Ergo AN OPEN ACCESS JOURNAL OF PHILOSOPHY

Moral and Moorean Incoherencies

ANDRÉS SORIA-RUIZ
Logos, University of Barcelona

NILS FRANZÉN
Umeå University

It has been argued that moral assertions involve the possession, on the part of the speaker, of appropriate non-cognitive attitudes. Thus, uttering 'murder is wrong' invites an inference that the speaker disapproves of murder. In this paper, we present the result of 4 empirical studies concerning this phenomenon. We assess the acceptability of constructions in which that inference is explicitly canceled, such as 'murder is wrong but I don't disapprove of it'; and we compare them to similar constructions involving 'think' instead of 'disapprove'—that is, Moore paradoxes ('murder is wrong but I don't think that it is wrong'). Our results indicate that the former type of constructions are largely infelicitous, although not as infelicitous as their Moorean counterparts.

The last twenty years have witnessed the rise of empirically oriented philosophy of language. But despite the fact that a central concern of metaethics is the meaning of moral terms, there are few empirical studies about the meaning of such terms.[1] Our purpose is to remedy this, by beginning to empirically investigate some of the central claims about moral discourse made within metaethical research.

In this paper, we investigate expressivism and the empirical claim that moral statements express non-cognitive states of mind. We present the first piece of formal empirical evidence in favor of the existence of a connection between the use of moral language and the practical attitudes of speakers. As motivational

[1] Some recent exceptions are Baumgartner et al. (2022), Soria-Ruiz et al. (2022), Willemsen and Reuter (2021).

Contact: Andrés Soria-Ruiz <andressoriaruiz@gmail.com>
Nils Franzén <nils.franzen@umu.se>

https://doi.org/10.3998/ergo.4665

internalists highlight, judging that murder is wrong involves being in a certain motivational, desire-like state towards murdering (see, e.g., Björnsson et al. 2015; Rosati 2016; Svavarsdottir 2006, for some overviews). We propose to treat that claim as a thesis about natural language, specifically as the thesis that assertions containing *moral adjectives carry an inference that the speaker is in a state of (dis-) approval towards whatever the moral adjective is applied to*. Why do assertions with moral adjectives carry such inferences, and what type of inference is this? These are the questions that we tackle in this paper.

To begin answering those questions, we present the results of four empirical studies, focused on the moral adjective 'wrong'. Their purpose is to assess the acceptability of constructions in which that inference is canceled, that is, sentences like 'murder is wrong but I don't disapprove of it'.

Our experiments compare the acceptability of those constructions to the acceptability of similar constructions involving the verb 'think', that is, Moore's paradoxes ('murder is wrong but I don't think that it is wrong'). The purpose of that comparison is to shed light on the Parity Thesis, the view that moral sentences express non-cognitive attitudes (like disapproval) in the same way that declarative sentences express cognitive attitudes (like belief, Schroeder 2008: 3). Jack Woods (2014) has made the following point: If declarative sentences containing, for example, 'wrong', express disapproval just as the use of any declarative sentence expresses belief, then one would expect that conjoining one such sentence with a negative ascription of disapproval should result in infelicity, just like conjoining any declarative sentence with a negative ascription of belief results in a Moore paradox.

Woods furthermore argues that this prediction is not borne out, since he finds constructions like 'murder is wrong but I don't disapprove of it' by and large acceptable. Taking Woods's observations as a starting point for our empirical investigation, our results paint a more complicated picture: Similarly to Moore's paradoxes, constructions like 'murder is wrong but I don't disapprove of it' are significantly less acceptable than similar constructions mentioning a different person than the speaker ('murder is wrong but Ann doesn't disapprove of it'). However, this effect is weaker for 'disapproval'-constructions than for Moore's paradoxes. Moreover, the infelicity of 'disapproval'-constructions is bolstered when we rule out so-called *exocentric* readings of moral sentences (Hare 1952; Lasersohn 2005; Stephenson 2007).

Here is a preview: We ran 4 studies.[2] The first, Preliminary Study was designed to compare the acceptability of Moore's paradoxes built with moral vs. non-moral predicates ('murder is wrong/common but I don't think that it is wrong/common'). Since no significant difference was found between these, in Study 1 we compared the acceptability of Moore paradoxes with moral predi-

[2] All the studies were pre-registered with the Open Science Framework (see below).

cates ('murder is wrong but I don't think that it is wrong') to similar constructions involving 'disapprove' instead of 'think' ('murder is wrong but I don't disapprove of it'). In Studies 2 and 3, we ran two variations of Study 1: In Study 2, we tested whether subjects might be interpreting the moral predicates in a way that allows for exceptions. To control for that potential interpretation, we added the qualification 'under any circumstance' throughout our test items ('murder is wrong under any circumstance but I don't disapprove of it under any circumstance'). However, we did not find this variation to have an impact on our results. In Study 3, we tested whether subjects might be interpreting the moral sentences *exocentrically*, that is, expressing someone else's moral views. To control for that, we added the qualification 'that's my opinion' throughout our test items ('murder is wrong, that's my opinion, but I don't disapprove of it'). Participants did find these constructions significantly less acceptable than those in Study 1.

We take our results to provide initial evidence in favor of the existence of a relatively robust connection between moral language and non-cognitive attitudes.

1. Background (I): The Disapproval Inference

In this section, we argue that the inference triggered by moral adjectives cannot be straightforwardly assimilated to other, well-known natural language inferences, such as entailment, presupposition, or implicature. In particular, it is not clear whether the inference can be canceled, which suffices to motivate an empirical investigation of this issue.[3] The linguistic correlate of the claim that moral judgment involves motivation is the claim that moral language triggers an inference that the speaker is in a motivational state, for example, a state of (dis-) approval. We will henceforth call this the *dis/approval inference*, (DI). We represent this inference with the symbol $\rightsquigarrow$:

1. Murder is wrong. $\rightsquigarrow$ the speaker disapproves of murder

DI does not fit squarely with other well-known types of natural language inferences, in particular entailments, presuppositions, and (conventional and conversational) implicatures.[4]

[3] The reader familiar with these inferences types and the tests standardly used to distinguish them may skip this section.

[4] A precursor to the following discussion is found in (Väyrynen 2013). Väyrynen is interested in the evaluative content of thick terms ('cruel', 'generous'), which he attempts to track through linguistic tests like the ones we describe here. See also Zakkou (2021) and Mandelkern (2021) for related discussions. Recently, Coninx et al. (2022) have empirically studied the implications of the folk concept of pain through similar devices.

First, consider entailments. A sentence such as (2) entails that someone had coffee.

2. Jakob had coffee. ↝ someone had coffee

Any entailment *e* of a sentence *s* is such that *e* cannot be false unless *s* is, on pain of contradiction. This implies that conjoining *s* with the negation of *e* results in a contradiction:

3. Jakob had coffee and no one had coffee.

Importantly, the contradictory flavor of (3) remains when it is embedded under the antecedent of a conditional (Yalcin 2007):

4. If Jakob had coffee and no one had coffee, then . . .

DI differs from entailments in this respect. First, it is contentious whether conjoining (1) with the negation of the purported inferences is felicitous—in fact, that is the topic of this paper.

5. Murder is wrong but I don't disapprove of it.

Secondly, the oddness of these constructions goes away when they are embedded under the antecedent of a conditional.

6. If murder is wrong but I don't disapprove of it, then …

(6) is markedly different from (4), in the sense that whereas (4) has a (saliently) impossible antecedent, (6) does not. It is possible to suppose that a certain activity is morally wrong even though I do not share that opinion. Regarding (6) for example, I might just have never thought of the moral status of murder and have no attitudes about it, and this is consistent with murder being wrong. Thus, the negation of DI is not logically inconsistent with the sentence that triggers it, suggesting that DI is not an entailment.

Secondly, consider semantic presuppositions. As is well known, presuppositions project out of entailment-canceling environments, such as negation, questions, or modals. In this respect, the DI of (1) is very much *unlike* a presupposition, since it does not project in these environments. We take this contrast to suggest that the DI is not a presupposition.[5]

[5] This observation also suggests that DI is not a conventional implicature, since conventional implicatures also project from entailing-canceling environments (see, e.g., Potts 2004).

7. Murder is not wrong. $\not\rightsquigarrow$ the speaker disapproves of murder

Finally, consider conversational implicatures. Conversational implicatures are standardly characterized as inferences triggered by utterances that flout one or more Gricean conversational maxims (Grice 1975/1993). A standard example are scalar implicatures, triggered when a speaker chooses a lexical item that sits below the top item on a scale of informativity, thereby suggesting that the top item does not apply.

8. Marcin ate some of the cookies. $\rightsquigarrow$ Marcin didn't eat all the cookies.

The inference in (8) can be canceled, which is a hallmark of conversational implicature:

9. Marcin ate some of the cookies, in fact he ate all of them.

DI also behaves differently according to this test. We take it that there is an intuitive contrast between the cancellation of a stereotypical conversational implicature, like (9), and the cancellation of DI in (5) (repeated below).

5. Murder is wrong but I don't disapprove of it.

While (9) is unproblematic, the acceptability of (5) is up for grabs.

2. Background (II): Expressivism and the Parity Thesis

Another potential diagnosis for the DI is that moral sentences *express* non-cognitive states, such as disapproval. The word "expresses", as attached to words and discourse, has been used to designate several different phenomena within philosophy of language and metaethics (see, e.g., Schlenker 2007, for an alternative use). Here, we are primarily interested in the relationship that is typically taken to obtain between an assertion of a proposition and the belief in that proposition. For instance, uttering the declarative sentence 'it is raining' *expresses* the speaker's belief that it is raining (e.g., Grice 1993: 42; Searle 1979: 3–5).

Within metaethics, it is common to point to the relationship of expression in the course of articulating a core tenet of metaethical *expressivism*, that is, the family of views according to which moral statements do not serve to describe the world, but rather to express some positive or negative attitude of the speaker. That core tenet can be articulated as the 'Parity Thesis', as formulated by Schroeder, which he and others take to be a defining feature of expressivism:

> **Parity**: '[M]oral sentences are related to noncognitive, desire-like states of mind in the same way that ordinary descriptive sentences are related to ordinary beliefs—they express them'. (Schroeder 2008: 3)

One can think of the Parity Thesis as one particular hypothesis about why statements to the effect something is wrong typically trigger the DI. Under this hypothesis, they do so because 'wrong'-statements express disapproval, similar to how assertions express beliefs. Call the latter inference the *belief inference* (BI).

1. Murder is wrong. ⇝ the speaker disapproves of murder (DI)

10. It's raining. ⇝ the speaker believes that it's raining (BI)

At a first approximation, the relationship of expression pertaining between assertions and beliefs seems to share important properties with the DI. First, the BI is not an entailment, as evinced by the felicity of (11) (Yalcin 2007):

11. If it's raining and I don't believe that it is raining, then I am misinformed.

Secondly, the BI does not seem to be a conventional implicature or presupposition, since it does not project out of negations and questions. While an assertion to the effect that it is raining communicates that the speaker believes that it is raining, an assertion to the effect that it is *not* raining does not communicate that.

Thirdly, another salient feature of the BI is that it is not cancellable, suggesting that BI is not an implicature (pace Schlenker 2016). Non-cancellability is evinced by the infelicity of Moore paradoxes like (12):

12. It's raining but I don't believe that it is raining.

The first two features are clearly shared by DI, but what about the third? Here is where Woods (2014) enters the picture. Woods (2014) draws attention to the fact that expressivism, through the Parity Thesis, makes testable predictions that can, and should, be tested empirically against our linguistic intuitions: if moral assertions express non-cognitive attitudes in the same way that assertions express beliefs, then the relationship holding between assertions and belief ascriptions should also hold between moral assertions and ascriptions of the relevant non-cognitive attitudes.

In particular, just as the right combinations of sentences and belief ascriptions generate Moore paradoxes such as (12), the right combinations of moral sentences and ascriptions of non-cognitive attitudes ought to generate analogously unacceptable constructions, as in (5):

5. Murder is wrong but I don't disapprove of it.

As pointed out above, our intuitions with respect to (5) are not very clear. Are these sentences felicitous or infelicitous? Accordingly, one finds both views represented in the literature. Woods (2014), argues that constructions similar to (5) are not unacceptable in the same way as (12), and thus that this prediction is not borne out. He argues that this constitutes a major blow to the Parity Thesis, and thus to expressivism (cf. also Franzén 2020; Väyrynen 2022). By contrast, other theorists have claimed that it is 'semantically inappropriate' to call an action wrong while not disapproving of it (Copp 2009: 187; cf. Joyce 2016: 29).

The main purpose of this paper is to experimentally test the acceptability of sentences like (5) above. In light of the previous discussion, it should be clear why we think this is important. Testing the acceptability of sentences like (5) is a way of testing the cancellability of the DI, and thereby, *inter alia,* of testing the feasibility of the Parity Thesis and expressivism.

It should also be noted that we focus on variations of expressivism which take the non-cognitive attitude expressed by wrong-statements to be disapproval. This version of expressivism is perhaps most closely associated with Simon Blackburn (1993). Versions of expressivism which take moral statements to express some other kind of attitude are not covered directly by our investigation. One reason for this is that Blackburn's kind of expressivism is the one that most naturally aligns with DI, the latter which we take to be an important phenomenon to investigate quite independently of any concerns about expressivism.[6]

[6] Moreover, we think that (dis-)approval-expressivism is among the best candidates for validating the Parity Thesis. For instance, consider the other major kind of contemporary expressivism, according to which moral statements express intentions or plans (Gibbard 2003). While, as we have pointed out, intuitions concerning the cancellability of the disapproval inference are divided, it seems clearer to us that one can felicitously assert:

(i) Murder is wrong, but I nevertheless plan to do it.

Having said this, it should be recognized that there are attitudes in the vicinity, like disliking, abhorring and despising, which could also form the basis of expressivist theories about moral language. While we know of no contemporary expressivism that takes one of these attitudes as a point of departure, it would be interesting to see how they fare in comparison to disapproval in eliciting Moorean infelicity. We leave this for future work. Another approach is that of Franzén (2020), who takes Mooren infelicity to be generated by constructions involving "to find", as in "I find lying morally wrong". Franzén's argument that states of finding are non-cognitive in nature raises complications that we want to avoid in the present context.

3. An Empirical Study of the Disapproval Inference

The discussion in §1 surrounding DI arrived at the following hypothesis: if DI is a conversational implicature, it should be cancellable. Thus, constructions like (5) ('murder is wrong but I don't disapprove of it') should be acceptable. The discussion in §2 surrounding the predictions of expressivism as an empirical thesis about moral language arrived at the Parity Thesis, which predicts that constructions like (5) should lead to infelicity in the same way as Moore paradoxes, like (12), do.

To assess the defectiveness of these constructions, we propose to compare them to similar sentences where a third person is mentioned instead of the speaker. If these constructions are defective, one ought to find a contrast between those constructions when they mention the speaker ('murder is wrong but I don't disapprove/think . . .') and when they mention someone else ('murder is wrong but *Ann* doesn't disapprove/think . . .').[7]

We ran 4 experiments using an 'anti-inference test' paradigm (Hansen & Chemla 2017), in which participants were asked whether sentences of the form '*p* but not *q*' made sense or not. We assume that positive answers signal that participants do not see any oddness in that sentence. By contrast, if they reply 'it doesn't make sense', we assume that they take the sentence to be defective in some way.

The first, Preliminary Study (§3.1) was designed to compare the acceptability of Moore's paradoxes built with moral vs. non-moral predicates. Given that we found no difference between these constructions, in Studies 1–3 (§§3.2–4) we adopted Moore's paradoxes built with moral predicates as a baseline, and we compared their acceptability to that of counterparts of such sentences involving disapproval instead of belief (e.g., (5)). Studies 2 and 3 involved two different variations of Study 1, designed to control for possible interpretations of the items in Study 1 that might be driving these results. So in addition to analyzing the results of each study, we also compared the results of Study 1 vs. Study 2, on the one hand, and Study 1 vs Study 3, on the other.

Before moving on, let us highlight two features of our studies. First, we only tested the predicate 'wrong' throughout. Thus, our data concern inferences attributable to this predicate and this predicate only. Secondly, participants' opinion

[7] One may wonder why we didn't directly compare the acceptability of sentences like (5) and Moore paradoxes. The reason is that, by doing so, we could perhaps find evidence of a difference in their acceptability, but if we did not find such evidence, we could not conclude that sentences like (5) and Moore paradoxes are alike. To assess the (potential) similarities of these two types of sentences, we chose to compare both of them to their third-personal counterparts, which are expected to be acceptable.

about the acceptability of these constructions could be influenced by their first-order opinions about the moral sentences that they include. For example, someone who thinks that murder is wrong might reject a sentence like (5), not on the basis of the construction itself, but because they think that whoever fails to disapprove of murder cannot be in their right mind. To control for participants' first-order opinions about the object of evaluation, in our test items we introduced four different objects of evaluation: negative (murder), positive (volunteering), controversial (eating meat), and something that participants could not be opinionated about ('what she did'). This way we ensured that participants' first-order opinions were not driving their assessment of constructions like (5).

3.1. Preliminary Study

This study was meant to test whether Moore-paradoxes containing moral sentences, such as (13), are infelicitous in the same way that stereotypical Moore-paradoxes—containing nonmoral predicates, such as (14)—are. To this effect, we measured the acceptability of sentences like (13) and (14) by comparing both of these constructions with minimally different alternatives containing a proper name instead of the first-person pronoun (15), (16):

13. Murder is wrong but I don't think that it is wrong.

14. Murder is common but I don't think that it is common.

15. Murder is wrong but Ann doesn't think that it is wrong.

16. Murder is common but Bill doesn't think that it is common.

The purpose of this study was to ensure that we could use Moore paradoxes like (13) as baseline in subsequent studies.[8]

3.1.1. Methods

Participants. We recruited 80 self-reported English native speakers via Prolific. Participants were paid £0.45 for approximately 3 minutes of their time (9£/h).

[8] Design, predictions, and statistical models were pre-registered anonymously with the Open Science Framework and can be accessed at: https://osf.io/328z6/?view_only=a76d275dfc884e55bb5d3074115036b2.

Per our preregistration, we excluded from the analysis participants who failed to show sensitivity to Moore's sentences with 'common'. That is, we excluded participants who accepted more than 50% of first-person 'common'-sentences, or rejected more than 50% of third-person 'common'-sentences. This resulted in the analysis of 49 participants. This exclusion criterion worked as an attention check.

Materials and design. Test items were generated from the following template, where φ is an act-type (the "object of evaluation"), x is an individual, and *ADJ* is an adjective:

- φ-ing is *ADJ* but x does not think that φ-ing is *ADJ*

This is a sentence stating (i) that an act-type falls under a certain adjective and (ii) that a certain individual does not think that. Adopting this template, we manipulated 2 factors with 2 levels each:

- Adjective (levels: 'wrong'/'common');
- Person (levels: first or third person; i.e., 'I' or 'Ann', 'Bill', etc.).

The combination of these two factors gives rise to the four sentence-types shown on **Table 1**, illustrated with 'murder' as act-type:

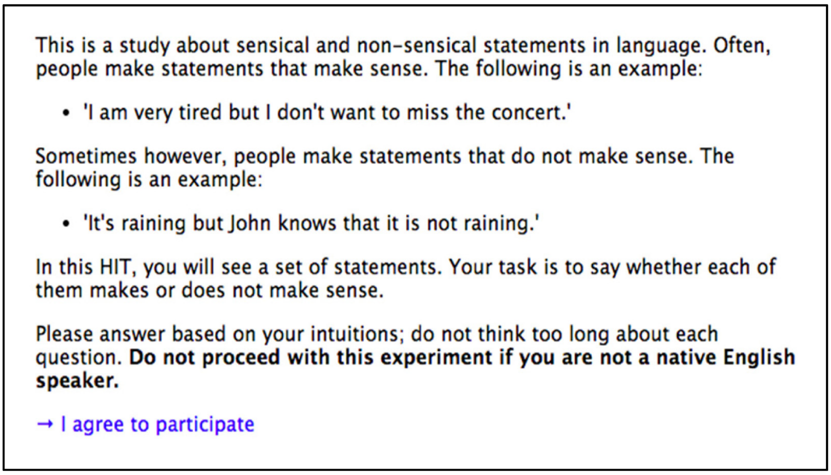
Person	**Adjective**	
	'wrong'	'common'
first	Murder is wrong but I don't think that it is wrong.	Murder is common but I don't think that it is common.
third	Murder is wrong but Ann doesn't think that it is wrong.	Murder is common but Bill doesn't think that it is common.

Table 1. 2×2 within subject design for the Preliminary Study.

The Preliminary Study consisted of 16 sentences generated by combining the 4 possible sentences types in **Table 1** with each of 4 act-types (murder, volunteering, eating meat, and 'what she did').

Procedure. The experiments were executed on PCIbex Farm (https://farm.pcibex.net/), to which Prolific workers were directed with a link. Participants saw an Instruction page, shown in **Figure 1**, and were shown items, as illustrated in

Figure 2. Importantly, the order of presentation of each of the 16 sentences was randomized for each participant.



This is a study about sensical and non-sensical statements in language. Often, people make statements that make sense. The following is an example:

- 'I am very tired but I don't want to miss the concert.'

Sometimes however, people make statements that do not make sense. The following is an example:

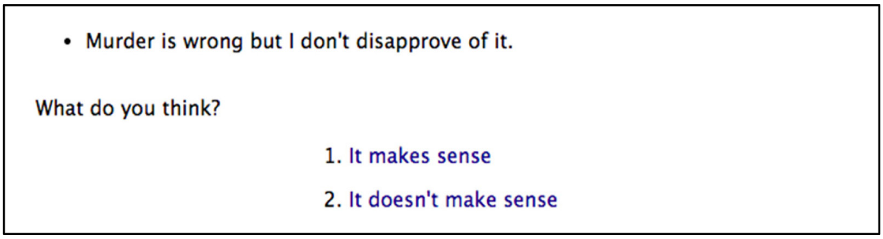
- 'It's raining but John knows that it is not raining.'

In this HIT, you will see a set of statements. Your task is to say whether each of them makes or does not make sense.

Please answer based on your intuitions; do not think too long about each question. **Do not proceed with this experiment if you are not a native English speaker.**

→ I agree to participate

Figure 1. Instruction page for the study



- Murder is wrong but I don't disapprove of it.

What do you think?

1. It makes sense
2. It doesn't make sense

Figure 2. Illustration of a critical trial used in the studies.

At the end, participants were asked for their native language, self-perceived gender (male/female/other), Prolific ID, and they were given the chance to leave a comment.

Predictions. For the Preliminary Study, we entertained two hypotheses:

> **Hypothesis (I)**: First-personal sentences are less acceptable than their third-personal counterparts. That is, we predict a main effect of Person;
>
> **Hypothesis (II)**: The difference in acceptability between first- and third-personal is smaller for 'wrong'-sentences than for 'common'-sentences. That is, we predict an interaction between Person and Adjective.

We would adopt Moore-paradoxes formed with 'wrong'-sentences as baseline for subsequent studies if we find evidence for **Hypothesis (I)** but not for **Hypothesis (II)**.

3.1.2. Results

The far-left plot on **Figure 3** shows the mean proportion of 'make sense' responses for *Person* and *Adjective* in the Preliminary Study. Data were analyzed using a logistic mixed regression that predicted participants' binary responses (1 if 'make sense', 0 otherwise) by *Person* (two levels: first and third), *Adjective* (two levels: 'common' and 'wrong') and their interaction.[9] Fixed effects were sum-coded. We included random by-subject and by-act-type intercepts, and random slopes by-*Person* and -*Adjective*.[10]

Here and throughout all the studies, our confirmatory analyses use Likelihood Ratio Tests where an omnibus model is compared to a simpler model, in which the relevant predictor is removed. In this case, these comparisons revealed a main effect of *Person*: first-person sentences were significantly less acceptable than third-person sentences across *Adjective* types ($\chi^2 = 65.441$; $p < 0.001$; ß = –2.02203); and no significant interaction between *Person* and *Adjective* was found ($\chi^2 = 0.5718$; $p = 0.4495$; ß = –0.09736).

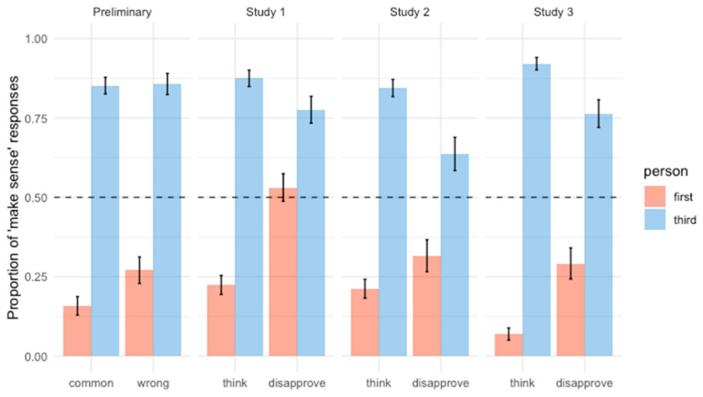


Figure 3. Results of the Preliminary Study and Studies 1–3. In the Preliminary Study there was a significant main Person effect, and no interaction between Person and Adjective. Throughout Studies 1–3, we found a main effect of Person as well as an interaction between Person and Attitude.

[9] All our analyses were carried on in the R environment (R Core Team 2014) using the lme4 software package (Bates et al. 2015)

[10] Due to lack of convergence, random slopes for act-type had to be removed.

3.1.3. Discussion

Since we found no interaction between the Person and Adjective factors, we proceeded to adopt the sentence pair (13), (15) as baseline for subsequent studies. Thus, in Studies 1–3 we compare the acceptability of (13) and (15) to that of similar constructions involving *disapproval* instead of *think*. Even though these results were somewhat predictable, it bears pointing out that we are offering—to our knowledge—the first piece of formal experimental evidence on Moore's paradox.

3.2. Study 1

In Study 1, we compared the acceptability of sentences like (13) and (5) (repeated below) by comparing each of these constructions with a minimally different alternative containing a proper name instead of the first-person pronoun ((15) and (17), respectively):

13. Murder is wrong but I don't think that it is wrong.

5. Murder is wrong but I don't disapprove of it.

15. Murder is wrong but Ann doesn't think that it is wrong.

17. Murder is wrong but Bill doesn't disapprove of it.

The objective throughout Studies 1–3 was to compare the degree of acceptability of first-person sentences like (13) and (5) to that of third-person sentences like (15) and (17), and to test whether the choice of attitude ('think'; 'disapprove') has an impact on their acceptability.[11]

3.2.1. Methods

Participants. We recruited 80 self-reported English native speakers via Prolific, who were paid £0.45 for 3 minutes approximately. Per our preregistration, we excluded from the analysis participants who failed to show sensitivity to Moore's paradox. That is, we excluded participants who accepted

[11] Design, predictions, and statistical models were pre-registered at: https://osf.io/6ghvs/?view_only=2bc9c83692b64517b427a9ea6786ab7c.

more than 50% of first-person 'think'-sentences, or rejected more than 50% of third-person 'think'-sentences. This resulted in the analysis of 48 participants.

Materials and design. Test items were generated from the following template, where φ is an act-type (the "object of evaluation"), x is an individual and A is an attitude verb:

- φ-ing is wrong but x does not A {that φ-ing is wrong / φ-ing}

This is a sentence stating (i) that an act-type is wrong and (ii) that an individual does not have a certain attitude towards that act-type. We manipulated 2 factors with 2 levels each:

a) Attitude (levels: 'think' and 'disapprove');
b) Person (levels: first or third person; i.e., 'I' or 'Ann', 'Bill', etc.).

This 2×2 within-subject design generates the 4 sentence-types shown in **Table 2**.

Person	**Attitude**	
	think	disapprove
first	Murder is wrong but I don't think that it is wrong.	Murder is wrong but I don't disapprove of it.
third	Murder is wrong but Ann doesn't think that it is wrong.	Murder is wrong but Bill doesn't disapprove of it.

Table 2. 2×2 within subject design for Study 1.

Study 1 consisted of 16 sentences generated by combining the four possible sentences types in **Table 2** with each of four act-types (murder, volunteering, eating meat, and 'what she did').

Procedure. The study was carried out in PCIbex Farm. The procedure was identical to the Preliminary Study (§3.1.1).

Predictions. We have two hypotheses (the first one is **Hypothesis (I)** from the Preliminary Study; the second one is the same as **Hypothesis (II)** except that we are now looking at the interaction between Person and Attitude, instead of Person and Adjective):

> **Hypothesis (I):** First-personal sentences are significantly less acceptable than their third-personal counterparts. That is, we predict a main effect of Person;
>
> **Hypothesis (II*)**: The difference in acceptability between first- and third-personal sentences is smaller for 'disapproval'-sentences than for 'think'-sentences. That is, we predict an interaction between Person and Attitude.

Both of these hypotheses are related to the Parity Thesis, but in different ways. **Hypothesis (I)** can be considered a corollary of Parity. If Parity is true, then one should find a contrast between first- and third-person constructions when looking at moral sentences paired with the relevant 'think'- and 'disapprove'-ascriptions. In other words, assuming that the phenomenon exists for 'think' (that is, Moore's paradox), introducing the relevant 'disapproval'-constructions should not make it go away.

Hypothesis (II*), by contrast, is not a corollary of Parity, but its denial might be seen as such: if Parity is true, one may expect to find *no evidence* of an interaction between 'think' and 'disapprove'. However, that is not a hypothesis that we can test, since we cannot conclude anything from the absence of an effect. At most, we can hypothesize that we will find an interaction, and then we can consider whether the existence of such an interaction implies that Parity is false. We return to this issue in Section 4.

3.2.2. Results

Figure 3 shows the mean proportion of 'make sense' responses for *Person* and *Attitude* in Study 1. A visual inspection suggests that 'make sense' responses are generally lower for first than for third person, and that this difference is weaker for 'disapprove'-sentences than for 'think'-sentences. A logistic regression[12] (see §3.1.2) statistically confirms these results, revealing a main effect of *Person* ($\chi^2 = 84.858$; $p < 0.001$; $ß = -1.4452$) as well as a significant interaction between *Person* and *Attitude* ($\chi^2 = 68.167$; $p < 0.001$; $ß = -0.6138$).

We also conducted a second logistic regression only on data from 'disapproval' trials (baseline level: first-person, treatment coded). This showed that the

[12] Due to lack of convergence, the model only included by-subject and by-object random intercepts, and random by-subject slopes for Person.

acceptability of first-personal 'disapproval'-sentences was not statistically different from chance (*ß* = 0.1427; *p* = 0.454).

3.2.3. Discussion

In Study 1, we found that first-person sentences are perceived as significantly less acceptable than third-person sentences. This confirms **Hypothesis (I)**. We also found that the contrast between first- and third-person constructions was significantly weaker for 'disapproval'- than for 'think'-sentences, thus confirming **Hypothesis (II*)** as well. In Studies 2 and 3, we aimed to control for two possible interpretations of (5) that might be driving these results.

3.3. Study 2

In Study 2, we assessed the possibility that speakers interpret 'wrong' in (5) with a *sotto voce* 'pro-tanto', so that it allows for circumstantial exceptions. The idea is that, when a speaker calls an action 'wrong', they might mean that it is wrong *in most cases*, or *in normal circumstances*. But then, it might be coherent for a speaker to utter (5). To see whether this interpretation played a role, we introduced the qualification 'under any circumstance' throughout our test items.[13]

3.3.1. Methods

Participants. We recruited 100 self-reported English native speakers via Prolific, who were paid £0.45 for 3 minutes approximately. As in Study 1, we excluded participants who were insensitive to Moore's paradox, which resulted in the analysis of 53 participants.

Materials and design. The materials were similar to that of Study 1, except for the fact that we introduced the aforementioned modification throughout our test items. Our design generates the 4 sentence-types shown in **Table 3**, illustrated with 'murder'.

[13] Design, predictions, and statistical models were pre-registered at: https://osf.io/568cz/?view_only=0fe4fcf94eb14f2eb6ac71b7920b147b.

Person	**Attitude**	
	think	disapprove
first	Murder is wrong under any circumstance but I don't think that it is wrong under any circumstance.	Murder is wrong under any circumstance but I don't disapprove of it under any circumstance.
third	Murder is wrong under any circumstance but Ann doesn't think that it is wrong under any circumstance.	Murder is wrong under any circumstance but Bill doesn't disapprove of it under any circumstance.

Table 3. 2×2 within subject design for Study 2.

Study 2 consisted of 16 sentences generated by combining the 4 possible sentences types in **Table 3** with each of 4 act-types (murder, volunteering, eating meat, and 'what she did').

Procedure. The study was carried out in PCIbex Farm. The procedure was identical to the Preliminary Study (see §3.1.1).

Predictions. Our hypotheses are the same as in Study 1, **Hypotheses (I)** and **(II*)**, with the addition of **Hypothesis (III)** regarding the comparison between Studies 1 and 2:

> **Hypothesis (III)**: The difference in acceptability between first- and third-personal 'disapproval'-sentences is smaller in Study 1 than in Study 2. That is, there is an interaction between Person and Study (levels: Study 1 & 2) among the 'disapproval'-sentences.

3.3.2. *Results*

Figure 3 shows the mean proportion of 'make sense' responses for *Person* and *Attitude* in Study 2. Visually, 'make sense' responses are generally lower for first than for third person, and the difference between first and third person looks weaker for 'disapprove' than for 'think'. Statiscally, a logistic regression[14] (see §3.1.2) revealed a main effect of *Person* ($\chi^2 = 44.754$; $p < 0.001$; $ß = -1.447198$) as well as a significant interaction between *Person* and *Attitude* ($\chi^2 = 33.58$; $p < 0.001$; $ß = -0.526604$).

[14] Due to model convergence, we had to drop the random slopes per attitude, keeping only by-subject random slopes per person.

A second logistic regression only on data from 'disapproval' trials (baseline level: first-person, treatment coded) showed that the acceptability of first-personal 'disapproval'-sentences was significantly below chance (*ß* = –1.5406; p = 0.00173).

In order to evaluate **Hypothesis (III)**, we considered the 'disapprove' trials from Studies 1 and 2. We fit a third logistic regression predicting participants' response by *Person*, *Study* (two between-subjects levels: *Study 1* & 2), and their interaction. We included by-subject random intercepts and by-person random slopes. No significant interaction was found between *Person* and *Study* (χ^2 = 0.0448; p = 0.8324; *ß* = 0.03348).

3.3.3. *Discussion*

In Study 2, we found that first-person sentences are perceived as significantly less acceptable than third-person sentences. This confirms **Hypothesis (I)**. We also found that the contrast between first- and third-person constructions was weaker for 'disapproval'- than for 'think'-sentences, confirming **Hypothesis (II*)** too. We did not, however, find evidence for **Hypothesis (III)**, and thus we could not conclude that the addition of 'under any circumstance' played any role in the interpretation of the sentences in Study 1.

3.4. Study 3

Finally, in Study 3 we aimed to control for the possibility that speakers might be interpreting 'wrong' *exocentrically*. This is in line with a response made to Woods (2014) by Toppinen (2015). Exocentric readings of evaluative predicates are such that they denote an opinion other than the speaker's, for example, the general opinion of society. Exocentric interpretations of subjective predicates are well-attested and can be easily brought about by context.[15] To see whether exocentric interpretations play a role in the interpretation of (5), in Study 3 we inserted the

[15] Exocentric readings of PPTs are described in Stephenson ('how's that new brand of cat food you bought?'—'I think it's tasty because the cat has eaten a lot of it', 2007: 498) and Lasersohn ('How did Bill like the rides?'—Bill's mum: 'The merry-go-round was fun, but the water slide was a little too scary', 2005: 672), among others. This type of non-evaluative usage of moral terms has long been observed in metaethics. Ayer (1936: 136) claimed that moral can terms express normative or purely *sociological* propositions, where the latter describe the practices of a particular society without endorsing them. Similarly, Hare characterized what he called 'inverted-commas' uses of evaluative adjectives, where we are 'not making a value-judgment ourselves, but alluding to the value-judgements of other people' (1952: 124). See discussion in Section 4.

appositive clause 'that's my opinion' after the moral sentence on each test item. We assumed that this rules out an exocentric reading of the moral sentence.[16]

3.4.1. Methods

Participants. We recruited 60 self-reported English native speakers via Prolific, who were paid £0.45 for 3 minutes approximately. As in Study 1, we excluded participants who were insensitive to Moore's paradox, which resulted in the analysis of 54 participants.

Materials and design. Materials were again similar to that of Study 1, except for the fact that this time we introduced a different modification throughout our test items. This design generates the 4 sentence-types shown in **Table 4**.

Person	**Attitude**	
	think	disapprove
first	Murder is wrong, that's my opinion, but I don't think that it is wrong.	Murder is wrong, that's my opinion, but I don't disapprove of it.
third	Murder is wrong, that's my opinion, but Ann doesn't think that it is wrong.	Murder is wrong, that's my opinion, but Bill doesn't disapprove of it.

Table 4. 2×2 within subject design for Study 3.

Study 3 consisted of 16 sentences generated by combining the 4 possible sentences types in **Table 4** with each of 4 act-types used in previous studies (murder, volunteering, eating meat, 'what she did').

Procedure. The study was carried out in PCIbex Farm. The procedure was identical to the Preliminary Study (§3.1.1).

Predictions. Our hypotheses are the same as in Study 1, **Hypothesis (I)** and **(II*)**, *plus* a cross-study hypothesis comparing Study 1 and Study 3—we note this as **Hypothesis (III*)**:

> **Hypothesis (III*)**: The difference in acceptability between first- and third-personal 'disapproval'-sentences is smaller in Study 1 than in Study 3. That is, there is an interaction between Person and Study (levels: Study 1 & 3).

[16] Design, predictions, and statistical models were pre-registered at: https://osf.io/u9m76/?view_only=a4a7e92baefb4ced8dcd6841d8a3e934.

3.4.2. Results

Figure 3 shows the mean proportion of 'make sense' responses for *Person* and *Attitude* in Study 3. Visually, 'make sense' responses are lower for first- than for third-person, and that difference looks weaker for 'disapprove'-sentences.

Statiscally, a logistic regression[17] (see §3.1.2) revealed a main effect of *Person* ($\chi^2 = 76.778$; $p < 0.001$; *ß* = –2.77254) as well as a significant interaction between *Person* and *Attitude* ($\chi^2 = 75.095$; $p < 0.001$; *ß* = –1.01261).

A second logistic regression only on data from 'disapproval' trials (baseline level: first-person, treatment coded) showed that the acceptability of first-personal 'disapproval'-sentences was significantly below chance (*ß* = –1.7284; p = 0.000488).

A third logistic regression (see §3.3.2) revealed a significant interaction between *Person* and the results of Study 1 & 3 ($\chi^2 = 9.8351$; $p = 0.001712$; *ß* = 1.21328).

3.4.3. Discussion

In Study 3, we found again that first-person sentences are perceived as significantly less acceptable than third-person sentences, which confirms **Hypothesis (I)**. We also found that the contrast between first- and third-person constructions was weaker for 'disapproval'- than for 'think'-sentences, confirming **Hypothesis (II*)** too. Finally, we found a stronger contrast between first- and third-person 'disapproval'-sentences in Study 3 than in Study 1, which is evidence for **Hypothesis (III*)**, thus suggesting that ruling out exocentric readings of 'wrong' played a role in our results.

4. General Discussion

4.1. Our Results and the Parity Thesis

Our results offer the first formal experimental evidence in favor of a connection between moral language and the ascription to the speaker of a non-cognitive attitude. First, and setting aside the Preliminary Study, in all three studies it was found that first-person sentences are significantly degraded in comparison to their third-person counterparts. This is so for 'think'- as well as for 'disap-

[17] Again, due to model convergence, we dropped the random slopes per attitudes, keeping only by-subject random slopes per person.

prove'-sentences. Thus, our results in all three experiments lend strong support to **Hypothesis (I)**.

Secondly, we also found an interaction in all three studies between the Person and Attitude factors, suggesting that the contrast in acceptability between first- and third-person constructions is significantly weaker for 'disapprove'- than for 'think'-sentences. In other words, we found evidence for **Hypothesis (II*)** as well.

Finally, cross-experimental comparisons yielded two results: First, we did not find evidence for **Hypothesis (III)**. That is, we did not find a significant interaction between the main Person effect for 'disapproval'-sentences and the results of Studies 1 and 2. Thus, we could not conclude that generic or *pro-tanto* interpretations play a role in the interpretation of the constructions that we tested. Secondly, we did find evidence for **Hypothesis (III*)**, that is, we found a significant interaction between the main effect of Person in 'disapproval'-sentences and Studies 1 and 3, due to a stronger main effect in Study 3. This suggests that exocentric readings play a role in the interpretation of moral sentences.

It is illuminating to frame these results in light of Woods's (2014) discussion. Recall that Woods claims that sentences like (5) ('disapproval'-sentences, repeated below) are not as bad as their corresponding Moorean counterparts ('think'-sentences):

5. Murder is wrong but I don't disapprove of it.

We think that a charitable interpretation of Woods's claim is along the lines of **Hypothesis (II*)** above. Under this interpretation, Woods's claim would predict a difference in acceptability between first- and third-personal 'disapproval'-sentences, on the one hand, and first- and third-personal 'think'-sentences, on the other (in virtue of a wider gap between 'think'- than 'disapproval'-sentences). The fact that we found a significant interaction between the Person and Attitude factors in all three studies bears this out: 'disapproval'-sentences are bad, but not as bad as Moore paradoxes. According to Woods, this spells trouble for the Parity Thesis and for the attempt to explain the DI by appealing to it. Judging from our results, DI is not *as strong as* BI, the relationship between assertions and beliefs.

That being said, we also found that sentences like (5) are significantly degraded in comparison to their third-personal counterparts, just like Moore paradoxes. That is, our results bear out **Hypothesis (I)**. To the extent that one may see **Hypothesis (I)** as a corollary of the Parity Thesis, we take this to lend support to Parity. At least, our results bear out the intuitions of those who claim that conjoining a moral judgment with a denial that one is in the relevant non-cognitive mental state is relatively infelicitous.

Finally, our results differ from experiment to experiment. This offers a possible diagnosis of what's driving the interaction between the Person and Attitude factors, that is, of why the effect is stronger for 'think'- than for 'disapproval'-sentences. As noted above, we did not find an interaction between Experiment 1 & 2 and the main Person effect, so we cannot conclude that a 'pro-tanto' interpretation of (5) had an impact on the acceptability of these sentences. But we did find an interaction between Experiment 1 & 3 and the main Person effect for 'disapproval'-sentences. That is, when exocentric readings are ruled out, participants found the relevant 'disapproval'-sentences significantly less acceptable than when such interpretations were available. The rest of the discussion will focus on this finding.

4.2. The Role of Exocentric Interpretations

Recall that the idea behind Study 3 was to check whether the acceptability of sentences like (5) might be explained by the fact that speakers reach for an exocentric reading of the moral predicate. We tried to control for this by testing a variation of (5), which should not allow for such readings:

18. Murder is wrong, that's my opinion, but I do not disapprove of it.

Our results show that indeed, a stronger effect was found when we ruled out exocentric interpretations in this way.

The relevance of this line of reasoning for the current discussion should be clear, but let us nevertheless spell it out. Exocentric readings have been characterized as expressing the opinion of someone other than the speaker (see n. 15). Equipped with the idea that ethical terms have such secondary reading, one could claim that some speakers find (5) acceptable because they reach for that interpretation. With (18), we tried to force the autocentric reading, the hypothesis being that speakers would then find the sentence infelicitous when the potential for accessing the exocentric reading was eliminated. And (18) is indeed found less felicitous than (5). This suggests that at least part of the reason why some speakers find (5) acceptable is that they are accessing an exocentric interpretation.

However, there are some potential problems with this line of argument. Woods, who argued for the acceptability of statements like (5), has also argued that this acceptability *cannot* be explained by the availability of an exocentric reading of the moral predicate (Woods 2014: 7–8). Woods makes two relevant points about this. First, Woods points out that

19. Murder is wrong but I don't think that it is wrong.

is infelicitous (something which is also corroborated by our experiment, see §3.2 above). But on the hypothesis that exocentric readings are easily available, one would expect such a reading to be possible in the first conjunct of (19) as well. Since this is not what happens, there is something wrong with our appeal to exocentric interpretations to account for the contrast between (5) and (18).

Here, a potential response would be that (19) is relevantly dissimilar to (5) in that the former has two occurrences of 'wrong'. To find a felicitous reading of (19), speakers would have to interpret the first occurrence of 'wrong' exocentrically, and the second autocentrically. It is not surprising, one can argue, that finding such divergent interpretations of two occurrences of the same predicate in one and the same sentence is difficult. We can see this by turning to non-controversial cases of context-dependent expressions:

20. Mary is tall but I don't think she is tall.

It is difficult to assign a different interpretation to each occurrence of 'tall' in (20). We think a similar diagnosis can be given for 'wrong', disallowing the possibility of interpreting one occurrence exocentrically and the next autocentrically. For this reason, it is to be expected that speakers find (19) less felicitous than (5).

Woods's second argument against the exocentricity hypothesis is that other predicates that allow for exocentric readings are nevertheless infelicitous in comparable constructions. Woods exemplifies this with 'delicious'. As other PPTs, 'delicious' allows for exocentric readings (see fn. 15). Nevertheless, the following construction is infelicitous:

21. Broccoli is delicious but I don't like it.

The point is that, if the availability of exocentric readings is what makes (5) felicitous for some speakers, then we should expect that a similar mechanism is activated for (21). But it is not. Moreover, note that the hypothesis given above for why (19) is infelicitous is not available for (21), since (21) contains only one occurrence of 'delicious'.

This point deserves more attention than we have space for here, but let us make two brief comments: it should first be acknowledged that our data only concerns 'wrong'. Regarding this predicate, we observe that exocentric readings are possible, and we use this possibility as a diagnosis for why a sentence like (5) might be acceptable. But our claim about the availability of exocentric readings of 'wrong' doesn't imply anything about the availability of exocentric readings of 'delicious', or any other predicate for that matter. Secondly, we are inclined to think that an exocentric reading of 'delicious' is more difficult to access than an exocentric reading of 'wrong'. In (tentative) support of this, note the following contrast:

22. Around here, pork is delicious.

23. Around here, eating pork is wrong.

It is easier to access an exocentric reading of the embedded sentence in (23) than in (22). Plausibly, this has to do with exocentric readings of taste predicates requiring a salient individual on which to locate the gustatory preference, whereas moral views are more naturally assigned to amorphous groups, that is, "people around here" or "society in general". Of course, this is just a hypothesis, and further empirical studies would be needed to show that this is really what explains the putative contrast between 'wrong' and PPTs like 'delicious'. In the present context, these considerations at least show that it is not evident that exocentric readings are equally accessible for 'delicious' as for 'wrong'.

5. Conclusion

To summarize, our study provides the first piece of formal empirical linguistic evidence in favor of a connection between moral language and the possession of non-cognitive attitudes on the part of the speaker. Our results show that pairing a moral sentence like 'murder is wrong' with the denial that the speaker disapproves of the action is significantly less acceptable than pairing that moral sentence with a similarly negative attitude ascription to someone else. These results mimic, although to a lesser extent, the result we get for Moorean constructions formed with the same moral sentences.

Moreover, we found that the infelicity of first-person 'disapproval'-constructions is stronger when exocentric interpretations are explicitly ruled out. In other words, we offer evidence that a qualified variation of (5) exhibits stronger paradoxical features. This is not surprising, if one thinks that part of the reason why speakers find constructions like (5) acceptable is that they are accessing an exocentric reading of its first conjunct, according to which it doesn't express the speaker's moral judgment, but someone else's.

Our results do not suffice, however, to conclusively decide whether the Parity Thesis is true nor to decide between a semantic or a pragmatic account of DI. Regarding the Parity Thesis, it really depends on how one understands it: if one takes it to entail **Hypothesis (I)**, according to which first-person 'disapproval'-sentences are significantly worse than their third-personal counterparts (just like Moore paradoxes), then we have found evidence for it. But if Parity is taken to demand that *no difference* in acceptability be found between "Moral" and Moorean constructions, that they pattern exactly the same, this prediction is not borne out.

Regarding the DI, our study delivers an equally mixed result, and leaves the issue in somewhat of a puzzle. On the one hand, as shown in Section 1, the inference does not seem to yield to standard linguistic diagnoses. On the other hand, the phenomenon is only partially similar to the expression-relationship, as shown in our experimental comparison with Moore's paradox. In addition, the observation that sentences like (5) are broadly infelicitous, might suggest that a pragmatic account is not supported. But properly assessing that possibility would require further studies. One—in our view—attractive path to decide this question would be to adapt Willemsen and Reuter (2021)'s paradigm, where they contrast certain constructions involving thick terms with paradigmatic cases of entailment and implicature cancellation, taken as control. We leave this work for the future.

Acknowledgments

We would especially like to thank Mora Maldonado for invaluable help with the experimental design, programming, and analyses. We are very grateful to Alex Dinges, Nate Hansen, Guido Löhr, Emanuel Viebahn, Julia Zakkou and Eduardo Pérez Navarro, two anonymous referees and an editor for comments on earlier drafts of this paper. We are also grateful to audiences at ECAP10, The 2021 Value in Language Workshop (Bratislava), and IFILNOVA Open Seminar (Lisbon). Andrés Soria-Ruiz acknowledges funding from COST Action CA17132, funded by the Horizon 2020 Framework Programme of the European Union. Nils Franzén acknowledges funding from the Swedish Research Council, grant number 2019-02905.

References

Ayer, A. J. *(1936). Language, Truth and Logic*. V. Gollancz.

Bates, D., M. Mächler, B. Bolker, and S. Walker (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*, 1–48. https://doi.org/10.18637/jss.v067.i01

Baumgartner, L., P. Willemsen, and K. Reuter (2022). The Polarity Effect of Evaluative Language. *Philosophical Psychology*. Advance online publication. https://doi.org/10.1080/09515089.2022.2123311

Björnsson, G., C. Strandberg, R. F. Olinder, J. Eriksson, and F. Björklund (Eds.) (2015). *Motivational Internalism*. Oxford University Press.

Blackburn, S. (1993). *Essays in Quasi-Realism*. Oxford University Press.

Coninx, S., P. Willemsen, and K. Reuter (2022). An Experimental-Linguistic Study of the Folk Concept of Pain: Implication, Projection, & Deniability. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *44*, 388–94.

Copp, D. (2009). Realist-Expressivism and Conventional Implicature. *Oxford Studies in Metaethics*, *4*, 167–202.

Franzén, N. (2020). Evaluative Discourse and Affective States of Mind. *Mind*, *129*(516), 1095–126. https://doi.org/10.1093/mind/fzz088

Gibbard, A. (2003). *Thinking How to Live* (1st paperback ed.). Harvard University Press.

Grice, P. (1993). *Studies in the Way of Words* (3rd print ed.). Harvard University Press.

Hansen, N. and E. Chemla (2017). Color Adjectives, Standards, and Thresholds: An Experimental Investigation. *Linguistics and Philosophy*, *40*, 1–40.

Hare, R. M. (1952). *The Language of Morals*. Clarendon Press.

Joyce, R. (2016). *Essays in Moral Skepticism*. Oxford University Press.

Lasersohn, P. (2005). Context Dependence, Disagreement, and Predicates of Personal Taste. *Linguistics and Philosophy*, *28*, 643–86.

Mandelkern, M. (2021). Practical Moore Sentences. *Noûs*, *55*(1), 39–61. https://doi.org/10.1111/nous.12287

Potts, C. (2004). *The Logic of Conventional Implicatures*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199273829.001.0001

R Core Team (2014). R: A Language and Environment for Statistical Computing.

Rosati, C. S. (2016). Moral Motivation. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2016 ed.). https://plato.stanford.edu/archives/win2016/entries/moral-motivation/

Schlenker, P. (2007). Expressive Presuppositions, *33*, 237–45. https://doi.org/10.1515/TL.2007.017

Schlenker, P. (2016). The Semantics–Pragmatics Interface. In M. Aloni and P. Dekker (Eds.), *The Cambridge Handbook of Formal Semantics* (664–727). Cambridge Handbooks in Language and Linguistics. Cambridge University Press. https://doi.org/10.1017/CBO9781139236157.023

Schroeder, M. (2008). *Being For*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199534654.001.0001

Searle, J. R. (1979). *Expression and Meaning: Studies in the Theory of Speech Acts*. Cambridge University Press.

Soria-Ruiz, A., M. Maldonado, and I. Stojanovic (2022). Good and Ought in Argumentation: COVID-19 as a Case Study. In S. Oswald, M. Lewiński, S. Greco, and S. Villata (Eds.), *The Pandemic of Argumentation* (43–64). Argumentation Library. Springer International. https://doi.org/10.1007/978-3-030-91017-4_3

Stephenson, T. C. (2007). Towards a Theory of Subjective Meaning (Doctoral dissertation). Massachusetts Institute of Technology.

Svavarsdottir, S. (2006). How Do Moral Judgments Motivate? In J. Dreier (Ed.), *Contemporary Debates in Moral Theory* (6–163). Wiley-Blackwell.

Toppinen, T. (2015). Relational Expressivism and Moore's Paradox. *Journal of Ethics and Social Philosophy*, *9*, 1–8. https://doi.org/10.26556/jesp.v9i2.173

Väyrynen, P. (2013). *The Lewd, the Rude, and the Nasty: A Study of Thick Concepts in Ethics*. Oxford Moral Theory. Oxford University Press.

Väyrynen, P. (2022). Practical Commitment in Normative Discourse. *Journal of Ethics and Social Philosophy*, *21*(2), Art. 2. https://doi.org/10.26556/jesp.v21i2.1484

Willemsen, P. and K. Reuter (2021). Separating the Evaluative from the Descriptive: An Empirical Study of Thick Concepts. *Thought: A Journal of Philosophy*, *10*, 135–46. https://doi.org/10.1002/tht3.488

Woods, J. (2014). Expressivism and Moore's Paradox. *Philosophers' Imprint*, *14*(5), 1–12.

Yalcin, S. (2007). Epistemic Modals. *Mind*, *116*, 983–1026. https://doi.org/10.1093/mind/fzm983

Zakkou, J. (2021). Conventional Evaluativity. *Australasion Journal of Philosophy*. Advance online publication. https://doi.org/10.1080/00048402.2021.2013264