate. For instance, the term 'ideal' is plausible thought to encode some highly general descriptive content along the lines of "matching some pattern that fits certain purposes". Importantly, we believe there is still a crucial difference between the rather specific descriptive content of thick terms like 'courageous' and 'manipulative' and the highly general descriptive content of terms like 'ideal' and 'terrible'.
  10. We selected highly frequent thin terms, including 'good', 'great', and 'bad' (2nd, 4th, and 22nd most frequently used adjectives in American English in the Corpus of Contemporary American English).
  11. Whereas in Study 1 we used thick term attributions to persons, in Study 2 thin and thick terms were attributed to behavior. Previous studies have revealed no differences between both conditions.
  12. A reviewer for this journal has pointed out that other terms like 'courageous' can similarly switch from being a virtue to being a vice. For instance, we are likely not to consider a Nazi soldier to be courageous. However, it seems to us that the term 'courageous' requires far more stage-setting and a particular context for it to become a vice. In contrast, people regularly deliberate about how honest they can or should be. This asymmetry is also reflected in the frequency with which we say that someone is 'too honest' compared to 'too courageous'. Whereas we find 932 uses of 'too honest' (0.18% of all uses of 'honest') on the Corpus of Contemporary America English, 'too courageous' was only listed 26 times (0.04%).
  13. Truthfulness is one of the central maxims in Gricean and neo-Gricean frameworks (Carston, 2004; Horn, 2004). Also, truthfulness is a key element in many discussions on the norm of assertion (Kneer, 2018; Marsili & Wiegmann, 2021; Reuter & Brössel, 2019).
  14. We would like to thank a reviewer for this journal for suggesting this alternative explanation of the polarity effect.



## Acknowledgements

We would like to thank Bianca Cepollaro and Ethan Landes, as well as two anonymous reviewers for this journal for their helpful comments. We are grateful for the feedback we have received at several conferences and workshops, e.g., the Annual Conference of the Cognitive Science Society 2022, First European XPhi Conference, the 10th Annual Conference of the Sociedad de Lógica, Metodología y Filosofía de la Ciencia en España, and the XPhi Lab at the University of Zurich.

## Disclosure statement

The authors declare not to have any conflicting interests.

## Funding

The research of Lucien Baumgartner, Pascale Willemsen, and Kevin Reuter was funded by the Swiss National Science Foundation (SNSF), grant number PCEFP1 181082. Pascale Willemsen also received generous support from the SNSF, grant number PZ00P1 201737.

## ORCID

Lucien Baumgartner  <http://orcid.org/0000-0003-1698-0114>

Pascale Willemsen  <http://orcid.org/0000-0002-4563-1397>

Kevin Reuter  <http://orcid.org/0000-0003-2404-1619>

## References

- Anderson, R., Crockett, M., & Pizarro, D. (2020). A theory of moral praise. *Trends in Cognitive Sciences*, 24(9), 694–703. <https://doi.org/10.1016/j.tics.2020.06.008>
- Beebe, J., & Buckwalter, W. (2010). The epistemic side effect-effect. *Mind & Language*, 25(4), 474–498. <https://doi.org/10.1111/j.1468-0017.2010.01398.x>
- Blackburn, S. (1992). Through thick and thin. *Proceedings of the Aristotelian Society*, 66, 284–299.
- Brown, P., & Levinson, S. (1987). *Politeness: Some universals in language usage*. Cambridge University Press.
- Carston, R. (2004). Truth-Conditional content and conversational implicature. In C. Bianchi (Ed.), *The Semantics/Pragmatics Distinction* (pp. 65–81). CSLI Publications.
- Cepollaro, B. (2020). *Slurs and thick terms: When language encodes values*. Lexington Books.
- Cepollaro, B., & Stojanovic, I. (2016). Hybrid evaluatives: In defense of a presuppositional account. *Grazer Philosophische Studien*, 93(3), 458–488. <https://doi.org/10.1163/18756735-09303007>
- Chappell, S.-G. (2013). There are no thick concepts. In S. Kirchin (Ed.), *Thick Concepts* (pp. 182–196). Oxford University Press.
- Dancy, J. (1996). In defense of thick concepts. *Midwest Studies in Philosophy*, 20, 263–279. <https://doi.org/10.5840/msp19952016>

- Del Pinal, G., & Reuter, K. (2017). Dual character concepts in social cognition: Commitments and the normative dimension of conceptual representation. *Cognitive Science*, 41(3), 477–501. <https://doi.org/10.1111/cogs.12456>
- Eklund, M. (2011). What are thick concepts? *Canadian Journal of Philosophy*, 41(1), 25–49. <https://doi.org/10.1353/cjp.2011.0007>
- Elstein, D., & Hurka, T. (2009). From thick to thin: Two moral reduction plans. *Canadian Journal of Philosophy*, 39(4), 515–535. <https://doi.org/10.1353/cjp.0.0063>
- Gibbard, A. (1992). Thick Concepts and Warrant for Feelings, in Proceedings of the Aristotelian Society, supplementary volume 66: 267–283.
- Gotzner, N., & Mazzarella, D. (2021). Face management and negative strengthening: The role of power relations, social distance and gender. *Frontiers in Psychology: Experimental Approaches to Pragmatics*, 12. <https://doi.org/10.3389/fpsyg.2021.602977>
- Grice, H. (1989). Logic and conversation. In H. Grice (Ed.), *Studies in the way of words* (pp. 22–40). Harvard University Press.
- Guglielmo, S., & Malle, B. (2019). Asymmetric morality: Blame is more differentiated and more extreme than praise. *PLoS One*, 14(3), e0213544. <https://doi.org/10.1371/journal.pone.0213544>
- Hare, R. (1952). *The language of morals*. Clarendon Press.
- Hare, R. (1963). *Freedom and reason*. Clarendon Press.
- Horn, L. (1989). *A natural history of negation*. University of Chicago Press.
- Horn, L. (2004). Implicature. In L. Horn & G. Ward (Eds.), *The Handbook of Pragmatics* (pp. 2–28). Blackwell Publishing Ltd.
- Hume, D. (1739). *A Treatise of Human Nature*. In David Fate Norton and Mary J. Norton (eds.), Oxford: Oxford University Press.
- Kneer, M. (2018). The norm of assertion: Empirical data. *Cognition*, 177, 165–171. <https://doi.org/10.1016/j.cognition.2018.03.020>
- Knobe, J., Prasada, S., & Newman, G. (2013). Dual character concepts and the normative dimension of conceptual representation. *Cognition*, 127(2), 242–257. <https://doi.org/10.1016/j.cognition.2013.01.005>
- Kyle, B. (2020). The expansion view of thick concepts. *Noûs*, 54(4), 914–944. <https://doi.org/10.1111/nous.12289>
- Marsili, N., & Wiegmann, A. (2021). Should I say that? An experimental investigation of the norm of assertion. *Cognition*, 212, 104657. <https://doi.org/10.1016/j.cognition.2021.104657>
- Mazzarella, D., & Gotzner, N. (2021). The polarity asymmetry of negative strengthening: Dissociating adjectival polarity from face-threatening potential. *Glossa: A Journal of General Linguistics*, 6(1), 47. <https://doi.org/10.5334/gjgl.1342>
- Moore, G. E. (1903). *Principia Ethica*. Cambridge: Cambridge University Press.
- Putnam, H. (2002). *The collapse of the fact/value dichotomy and other essays*. Harvard University Press.
- Reuter, K. (2019). Dual character concepts. *Philosophy Compass*, 14(1), e12557. <https://doi.org/10.1111/phc3.12557>
- Reuter, K., & Brössel, P. (2019). No knowledge required. *Episteme*, 16(3), 303–321. <https://doi.org/10.1017/epi.2018.10>
- Reuter, K., Löschke, J., & Betzler, M. (2020). What is a colleague? The descriptive and normative dimension of a dual character concept. *Philosophical Psychology*, 33(7), 997–1017. <https://doi.org/10.1080/09515089.2020.1817885>
- Roberts, D. (2013). It's evaluation, only thicker. In S. Kirchin (Ed.), *Thick Concepts* (pp. 489–520). Oxford University Press.

- Scheffler, S. (1987). Morality through thick and thin: A critical notice of ethics and the limits of philosophy. *The Philosophical Review*, 96(3), 411–434. <https://doi.org/10.2307/2185227>
- Smith, M. (2013). On the nature and significance of the distinction between thick and thin concepts. In S. Kirchin (Ed.), *Thick Concepts* (pp. 97–120). Oxford University Press.
- Sterken, R. (2017). The meaning of generics. *Philosophy Compass*, 12(8), 1–13. <https://doi.org/10.1111/phc3.12431>
- Sullivan, A. (2017). Evaluating the cancellability test. *Journal of Pragmatics*, 121, 162–174. <https://doi.org/10.1016/j.pragma.2017.09.009>
- Sytsma, J., Bluhm, R., Willemsen, P., Reuter, K. (2019). Causal attributions and corpus linguistics. In Fischer E. & Curtis M. Eds., *Methodological advances in experimental philosophy* pp. 209–238. Bloomsbury Academic.
- Thakral, R. (2018). Generics and weak necessity. *Inquiry*, 1–28. <https://doi.org/10.1080/0020174X.2018.1426683>
- Väyrynen, P. (2013). *The lewd, the rude and the nasty*. Oxford University Press.
- Väyrynen, P. (2021). Thick ethical concepts. E. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. <https://plato.stanford.edu/archives/spr2021/entries/thick-ethical-concepts>
- Wiggins, D. (1993). Cognitivism, naturalism and normativity. In Haldane & Wright. Eds., *Reality, representation, and projection* pp. 279–300. Oxford University Press.
- Willemsen, P., Baumgartner, L., Cepollaro, B., & Reuter, K. (2022). Evaluative deflation, social expectations, and the zone of moral indifference. *SSRN Electronic Journal*, <https://doi.org/10.2139/ssrn.4107428>
- Willemsen, P., Baumgartner, L., Frohofer, S., & Reuter, K. (2023). Examining evaluativity in legal discourse: A comparative corpus-linguistic study of thick concepts. In S. Magen & K. Prochownik (Eds.), *Advances in experimental philosophy of law*. Bloomsbury Publishing.
- Willemsen, P., & Kirfel, L. (2019). Recent empirical work on the relationship between causal judgments and norms. *Philosophy Compass*, 14(1), e12562. <https://doi.org/10.1111/phc3.12562>
- Willemsen, P., & Reuter, K. (2020). Separability and the effect of valence. In M. Mack, Y. Xu, & B. C. Armstrong (Eds.) *Proceedings of the 42th Annual Conference of the Cognitive Science Society 2020* (pp. 794–800). Cognitive Science Society.
- Willemsen, P., & Reuter, K. (2021). Separating the evaluative from the descriptive: An empirical study of thick concepts. *Thought*, 10(2), 135–146. <https://doi.org/10.1002/tht3.488>
- Williams, B. (1985). *Ethics and the limits of philosophy*. Harvard University Press.
- Zakkou, J. (2018). The cancellability test for conversational implicatures. *Philosophy Compass*, 13(12), e12552. <https://doi.org/10.1111/phc3.12552>
- Zakkou, J. (2021). Conventional evaluativity. *Australasian Journal of Philosophy*, 1–15. <https://doi.org/10.1080/00048402.2021.2013264>