Does Evaluative Language Provide Reasons to Act?

An Empirical Study of the Action-Guiding Potential of Evaluative Concepts

Pascale Willemsen (Pascale.Willemsen@uzh.ch)
University of Zurich, Department of Philosophy, Zollikerstrasse 117, 8008 Zurich, Switzerland

Judith H. Martens (Judith.Martens@uantwerpen.be)
University of Antwerp, Department of Philosophy, Rodestraat 14, 2000 Antwerp, Belgium

Abstract

What is the difference between language that describes the world and language that evaluates it? It has been suggested that an essential, distinguishing feature of evaluative language is its potential to guide actions by providing us with reasons to act. Calling an action "cruel" not only evaluates it negatively, its cruelty also provides us with a reason to refrain from it. Descriptive language, in and by itself, is relatively inert in this respect. In this paper, we examine whether this undisputed assumption is empirically adequate. We present three preregistered studies that demonstrate that evaluative language provides reasons for action when an agent contemplates how she should act, and also in conversational contexts. However, we also demonstrate that the speaker can easily deny the intention to provide such reasons to act.

Keywords: thick and thin concepts; evaluative language; action-guidingness; motivation; reasons for action

Introduction

Philosophers typically assume that there is a key difference between evaluative and descriptive terms. Descriptive terms describe what the world is like, and they can do so correctly or incorrectly. Evaluative terms, in contrast, communicate the speaker's stance or attitude towards the object he or she is evaluating, and they give away the speaker's norms and values. It has been further suggested that evaluative language provides *practical reasons*. A practical reason is a consideration that supports or counts in favour of performing a certain action (or having a certain attitude) or to refrain from that action.

While this assumption seems widely accepted and never explicitly questioned, no evidence has been presented in its support. In this paper, we present three preregistered studies suggesting that evaluative concepts indeed differ from descriptive concepts in that the former are readily interpreted as providing practical reasons to act. One of the most natural situations in which evaluative language is likely to realise its action-guiding potential is practical deliberation (*Study 1A*). When an agent contemplates which course of action they should choose, evaluating one of those actions in negative terms like "cruel", "selfish", or "bad" makes participants infer that the agent will count this as a reason not to choose that action. For descriptive terms, no such inference is made. The action-guiding potential is further realised in conversational context (*Study 1B*). When we hear a speaker say, "What you did there was cruel", we automatically infer that the speaker means to express their belief that something speaks against the behaviour in question and a different course of action is preferable. Descriptive terms like ”red" and "uncommon" do not have the same action-guiding potential.

As Study 1A and 1B demonstrate, participants reliably interpret statements containing evaluative terms as intended to provide reasons to act or to refrain from acting. Thus, evaluative language has action-guiding potential in the contexts we provided—namely, contemplating alone or with another person which course of action to choose. Also, this action-guiding potential is more pronounced compared to descriptive language. Study 2 tests how essential and irrefutable that action-guiding potential is. The fact that participants inferred that a speaker who uses evaluative language communicates reasons to act suggests that evaluative language somehow *implicates* these reasons. How is this implication communicated? And is it possible to use evaluative language without providing reasons to act? We address these questions in Study 2 using the cancellability test for conversational implicatures. Study 2 suggests that the action-guiding content of evaluative concepts is not lexically encoded but most likely to be conversationally implicated. A speaker can explicitly deny or cancel the implication that they wish to provide a reason to act by using an evaluative term. These findings challenge the traditional philosophical picture according to which there is a pretty strong and reliable connection between evaluative language and reasons to act.

Thin and Thick Evaluative Concepts, Action-Guidingness and Reasons for Action

Philosophers and linguists usually distinguish two types of evaluative terms and concepts: "thin" and "thick" terms (Eklund, 2011; Kirchin, 2013; Väyrynen, 2021). Thin terms and concepts evaluate an object as, for instance, "permissible", "right", "wrong", "good", "bad", or "blameworthy", yet they do not explicate in what way something is right or wrong. If a speaker evaluates an instance of lying as wrong, they convey no information as to why they think so. Thick terms and concepts do not merely evaluate but also provide information on why something is evaluated positively or negatively. Typical examples are *ethical* thick terms and concepts, such as “rude”, “cruel”, “courageous”, or “trustworthy”. Calling an agent courageous evaluates them positively for being willing to take risks – “reckless” also ascribes willingness to take risks yet assigns a negative evaluation to it (see also Baumgartner et al., 2022, and Willemsen & Reuter, 2021 for empirical investigations).

While there is considerable disagreement about how to delineate the boundary between thin and thick terms and concepts[1] (e.g., Chappell, 2013; Smith, 2013; Väyrynen, 2021), philosophers see a very clear difference to descriptive terms and concepts. Evaluative terms and concepts, so it is argued, have a close connection to reasons for action that is mostly absent from descriptive language (Heuer & Lang, 2012; Kirchin, 2013; Moore, 2006; Wiland, 2013; Williams, 1985). Heuer (2012, p. 220) makes this point explicitly: “They provide reasons for action— they are action-guiding […]: The cruelty of an action is a reason not to perform it or to prevent it; that an action is kind is a reason in its favor.” This idea seems to match our ordinary intuitions. Intuitively, expressing a sincere, whole-hearted negative evaluation of eating meat as wrong or cruel can be expected to correlate with a vegetarian lifestyle (for empirical studies on the role of motivation to act in accordance with one’s moral judgments, see, e.g., Björklund et al., 2012, and Björnsson et al., 2015; Nichols, 2002).

It might be objected that descriptive terms can provide reasons for actions as well. The fact that one store sells a product for a “cheaper” price than another might be a reason to buy the product in that store. A t-shirt being “red” is a reason not to wear it at a funeral. The key difference between terms that are categorized as evaluative and descriptive terms like “cheap”, or “red” is that the action-guiding potential of descriptive terms depends much more strongly on situational circumstances. A shirt’s color is not per se a reason not to wear it but only in the context of a funeral. In contrast, the “cruelty” of an action is said to always be at least a pro-tanto reason not to do it.[2]

Philosophers connect the action-guiding potential to the *evaluative* part of a concept, which thin and thick concept share. From this theoretical consideration, we might infer that the reasons for actions are expressed in the same way for thin and thick concepts. Surprisingly, philosophers have not explicitly committed to a view on this matter. However, at least from an empirical perspective, there are reasons to expect differences in how strongly, how reliably, or in what way they provide reasons for action. In addition to merely evaluating, thick concepts provide descriptive details about the considerations on which that evaluation is based. For instance, calling an action cruel communicates the action causing harm, suffering, or pain. Those details might affect the strength of the action-guiding potential and, as a consequence, how easy or hard it is for a speaker to deny it. In the following, we explore whether thin and thick concepts differ in their action-guiding potential.

Study 1A: Providing *Yourself* with Reasons for Action

One of the most natural situations in which evaluative language is likely to realise its action-guiding potential is practical deliberation. Imagine a situation in which an agent needs to decide which of several options she should choose. If evaluative language is action-guiding, we should expect that the fact that one option is evaluated as cruel or bad counts against taking it, while the option being honest or right speaks in its favour. If philosophers are correct, we should also expect that descriptive concepts do not work in this way.[3] In the first study, we test five groups of concepts against each other. We use thick and thin ethical concepts, and for each we use negative and positive items. We contrast these four groups of evaluative concepts with descriptive concepts. The design as well as our empirical predictions were preregistered with the OpenScience Framework.

Methods

We created a 5×1 between-subject design, with the independent factor *Category* (Thin Negative; Thin Positive; Thick Negative; Thick Positive; Descriptive). Each Category was exemplified by four different terms.

Thin Negative: bad, wrong, blameworthy, negative
Thin Positive: good, right, praiseworthy, positive
Thick Negative: rude, cruel, manipulative, selfish
Thick Positive: friendly, compassionate, honest, generous
Descriptive: pragmatic, uncommon, ordinary, conventional[4]

Participants were assigned to only one term. Since we are not primarily interested in the differences between the terms within one group of concepts, we collapse results for the four terms belonging to the same category. Participants were presented with the following prompt:

[1] Terms such as “just” and “fair” are richer than thin concepts, yet are not clearly thick concepts either (for a discussion see e.g., Väyrynen, 2021). We ignore the details of this debate in the following.

[2] For recent discussions of descriptive, so-called value-associated, and dual-character concepts and how to tell them apart, see Reuter, Baumgartner, & Willemsen, 2023.

[3] This claim in isolation might suggest something false. Descriptive features of an action, such as being spontaneous, can speak in its favour or against it depending on what the situation requires. A spontaneous expression of love might be better for being spontaneous, yet a business decision might be worse for being spontaneous.

[4] We applied the selection criteria for thick concepts specified by Willemsen & Reuter (2020). In short, the thick and thin concepts we chose are primary, most prototypical, examples in the philosophical literature which, at the same time, seem sufficiently colloquial and frequent in daily conversations (other than “lewd”, “chaste” or “galant”). Descriptive terms had to fit the syntactic structure of our stimulus phrases. Therefore, clearly descriptive terms such as “round” and “blue” could not be used. To ensure that those terms we selected were sufficiently descriptive or neutral, we checked their sentiment values using various sentiment dictionaries.

> Sally is struggling with a decision on how she should act. The situation is tricky, and Sally has several alternative options. To decide which option she should choose, Sally makes a list of things that count against and in favour of each of these options, and also of things that speak neither against nor in favour of these options.
>
> Sally thinks about Option A and writes down *"Doing this would be [term]".*

We recruited 400 participants (M_{Age} = 34 years; 37.8% male, 61.8% female, 0.5% non-binary). Participants were recruited on Prolific Academics (pre-selection criteria: Age: over 18, Native Language: English; Approval Rate: 90%).

Participants answered the following question on a 9-point Likert item, ranging from '-4 = strongly against it' over '0 = neither against nor in favour of it' to '4 = strongly in favour of it': "To what extent do you believe this will count against or in favour of Option A?" We tested the following predictions:

(H1): For positive thick terms, ratings will be significantly above the neutral midpoint (0).

(H2): For positive thin terms, ratings will be significantly above the neutral midpoint (0)

(H3): For negative thick terms, ratings will be significantly below the neutral midpoint (0)

(H4): For negative thin terms, ratings will be significantly below the neutral midpoint (0)

(H5): Ratings for positive thick terms will be significantly higher than ratings for descriptive terms.

(H6): Ratings for negative thick terms will be significantly lower than ratings for descriptive terms.

Results & Discussion

We conducted a global one-way ANOVA with 5 levels, namely Thick Negative, Thick Positive, Thin Negative, Thin Positive, and Descriptive. The mean ratings are depicted in Figure 1. There was a significant main effect of Category, $F(4, 395) = 146.67, p < .001$.

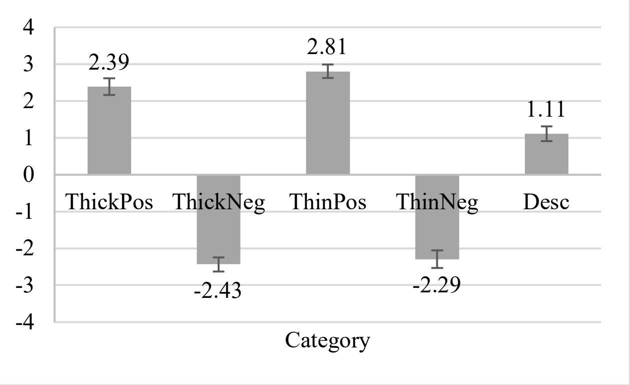

Figure 1: Mean ratings by concept class. Error bars represent the standard error around the mean

In line with our preregistered predictions, we conducted four one-sided t-tests against the scale's midpoint. Supporting **H1** and **H2**, mean ratings for positive thick concepts and for positive thin concepts were significantly above the neutral midpoint of 0 (positive thick: $M = 2.39$, t(76) = 10.62, $p < .001$; positive thin: $M = 2.81$, t(77) = 15.23, $p < .001$). Thus, participants believed that statements containing positive evaluative terms would count in favour of performing the action. In support of **H3** and **H4**, mean ratings for negative thick and thin terms were significantly below the neutral midpoint of 0 (negative thick: $M = -2.43$, t(83) = -12.69, $p < .001$; negative thin: $M = -2.29$, t(81) = -9.57, $p < .001$). As predicted (**H5** and **H6),** means for both thick positive and thick negative concepts differed significantly from means for descriptive concepts (positive: t(155) = 4.23, $p < .001$; negative: t(161) = -2.43, $p < .001$).

These results support the view that evaluative terms provide a reason for or, respectively, against performing the action, and that evaluative terms are indeed action-guiding.

Study 1B: Providing *Others* with Reasons for Action

Besides private, inner monologues in which an agent contemplates how they should act, evaluative concepts are often used in social interactions. To test whether evaluative concepts are considered action-guiding in conversational settings as well, we modified Study 1 accordingly.

Methods

We repeated the 5×1 between-subject design from Study 1A. Participants were presented with the same prompt, describing Sally contemplating how she should act. However, this time, it is not Sally who evaluates the action in private. Instead, it is now Amy who says to Sally, "Doing this would be [term]". Participants then answered the following question on a 9-point Likert item ranging from '-4 = strongly against it' over '0 = neither against nor in favour of it' to '4 = strongly in favour of it': *"To what extent do you believe that Sally will count this against or in favour of Option A?"*,

We recruited 405 participants (M_{Age} = 38 years; gender-balanced sample). Three participants had to be excluded for failing to finish the survey. We tested the same predictions **H1** to **H6** as in Study 1A.

Results & Discussion

We conducted a global one-way ANOVA with 5 levels, namely Thick Negative, Thick Positive, Thin Negative, Thin Positive, and Descriptive. The mean ratings are depicted in Figure 1. There was a significant main effect of Category, $F(4, 397) = 137.37, p < .001$.

In line with our preregistered predictions and parallel to Study 1A, we conducted four one-sided t-tests against the midpoint of the scale. In line with **H1** and **H2**, mean ratings for positive thick concepts and for positive thin concepts were significantly above the neutral midpoint of 0 (positive thick: $M = 2.39$, t(79) = 12.10, $p < .001$; positive thin: $M = 2.41$, t(80) = 16.41, *p* < .001). Thus, participants judged that statements containing positive evaluative terms count in favour of performing the action. These results match our findings in Study 1A.

Supporting **H3** and **H4**, mean ratings for negative thick and thin terms were significantly below the neutral midpoint of 0 (negative thick: *M* = -2.10, t(79) = -10.89, *p* < .001; negative thin: *M* = -2.16, t(79) = -10.69, *p* < .001). These results strengthen the view that negative evaluative terms provide a reason against performing the action and are in line with our findings in Study 1A.

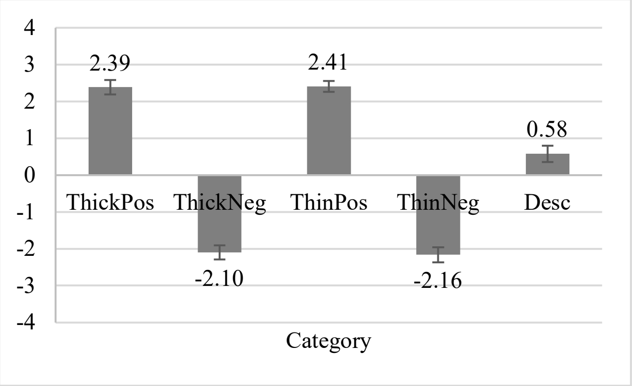

Figure 2: Mean ratings by concept class. Error bars represent the standard error around the mean

Finally, as in Study 1A, we also test whether thick terms differ from descriptive terms in virtue of their action-guiding potential, we compared mean ratings for both positive and negative thick concepts with descriptive concepts. Compatible with philosophical assumptions **H5** and **H6,** means for both thick positive and thick negative concepts differed significantly from means for descriptive concepts (thick positive: t(159) = 6.08, *p* < .001; thick negative: t(159) = -9.12, *p* < .001).

Study 2: Cancelling Reasons for Action

The first aim is to shed light on how reasons to act are conveyed. While philosophers have not been very explicit concerning the linguistic means by which the implication is communicated, we dare offer some informed guesses. Intuitively, one might think that a plausible mode of communication is lexical meaning. Evaluative concepts communicate, as part of their literal meaning, evaluative content and, by virtue of that evaluative content, also reasons to act. Once we have understood that "cruel" means something negative, we understand that this speaks against any cruel action. Such a position is likely to predict that action-guidingness is semantically entailed or presupposed. Therefore, any attempt to cancel it should result in a contradictory statement—just as, "Tom is a bachelor, but he is not a man/ not unmarried", would. Alternatively, one might think that action-guidingness is more loosely connected to evaluative terms, as it is conveyed as part of the context-dependent speaker meaning. Just as, "It is cold in here", could convey the request to close the window or turn up the heat, evaluative terms are interpreted with respect to what the speaker means to convey beyond what is literally said. In contrast to semantically entailed or presupposed implications, conversational implicatures like these can be cancelled (denied) without creating a contradictory statement: "It is cold in here, but don't close the window. We really need some fresh air."

The Cancellability Test for conversational implicatures (Grice, 1975; Sullivan, 2017; Zakkou, 2018) allows us to gain some first insights into how action-guidingness is communicated[5]. For this reason, we compare evaluative terms with standard examples of semantic entailments (SE) and generalised conversational implicatures (GCI), taken from the literature. We make the following predictions:

(H1): There will be a significant main effect of Category.

(H2): There is a significant difference between SE and GCI, such that Contradiction ratings are higher for SE than for GCI.

(H3): If action-guidingness is lexically encoded, the following hypothesis needs to hold: There is no significant difference between Evaluative Concepts and SE

(H4): If action-guidingness is pragmatically conveyed, the following hypothesis needs to hold. There is no significant difference between Evaluative Concepts and GCI.

While philosophers have not committed explicitly to either a lexical or pragmatic understanding of action-guidingness, they all seem to agree that action-guidingness is more essential and defining to evaluative than descriptive language. *The second aim of this experiment is to provide evidence for this view.*

One would assume that if action-guidingness can function as a useful tool to distinguish evaluative and non-evaluative language, then it should be more strongly and closely tied to the meaning of an evaluative term. It would follow that cancelling action-guidingness would seem only or more erratic and contradictory for evaluative terms than for descriptive terms. Therefore, in this study, we compare evaluative language (namely thin and thick concepts) with descriptive concepts, and we preregistered the following predictions:

(H5): Evaluative Concepts receive significantly higher contradiction ratings than Descriptive Concepts.

It has recently been found that evaluative concepts are asymmetrical in important respects. As, e.g., Willemsen and colleagues (2022) argue, the evaluation of a negative term is harder to cancel than the evaluation of a positive term. These findings inspire *two further exploratory inquiries* addressing

[5] For the application of this test to evaluative language, see, e.g. Almeida, Struchiner, & Hannikainen (2021); Baumgartner et al., 2022; Muth et al., 2020; Willemsen & Reuter, 2021.

two open questions. Since action-guidingness is closely intertwined with the evaluative content of the term, it is plausible that contradiction ratings also differ with polarity when we try to cancel action-guidingness. For this reason, we explore the following open question:

(H6): Do contradiction ratings for Positive Concepts (including Thin Positive and Thick Positive) differ significantly from contradiction ratings for Negative Concepts (including Thin Negative and Thick Negative)?

We also search for differences between thin and thick concepts and explore the following open question:

(H7): Do contradiction ratings for Thin Concepts (including Thin Negative and Thin Positive) differ from contradiction ratings for Thick Concepts (including Thick Negative and Thick Positive)?

Methods

We implemented a 7×1 between-subjects design, with the independent variable Category. We used four groups of evaluative concepts, namely thick positive, thick negative, thin positive, and thin negative, one group of descriptive concepts, and two control conditions, namely semantic entailments (SE) and generalised conversational implicatures (GCI). The stimuli for evaluative and descriptive terms are the same as in Study 1A and 1B. The stimuli for Generalised Conversational Implicatures and Semantic Entailments can be found in Willemsen & Reuter, 2021. We asked participants "Please imagine that Sally said the following sentence: […]". Here are some examples of the stimuli (sentences) that participants saw:

Evaluative and Descriptive: "What Robin did last week was [evaluative/descriptive term], but by that I am not saying that Robin should have acted [in a different way/in this way]."[6]

GCI (tried): "Robin tried to get into the club, but by that I am not saying that Robin failed to get into the club."

SE (couch): "This is a couch, but by that I am not saying that this is a piece of furniture."

Participants answer the question "Does Sally contradict herself?" on a 9-point Likert item, ranging from *'1 = definitely not'* to *'9 = definitely yes'*

As per our preregistration, we recruited 601 English native speakers (~80 per between-subject condition) with an approval rate of >80%, located in Australia, Canada, Ireland, the UK, or the US on Prolific Academics (gender-balanced sample, mean age 42 years). All participants were included in the analysis.

Results & Discussion

We conducted a global 7×1 ANOVA between-subjects Anova with the independent variable Category and the random factor Item. The mean ratings are depicted in Figure 3. There was a significant main effect of Group, $F(6, 593) = 44.00$, $p < .0001$, $\eta^2 = 0.31$ (supporting **H1**). As predicted (**H2**), a Bonferroni post-hoc test revealed that the mean contradiction ratings of Semantic Entailments ($M = 8.01$) were significantly higher than those of Generalised Conversational Implicatures ($M = 3.66$; $\Delta = 4.35$, [CI+ 3.14, CI- 5.56], $p < .001$). Testing **H3** and **H4**, Bonferroni post-hoc test showed that the mean contradiction ratings of Evaluative Terms (including Thick Negative, Thin Negative, Thick Positive and Thin Negative terms; $M = 3.80$) were significantly different from SE ($M = 8.01$, $\Delta = 4.22$, [CI+ -3.36, CI- 5.07], $p < .00$) but not from CI ($M = 3.66$; $\Delta = 0.14$, [CI+1, CI- -0.72], $p = 1$). These results support **H4** and speak against **H3**. Providing support for **H5**, mean contradiction ratings of Evaluative Terms ($M = 3.80$) differed significantly from those of Descriptive Terms ($M = 2.78$; $\Delta = 1.01$, [CI+ -1.90, CI- 0.13], $p < 0.00$).

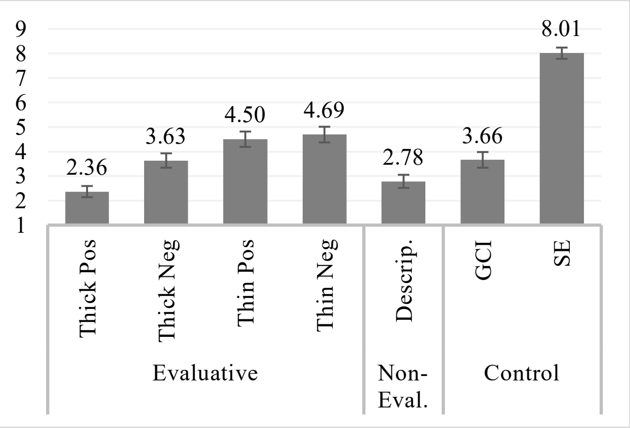

Figure 3: Mean contradiction ratings as a function of Category (Thick Positive, Thick Negative, Thin Positive, Thin Negative, Descriptive, Generalised Conversational Implicature, and Semantic Entailment). Error bars indicate the standard error around the mean.

Concerning **H6**, we did not formulate any concrete predictions, yet we examined whether our data show signs of a polarity effect. An independent sample t-test comparing positive with negative concepts revealed that there was a significant difference between contradiction ratings for positive and negative terms ($M_{positive} = 3.43$. $M_{negative} = 4.17$, t(339) = -2.43, $p = .016$). The action-guidingness of negative terms was harder to cancel than action-guidingness of positive terms. This effect is more pronounced for thick concepts ($\Delta = 1.27$ units) than for thin ones ($\Delta = 0.19$). There was also a significant difference between mean contradiction ratings for thick and thin concepts ($M_{thick} = 2.99$ $M_{thin} = 4.60$, t(339) = -1.61, $p < .001$), such that statements containing thin concepts were considered more contradictory when action-guidingness was denied **(H7)**.

[6] As for descriptive terms the valence of the action-guidingness is unclear (it can provide reasons to act in the same or a different way), we assigned half of our participants (n=10 per item) to the negative wording, the other half (n=10 per item) to the positive wording.

General Discussion

In this study, we provided some long-overdue empirical support for two philosophical assumptions: First, evaluative language is action-guiding and provides reasons for or against performing a specific action which is evaluated positively or negatively, respectively. Second, it is this action-guidingness that differentiates evaluative from descriptive language. To this end, we conducted three preregistered studies.

In **Study 1A**, we asked participants what they believe an agent is going to do with the insight that a certain course of action would be cruel, selfish, friendly, or honest. In **Study 1B**, we investigated the effect of another person making that evaluation. As we have hypothesised, participants believe that an agent evaluating an action in positive or negative terms is likely to consider this positive or negative evaluation a reason to act accordingly. Evaluations containing a positive term speak in favour of an action, while negative terms speak against it. This interpretation is equally available when a third-person utters the respective statement. Descriptive concepts, however, do not share that action-guiding potential to the same degree.

The aims of **Study 2** were a) to provide some initial evidence on whether action-guidingness is communicated by lexical or by pragmatic means, and b) to determine whether action-guidingness is more essential to evaluative than to descriptive language. The results suggest that the action-guiding potential of evaluative language can be explicitly cancelled without creating a contradictory statement. We find no differences between contradiction ratings of evaluative terms to conversational implicatures, yet a significant difference to semantic entailments. Of course, these results should not be taken as direct, indisputable evidence that action-guidingness is conversationally implicated. However, it is a plausible interpretation of our data that deserves further discussion. For the time being, and in the absence of evidence to the contrary, we believe that the burden of proof lies with scholars who think of action-guidingness as lexically encoded. For them, it might be difficult to explain the significant difference (4.22 units difference on a 9-point scale) to semantic entailments.

Here are two strategies available to defenders of a lexical view. Firstly, they might argue that contradiction ratings can be explained in a way which does not speak to how action-guidingness is encoded. We often have good reasons not to do things that could count as, say, honest or compassionate. An honest truth can be unnecessarily hurtful (no matter how honest we are), and giving a child all they want out of compassion is failing one's job as a parent. Participants in our study might have imagined a situation in which more general considerations outweighed the pro-tanto reason to act an evaluative concept would normally convey. Secondly, a more social explanation is available. Providing another person with reasons to act is not always considered appropriate. It depends strongly on one's own moral standing, the personal relationship, if others are listening in, etc. Participants might have imagined a context in which someone offers an opinion but considers it inappropriate to give the other person advice and request a change in behaviour. For the time being, we cannot rule out that either or a combination of these alternatives explains our data, and we suggest further research on this question.

The comparison between contradiction ratings for evaluative and descriptive language deserves some critical reflection. In line with the philosophical assumption that evaluative and descriptive language differ concerning their action-guiding potential, we *do* find a significant difference between the two groups. The difference also goes in the expected direction, with action-guidingness being harder to cancel for evaluative terms. However, two observations should shake philosophers' confidence in their own assumptions. First, contradiction ratings for both evaluative and descriptive terms are very low and do not exceed the neutral midpoint (5) of our scale. Overall, participants do not seem to find cancelling action-guidingness contradictory. Second, while the difference between evaluative and descriptive terms is significant, it is only small (1.01 units on a 9-point scale). From a theoretical standpoint, it is unclear how large a difference is required for action-guidingness to be a feature that distinguishes evaluative and descriptive language. However, we wish to suggest that 1.01 units on a 9-point scale might at best offer weak support of this idea.

Study 2 further revealed some interesting results that should come as a surprise to philosophers. First, inspired by previous research in experimental philosophy, we explored whether positive and negative evaluative terms differed in their cancellability behavior (see H6). Adding to the work of Baumgartner and colleagues (2022), Willemsen & Reuter (2021), and Willemsen and colleagues (2022), we find that polarity has an effect, and the action-guidingness is easier to cancel for positive than for negative terms. Future empirical research will need to confirm this finding and explain it by further exploring potential asymmetries between positive and negative language.

Secondly, Study 2 suggests that thin and thick concepts do not work alike. Cancelling action-guidingness seems harder for thin concepts than for thick concepts. At this point, we can only speculate why this effect occurs. It is possible that the differences in descriptive richness affect participants' inferences as to what the speaker means to convey. In addition to merely evaluating, thick concepts provide details about the considerations on which that evaluation is based. For instance, calling an action cruel communicates the action causing harm, suffering, or pain. Implicitly providing such details could draw attention to only one aspect of the action—its cruelty. In contrast, calling an action wrong or bad might be interpreted as a more general, overall evaluation of the action, all things considered. Whether this explanation is plausible will require further investigation. Such interpretations could be further explored by testing the action-guidingness of a term in contexts that vary in richness. Whether such differences in interpretation would matter philosophically deserves additional debate.

Acknowledgements

Pascale Willemsen was funded by the Swiss National Science Foundation (SNSF), grant number PZ00P1_201737.

We would like to express our gratitude to Glenn Anderau, Walker Siu Hong Au, Lucien Baumgartner, Ethan Landes, August Olsen, Kevin Reuter, as well as the participants of the Zurich XPhi Lab Meeting for their helpful feedback.

References

Almeida, G., Struchiner, N., & Hannikainen, I. (2021). Rule is a dual character concept. Available at SSRN 4018823.

Baumgartner, L., Willemsen, P., & Reuter, K. (2022). The polarity effect of evaluative language. *Philosophical Psychology*.

Björklund, F., Björnsson, G., Eriksson, J., Francen Olinder, R., & Strandberg, C. (2012). Recent Work on Motivational Internalism. *Analysis, 72(1),* 124–137. https://doi.org/10.1093/analys/anr118

Björnsson, G., Eriksson, J., Strandberg, C., Olinder, R. F., & Björklund, F. (2015). Motivational internalism and folk intuitions. *Philosophical Psychology, 28(5),* 715–734. https://doi.org/10.1080/09515089.2014.894431

Chappell, T. (2013). There are no thin concepts. In S. Kirchin (Ed.), *Thick Concepts.* Oxford University Press.

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

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Hrsg.), Speech Acts. BRILL. https://doi.org/10.1163/9789004368811_003

Heuer, U., & Lang, G. R. (edit.) (2012). *Luck, value, and commitment: Themes from the ethics of Bernard Williams.* Oxford University Press.

Kirchin, S. (2013). Thick concepts and thick descriptions. In S. Kirchin (Ed.), *Thick Concepts*. Oxford University Press.

Nichols, S. (2002). How psychopaths threaten moral rationalism: Is it irrational to be amoral? *The Monist, 85*, 285–304.

Moore, A. W. (2006). Maxims and thick ethical concepts. *Ratio, 19(2),* 129–147. https://doi.org/10.1111/j.1467-9329.2006.00315.x

Muth, C., Briesen, J., & Carbon, C. (2020). “I like how it looks but it is not beautiful”: Sensory appeal beyond beauty. *Poetics, 79*. Doi: 10.1016/j.poetic.2019.101376

Reuter, K., Baumgartner, L, & Willemsen, P. (2023). Tracing thick concepts through corpora. *Language & Cognition*.Smith, M. (2013). On the nature and significance of the distinction between thick and thin ethical concepts. . In S. Kirchin (Ed.), *Thick Concepts*. Oxford University Press.

Sullivan, A. (2017). Evaluating the cancellability test. *Journal of Pragmatics, 121*, 162–174. https://doi.org/10.1016/j.pragma.2017.09.009

Väyrynen, P. (2021). Thick Ethical Concepts. In E. N. Zalta (Edit), *The Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/archives/sum2019/entrieshick-ethical-concepts/

Moore, A. W. (2006). Maxims and thick ethical concepts. *Ratio, 19(2),* 129–147. https://doi.org/10.1111/j.1467-9329.2006.00315.x

Wiland, E. (2013). Williams on thick ethical concepts and reasons for action. In S. Kirchin (Ed.), *Thick Concepts.* Oxford University Press.

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

Willemsen, P., Baumgartner, L., Cepollaro, B., & Reuter, K. (2022). Evaluative deflation, social expectations and the zone of moral indifference. Available at SSRN: https://ssrn.com/abstract=4107428

Williams, B. (1985). *Ethics and the limits of philosophy*. Routledge.

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