# *Moral Disagreement and Moral Semantics*

# Moral disagreement and moral semantics

Justin Khoo
MIT

Joshua Knobe
Yale University

Forthcoming in *Noûs*

**Abstract**

When speakers utter conflicting moral sentences ("X is wrong"/"X is not wrong"), it seems clear that they disagree. It has often been suggested that the fact that the speakers disagree gives us evidence for a claim about the semantics of the sentences they are uttering. Specifically, it has been suggested that the existence of the disagreement gives us reason to infer that there must be an incompatibility between the contents of these sentences (i.e., that it has to be the case that at least one of them is incorrect). This inference then plays a key role in a now-standard argument against certain theories in moral semantics. In this paper, we introduce new evidence that bears on this debate. We show that there are moral conflict cases in which people are inclined to say both (a) that the two speakers disagree and (b) that it is not the case at least one of them must be saying something incorrect. We then explore how we might understand such disagreements. As a proof of concept, we sketch an account of the concept of disagreement and an independently motivated theory of moral semantics which, together, explain the possibility of such cases.

One of the most salient and fundamental facts about ordinary moral discourse is the fact that people disagree. Here is a simple example:

(1) Jim (*from American culture*): What Dylan did is morally wrong.
Yör (*from a very different culture*): No, what Dylan did isn't morally wrong.

In this case, it seems clear that Jim and Yör disagree. The key question now is how to capture this fact in a theoretical account of the semantics and pragmatics of moral statements.

One natural suggestion would be that the disagreement between the two speakers in some way arises from a conflict between the contents of their claims. To spell out this suggestion in a little bit more detail, we might say that part of what makes it the case that the two speakers disagree is that their claims are *exclusionary* in the following sense:

> Two claims are *exclusionary* (or have exclusionary content) iff it has to be the case that at least one of them is false.

There is certainly something intuitively compelling about this basic approach, and within the existing literature, it has led to a lively debate. Some theorists argue that we have reason to reject theories that do not posit exclusionary content in paradigm cases of moral disagreement (e.g., Moore 1922, Hare 1952, Horgan & Timmons 1990, Horgan & Timmons 1992, Smith 1994, Egan 2012), while others defend such theories by arguing for the plausibility of alternative ways of understanding the disagreement in such cases which are compatible with disagreeing speakers making non-exclusionary claims (e.g., Björnsson & Finlay 2010, Plunkett & Sundell 2013).

Our aim is to introduce some new evidence to this debate. Across a series of experimental studies, we show that people's judgments about exclusionary content systematically come apart from their judgments about disagreement. Specifically, in cases very much like the dialogue between Jim and Yör above, people show a tendency to say that the speakers do disagree but that their claims are not exclusionary.

This result turns the dialectic on its head. The previous debate was over whether it was problematic for a theory if it failed to predict exclusionary content in all cases of moral disagreement. We argue for the opposite conclusion: not only is it *not* problematic for a theory if it fails to predict exclusionary content in all cases of moral disagreement, but it *is* problematic for a theory if it *does* predict exclusionary content in all cases of moral disagreement. To accord with people's ordinary judgments, a theory should allow for the possibility of non-exclusionary moral disagreements.

We then sketch a theory that does accord with this aspect of people's ordinary judgments. We present a general account of the relevant sense of disagreement (§3) and an independently motivated theory of the semantics of moral claims (§4). Finally, we show that the combination of these two elements yields a view according to which there can indeed be non-exclusionary moral disagreements. We achieve this result by locating the disagreement between two speakers in their making incompatible proposals to change some aspect of their conversational context, rather than in their making claims whose contents are exclusionary.

# 1 The exclusion inference

We will be arguing that it is a mistake to suppose that all cases of moral disagreement must be cases that involve exclusionary content. Yet, though we believe that this view is mistaken, we do think that philosophers had good reasons for adopting it. We therefore begin by developing the best possible argument we can in favor of this view. Then, in the next section, we try to show that even this argument turns out to be unsuccessful.

To get a feel for the argument, consider the intuitive contrast between the following dialogues:

(2) a. Gabe: 3,677 is a prime number.
b. Amanda: No, 3,677 is not a prime number.

(3) a. Kara (in Australia): It's raining here.
b. Tim (in Boston): It's not raining here.

Looking at these two conversations, one sees a clear difference. In (2), we would ordinarily say that the two speakers disagree and that it would be appropriate for one to reject the other's claim using 'No.' By contrast, in (3), we would not ordinarily say that the speakers disagree, and it would not be at all appropriate for one to reject the other's statement using 'No.' The most natural way of explaining this difference between the two cases is that the speakers are making exclusionary claims in (2) but not in (3).

We are about to turn to moral examples, but before we do, we want to clarify the sense in which Gabe and Amanda disagree but Kara and Tim do not. An important question is whether by 'disagree' here we mean the notion of two speakers being a state of disagreement or engaging in the activity of disagreeing with one another (cf. Cappelen & Hawthorne 2009). There is a subtle difference between these notions, but the sense we are after is the activity notion of disagreement.[1] We take this point to be somewhat standard among

[1] We can begin to tease apart these notions with the following examples:

(i) A and B are in a discussion and both believe $p$. A then asserts $p$. B, trying to be polite, asserts $\neg p$.

(ii) A is in America talking with friends and B is in Russia talking with a completely different group of people. Both A and B believe $p$. A asserts $p$; B, trying to be polite to her audience, asserts $\neg p$.

In the first case, it seems like there *is* a sense in which A and B disagree, despite the fact that both believe $p$—they disagree in the activity sense in virtue of what each says. In the second case, there seems to be no sense in which A and B disagree, and this is because disagreeing in the activity sense requires being in a conversation with your disputant.

theorists making disagreement arguments, as such theorists typically presuppose that clear evidence that two speakers disagree is that they can appropriately reject each others' claims by saying "No" (cf. Stevenson 1937, Hare 1952, Gibbard 2003, Smith 1989, 1994, Wright 2001, Richard 2004, Lasersohn 2005, MacFarlane 2007, 2014, Stephenson 2007, von Fintel & Gillies 2008, 2011, Braun 2012, Willer 2013).[2] In what follows, we will use the expression "conversational disagreement" to pick out the relevant activity of disagreement that two people engage in by making claims to each other.

Now, turn back to the dialogue with which we began:

(1) a. Jim (from American culture): What Dylan did is morally wrong.
b. Yör (from a very different culture): No, what Dylan did isn't morally wrong.

It will be helpful to have a name for cases that are structurally like this one as we go forward—we will call them *moral conflict cases*. A moral conflict case is a conversation between two speakers in which one assertively utters a moral sentence and the other assertively utters its negation, and the two speakers thereby disagree with each other. We submit that the conversation between Jim and Yör in (1) is a moral conflict case.

The argument that moral conflict cases involve exclusionary content now is straightforward. One first observes that speakers in moral conflict cases disagree, and one needs an explanation for this fact. It seems that the best explanation is that these speakers are making claims with exclusionary content. Therefore, one should infer that these speakers are indeed making claims with exclusionary content. We will refer to this inference as the *exclusion inference*.

Perhaps unsurprisingly, the exclusion inference has played an enormously important role in existing work on the semantics of moral statements. It figures in work by a great many different philosophers and linguists, representing a striking consensus among researchers who are divided on many other issues; see for instance Moore (1922), Stevenson

[2]Some theorists run disagreement arguments by focusing on a sense of state disagreement that two speakers/thinkers may be in even if they are not in conversation with each other (cf. MacFarlane 2014: 182). We will not have much to say about this kind of disagreement argument. One issue is that the only way to verify whether two conversationally isolated speakers disagree is to directly consult our intuitions about whether they disagree. In such cases, we can't independently verify whether two such speakers disagree by consulting our intuitions about rejection. A second issue is that if the people disagreeing are not saying anything, then there is no obvious semantic upshot of their being in a state of disagreement. For these reasons, throughout this paper, we will focus entirely on activity disagreements in which the speakers make various claims. We will consider the question of how the state and activity senses of disagreement are related in §3.

(1937), Hare (1952), Lyons (1976), Williams (1986), Horgan & Timmons (1990, 1992), Smith (1989, 1994), Gibbard (2003), Streiffer (2003), Kölbel (2004), Richard (2004), Huemer (2005), Lasersohn (2005), MacFarlane (2014), MacFarlane & Kolodny (2014), Egan (2007, 2012), Dreier (2009).[3] Moreover, the exclusion inference has done a lot to shape recent work in moral semantics, with researchers showing considerable ingenuity in constructing theoretical frameworks that predict exclusionary content in moral conflict cases (e.g., Gibbard 2003, Egan 2012, MacFarlane 2014).

Before we begin arguing against the exclusion inference, we want to briefly draw attention to a fact that we hope will be acknowledged by researchers on both sides of the debate. Specifically, everyone should agree that not every case of disagreement has to be a case involving exclusionary content. There can be cases of various sorts in which two speakers disagree without making exclusionary claims. For one simple example, consider a case of imperative disagreement:

(4) a. Cody: Let's get a coffee.
b. Sally: No, let's get a beer.

In this case, Sally disagrees with Cody about whether to get a coffee, but the disagreement is not best understood in terms of exclusionary content. Since neither speaker makes any assertion, there are no claims here whose contents could be exclusionary. Or, for another example, consider cases of implicature denial (cf. Horn 1985, Horn 1989):

(5) a. Cody: John ate some of the cookies.
b. Sally: No, John ate *all* of the cookies.

Here too, it does seem that the two speakers disagree, but it does not appear that the actual contents of their claims are exclusionary. (Rather, Sally is disagreeing with the implicature of Cody's utterance.)

One way in which defenders of the exclusion inference could respond to these cases is by advancing a general principle that would tell us which cases of disagreement have to involve exclusionary content and which do not. In our view, however, it would not be fair to demand a general principle of this kind. Defenders of the exclusion inference can simply respond as follows: 'We are not claiming to have deduced a conclusion from a

[3]For instances of this argumentative strategy in other areas of semantics, see Wright (2001), Richard (2004), Lasersohn (2005), MacFarlane (2005, 2007, 2011, 2014), MacFarlane & Kolodny (2010), Stephenson (2007), von Fintel & Gillies (2011), Braun (2012), Willer (2013).

more general principle. Rather, we are making an inference to the best explanation. We observe disagreement across a range of different moral conflict cases, and we look for the best explanation of this disagreement. In the cases we are examining, we do not find any of the special features that explain the disagreements in the cases just discussed (imperatives, implicatures, etc.). Thus, we conclude that the best explanation is that the two speakers disagree.'

This response strikes us as a very reasonable one. Accordingly, in the argument that follows, we focus exclusively on paradigm cases of moral conflict. We try to show that the inference from disagreement to exclusionary content fails even in these cases.

## 2 Disagreement without exclusion

The argument we have been discussing rests on certain claims about how language actually works, and to put those claims to the test, we conducted a series of experimental studies. Before moving on to the details of those studies, however, we need to say a few brief words about the role of this experimental evidence in our argument as a whole.

First, a note about the questions these experiments are designed to address. Our aim is to explore questions about moral *language*. That is, we are concerned with questions about the semantic and pragmatic properties of the linguistic expressions people use to make moral claims. We claim that the results of the present experiments bear on questions of this type. Of course, there are also other important philosophical questions in the vicinity, including normative questions and metaphysical questions. We are not claiming that the experimental evidence provided here bears directly on those other questions.

Second, a note about the method we use to address these questions. We collect data about people's ordinary judgments and then use those data to argue for claims about moral language. It should be emphasized, however, that our argument does not in any way go through the assumption that people's ordinary judgments are *true*. Given the questions we are asking, we would only be warranted in concluding that people's ordinary judgments were true to the extent that we assumed that people had correct beliefs about certain normative and metaphysical matters. But we do not need to make that assumption; all we need to assume is that people have correctly grasped the meanings of certain English expressions. Drawing on this assumption, we can then use facts about their judgments to argue for claims about moral language.

To see the force of this point, consider a case that is closely analogous to the one we

examine here. Suppose that a team of researchers comes to the United States to study the semantics of the English word 'here.' As part of this research, they run an experiment in which they present native English speakers with the following case:

> Imagine that a person in one location says, "The speed of light here is 186,000 miles per second."
>
> Now imagine that a person in a different location says, "The speed of light here is not 186,000 miles per second."
>
> Now please tell us whether you agree or disagree with the following statement:
>
> *At least one of these two speakers must be saying something incorrect.*

The researchers might use people's responses to this question as part of an argument for a claim about the semantics of 'here.' But clearly, this argument would not have to rely on the assumption that people's responses were actually correct. For people to arrive at correct judgments, they would have to have certain correct beliefs about physics, but the argument does not assume that they do have such beliefs. All it assumes is that they have correctly grasped the meaning of a particular English word.

The present studies are in some ways more complex, but they nonetheless rely on this same basic method. We ask participants a series of questions. Then, drawing on facts about the patterns in their responses, we argue for certain conceptual and semantic claims. However, the argument does not rely in any way on the assumption that the responses given by participants are true. In the experiments we report here, participants give responses that some meta-ethicists would regard as false. Perhaps those meta-ethicists are right, and the participants' responses in these cases are indeed false. No matter. The argument should go through either way.

### 2.1 Experimental stimuli

A number of existing studies have explored people's intuitions about the conditions under which moral claims have exclusionary content (Beebe & Sackris 2014, Fisher *et al.* 2011, Goodwin & Darley 2008, 2012, Nichols 2004, Sarkissian *et al.* 2011). Although these studies use very similar methodologies, our own work will closely follow the methodology used by Sarkissian *et al.* (2011), and we therefore begin with a brief description of that earlier study and its results.

Participants in the study were given a vignette about an agent who performs a moral transgression. For example:

> Dylan buys a new knife and tests its sharpness by randomly stabbing a passerby on the street.

Participants were instructed to imagine two different speakers who made apparently conflicting claims about the moral status of Dylan's action. One utters the sentence "What Dylan did was morally wrong," while the other utters the sentence "What Dylan did was morally permissible." Participants were then asked a question to determine whether they thought these two claims were exclusionary. Specifically, they were asked whether at least one of the two speakers had to be wrong.

Each participant was randomly assigned to one of three possible conditions which differed only with regard to the identity of the speakers. In the **Same-culture** condition, both speakers were American university students. In the **Other-culture** condition, the speaker who thought the action was morally wrong was an American university student, while the speaker who thought it was permissible was a member of a traditional warrior culture in the Amazon. In the **Extraterrestrial** condition, the speaker who thought it was morally wrong was again an American university student, while the speaker who thought it was permissible was an extraterrestrial being with a psychology that differs radically from that of a human being. (Note that these conditions differ only in the identity of the speakers who make the judgments, not in the identity of the actual agent whose action is being judged.)

As Figure 1 shows, participants in these different conditions had quite different intuitions about whether or not the claims were exclusionary. Participants tended to agree that the claims were exclusionary in the **Same-culture** condition, were approximately at the midpoint in the **Other-culture** condition, and disagreed in the **Extraterrestrial** condition.

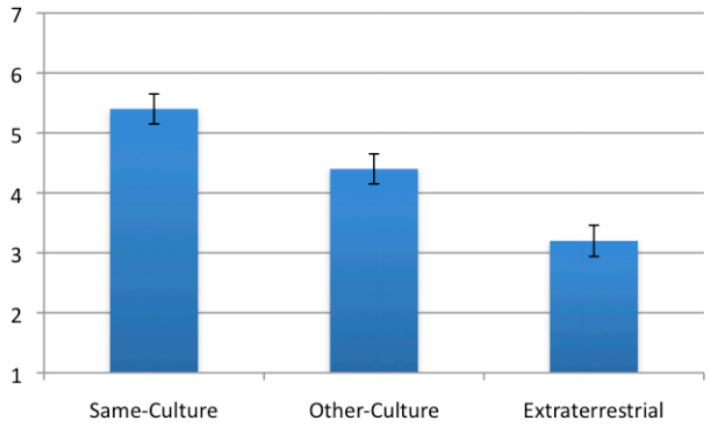


Figure 1. Mean agreement with the claim that 'At least one must be wrong' by condition. Error bars show standard error of the mean.

We should note in passing that this is precisely the pattern of results that would be predicted by certain versions of view traditionally known as 'moral relativism' (e.g., Harman 1975, Dreier 1990)[4], as well as contextualist theories of moral expressions (e.g., Björnsson & Finlay 2010, Finlay 2009, 2014). Suppose that each speaker's claim is evaluated relative to the norms of the speakers' cultures. Then, if two speakers are from very similar cultural backgrounds, one would expect that their claims to be evaluated relative to very similar norms. By contrast, if two speakers are from very different cultures, their claims would be evaluated relative to very different norms. The view thereby predicts ever lower ratings of exclusionary content as the speakers' cultural contexts become ever more different.

Let us now put this whole issue to one side. We will not be concerned with the question as to whether participants are actually right in their judgment that some of these claims are not exclusionary. Rather, we are simply using this experimental paradigm to get at certain conceptual and semantic questions. Above all, we will be concerned with the relationship between judgments of exclusionary content and judgments of disagreement. In cases like these, where participants think that the claims are not exclusionary, will they still think that the two speakers disagree?

[4]Although since Dreier's view is a version of speaker relativism (the view that speakers' moral claims are evaluated relative to the norms they themselves accept), some modification would be needed to get this result.

## 2.2 Experiment 1: Disagreement vs. exclusion

In this first experiment, we aim to see whether ordinary speakers' intuitions confirm the exclusion inference. To do so, we took the basic methodology of Sarkissian *et al.* (2011) but had each participant answer either a question about whether the two claims were exclusionary or a question about disagreement. We wanted to know whether there would be a significant difference between participants' responses to these different questions.

To resist the conclusion we draw here, one might question either our measure of exclusionary content or our measure of disagreement. Accordingly, below, we will discuss possible objections to our measure of exclusionary content (in the discussion section of Experiment 3) and to the measure of disagreement (in Experiment 2).

### 2.2.1 Methods

Two hundred forty-eight people filled out a brief questionnaire.[5] Each participant was randomly assigned to receive one of three possible vignettes (**Same-culture**, **Other-culture**, or **Extraterrestrial**) and to be asked one of two possible questions (**Incorrectness** or **Rejection**).

Participants were first introduced to the character who would appear in their vignette (an individual from their own culture, from another culture, or an extraterrestrial). They then received the following vignette.

> Sam is having a discussion with one of his classmates. Eventually, the conversation turns to a recent event. A person named Dylan bought an expensive new knife and tested its sharpness by randomly stabbing a passerby on the street.
>
> Sam says, about this case, "Dylan didn't do anything morally wrong."
>
> As it happens, Jim is listening in on Sam's conversation, and believes that Dylan did do something morally wrong. Jim jumps into Sam's conversation and says, "No, Dylan did do something morally wrong."

The vignettes in the **Other-culture** and **Extraterrestrial** conditions were exactly the same, except that Sam and his classmate were replaced by two individuals from a tribe in the Amazon (in the **Other-culture** condition) or by two extraterrestrials (in the **Extraterrestrial** condition).

[5] Participants were recruited using Amazon's Mechanical Turk. Sample was 74% male, mean age 28.

Within each vignette, participants were randomly assigned to receive one of two questions. Participants in the incorrectness conditions received a question of the form:

> Given that Jim and Sam have different judgments about this case, we would like to know whether you think at least one of their judgments must be incorrect, or whether you think both of them could actually be correct.
>
> Please tell us to what extent you agree or disagree with the following statement:
>
> *Since Jim and Sam have different judgments about this case, at least one of their judgments must be incorrect.*

Participants in the rejection conditions received a question of the form:

> Given that Jim and Sam have different judgments about this case, we would like to know what you think about Jim's response to Sam's claim.
>
> Please tell us to what extent you agree or disagree with the following statement:
>
> *Since Jim and Sam have different judgments about this case, it was appropriate for Jim to reject Sam's claim by saying "No".*

In all cases, participants answered on a scale from 1 ('Completely Disagree') to 7 ('Completely Agree').

### 2.2.2 Results

The mean responses for each question on each vignette are displayed in Figure 2:

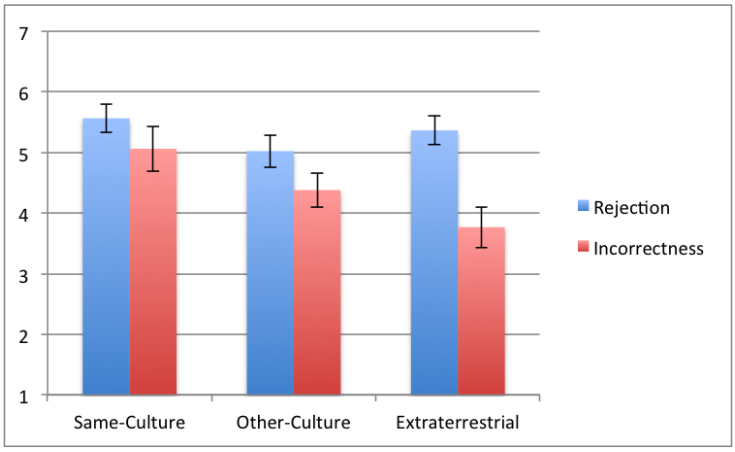


Figure 2. Mean responses by vignette and question in Experiment 1. Error bars show standard error of the mean.

The key thing to note here is the difference between responses to the incorrectness question and responses to the rejection question. Looking across the three different vignettes, we find an overall effect such that participants were more inclined to say that rejection was appropriate than they were to say that at least one of the two speakers had to be incorrect.[6]

We then explored the difference between the different vignettes (Same-culture, Other-culture, and Extraterrestrial). For the incorrectness question, people's judgments showed a significant difference between vignettes, much like the one seen in existing studies.[7] By contrast, for the rejection question, there was no significant difference between vignettes. Instead, participants tended to give high ratings across the board.[8]

As a result, the difference between the rejection question and the incorrectness question emerged in some vignettes but not in others. For the same-culture condition, there was no difference between responses to the two questions.[9] For the other-culture condition, the difference did not reach significance.[10] For the extraterrestrial condition, there was a highly significant difference, such that participants were inclined to say that rejection would be appropriate but were somewhat disinclined to say that at least one of the speakers had to be incorrect.[11]

### 2.2.3 Discussion

In this first study, we looked at a measure of exclusionary content (incorrectness judgments) and a measure of disagreement (rejection judgments). Friends of the exclusion inference might predict that these two kinds of judgments would show the same pattern, but that is not at all what we found. Instead, we found that the two types of judgments patterned

[6]The data were subjected to a 3 (vignette) x 2 (question) ANOVA. There was a significant main effect of vignette, $F(2, 248) = 3.8, p < .05$, and a significant main effect of question, $F(1, 248) = 15.2, p < .001$, but no significant interaction, $F(2, 248) = 2.1, p = .12$.

[7]Looking just at participants within the incorrectness conditions, we conducted a one-way ANOVA to examine the impact of vignette. There was a significant effect of vignette, $F(2, 110) = 3.6, p < .05$. Tukey's post-hoc tests found that the difference between same-culture and extraterrestrial conditions was statistically significant, $p < .05$, but that the other pairwise comparisons fell short of significance, both $p$s $> .3$.

[8]Looking just at participants within the rejection conditions, we conducted a one-way ANOVA to examine the impact of vignette. There was no significant effect, $F(2, 132) = 1.2, p = .29$.

[9] A planned comparison was used to compare incorrectness judgments ($M = 5.1, SD = 2.1$) with rejection judgments ($M = 5.6, SD = 1.7$), $t(83) = 1.2, p = .24$.

[10] A planned comparison was used to compare incorrectness judgments ($M = 4.4, SD = 1.8$) with rejection judgments ($M = 5.0, SD = 1.74$), $t(81) = 1.7, p = .10$.

[11] A planned comparison was used to compare incorrectness judgments ($M = 3.8, SD = 2.1$) with rejection judgments ($M = 5.4, SD = 1.5$), $t(78) = 3.9, p < .001$.

very differently.

Across all three vignettes, we found strong agreement with the claim that rejection would be appropriate. However, it does not appear that the best way to capture people's intuitions is to infer that in every moral conflict, the disagreeing parties make exclusionary claims. In fact, the opposite is true. People's responses show a clear divergence between intuitions about disagreements and intuitions about exclusionary content. Hence, the results of our first experiment challenges the exclusion inference: there seem to be cases in which speakers disagree by making non-exclusionary claims. Thus, any theory that predicts in every moral conflict that the two speakers make exclusionary claims will be going against people's ordinary intuitions.

## 2.3 Experiment 2: The direct method

In the previous experiment, we assessed intuitions about disagreement using an indirect method, namely, asking participants whether it would be appropriate for one speaker to reject the other's claim. As mentioned above, this is a very standard approach, which has been applied in numerous important papers in the existing literature. Still, one might well have doubts about this indirect method. That is, one might worry that people's answers to questions about the appropriateness of rejection are only an imperfect measure of their intuitions about disagreement. To address this worry, we conducted a second study in which we abandoned the indirect method and simply asked participants directly whether the two speakers disagree.

### 2.3.1 Methods

Five hundred twenty-one people filled out a brief questionnaire.[12] Participants were randomly assigned to receive a vignette about a speaker from a particular culture (**Same Culture**, **Other Culture**, or **Extraterrestrial**) and also to a particular question (**Incorrectness** or **Disagreement**). The procedure was exactly the same as that used in Experiment 1, with two exceptions.

First, we omitted the word 'No' from the statement made by the second speaker. (Thus, in this experiment, that speaker simply said, 'Dylan did do something morally wrong.') Second, and more importantly, there was a change in the questions participants received.

[12]Participants were recruited using Amazon's Mechanical Turk. Sample was 61% male, mean age 36.

Participants in the incorrectness condition were again asked to evaluate a statement of the form:

> *Since Jim and Sam have different judgments about this case, at least one of their judgments must be incorrect.*

However, this time, participants in the disagreement condition were directly asked to evaluate a statement of the form:

> *In making the claims they do, Jim and Sam disagree.*

Once again, all statements were evaluated on a scale from 1 ('Completely Disagree') to 7 ('Completely Agree').

### 2.3.2 Results

The mean responses for each question on each vignette are displayed in Figure 3:

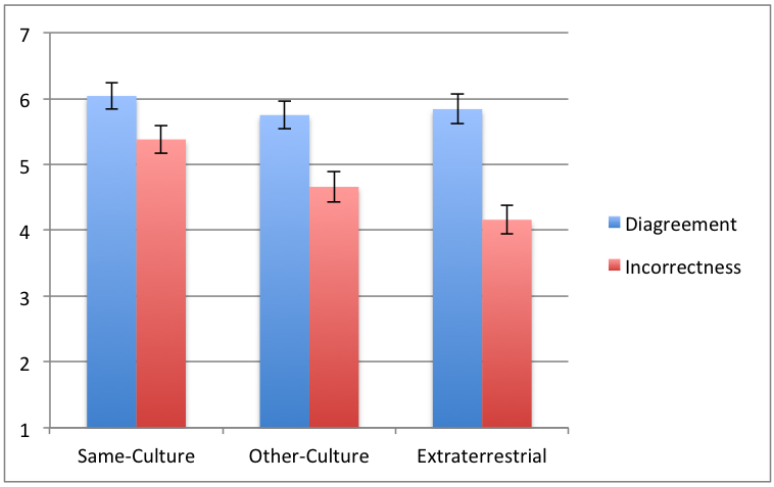


Figure 3. Mean responses by vignette and question in Experiment 2. Error bars show standard error of the mean.

As the figure shows, we again find a difference between judgments on our measure of disagreement and those on our measure of exclusionary content. Specifically, there was once again a significant effect whereby ratings were higher overall for the question as to

whether the two speakers disagreed than for the question as to whether at least one of them had to be incorrect.[13]

Just as in Experiment 1, we find ever greater differences between disagreement judgments and incorrectness judgments as we move to cultures that are ever more different from each other. However, because of the larger sample size used in this second experiment, this effect is actually statistically significant within each of the cultures considered separately. That is, there was a significant difference between judgments on the two questions in the same-culture condition,[14] in the other-culture condition[15] and in the extraterrestrial condition.[16]

### 2.3.3 Discussion

The results of these first two experiments are highly convergent. Regardless of whether disagreement is assessed indirectly (via intuitions about the appropriateness of rejections) or directly (via questions about whether the two parties 'disagree'), we find that participants are significantly more inclined to say that the speakers disagree than they are to say that the claims have exclusionary content. Again, the lesson seems to be that a theory predicting that in every moral conflict the two speakers make exclusionary claims will be going against ordinary people's intuitions.

## 2.4 Experiment 3: Moral vs. non-moral

In the first and second experiments, we found a surprising divergence between incorrectness judgments and disagreement judgments. We now want to ask whether that divergence reflects something special about moral claims or whether the same phenomenon would also arise for non-moral descriptive claims.

To address this question, we conducted a third study in which participants were assigned either to read about two speakers who expressed opposing moral claims (as in Experiment

[13] The data were subjected to a 3 (vignette) x 2 (question) ANOVA. There was a significant main effect of question, $F(1, 515) = 41.8, p < .001$, as well as a significant main effect of culture, $F(2, 515) = 6.1, p < .01$. The question x culture interaction fell just short of significance, $F(2, 515) = 2.9, p = .056$.

[14] A planned comparison was used to compare disagreement judgments ($M = 6.04, SD = 2.0$) with incorrectness judgments, ($M = 5.4, SD = 2.1$), $t(190) = 2.3, p = .02$.

[15] A planned comparison was used to compare disagreement judgments ($M = 5.8, SD = 1.9$) with incorrectness judgments, ($M = 4.7, SD = 2.0$), $t(155) = 3.5, p = .001$.

[16] A planned comparison was used to compare disagreement judgments ($M = 5.8, SD = 1.9$) with incorrectness judgments, ($M = 4.2, SD = 2.1$), $t(170) = 5.4, p < .001$.

1) or about two speakers who expressed opposing non-moral descriptive claims. The key question was whether the divergence we observed in the first experiment would arise only for the moral claims or whether it would arise for the non-moral claims as well.

### 2.4.1 Methods

Two hundred ninety-nine participants filled out a brief questionnaire.[17] Each participant was randomly assigned to receive either the moral vignette or the non-moral vignette and to receive either the incorrectness question or the rejection question.

All participants received a brief introduction to the extraterrestrials. Participants in the **moral** conditions received exactly the materials used within the extraterrestrial condition of Experiment 1. Thus, they received a vignette about a moral transgression, and asked either a question about incorrectness or a question about rejection.

Participants in the **non-moral** conditions instead received the following vignette:

> Two Pentars, Zog and Zar, are having a discussion. Eventually, the conversation turns to the famous French general Napoleon Bonaparte.
>
> Zog says, "Napoleon always used to go into battle on a helicopter."
>
> As it happens, Jim is listening in on the conversation, and believes that Napoleon never went into battle on a helicopter. Jim jumps into Zog and Zar's conversation and says, "No, Napoleon never went into battle on a helicopter."

These participants were then asked whether, given that Zog and Jim had opposite opinions, at least one of them must be incorrect (in the incorrectness condition) or whether it was appropriate for Jim to reject Zog's claim by saying 'No' (in the rejection condition).

### 2.4.2 Results

The mean responses for each condition are displayed in Figure 4.

[17] Participants were recruited through Amazon's Mechanical Turk. 66% male, mean age 29.

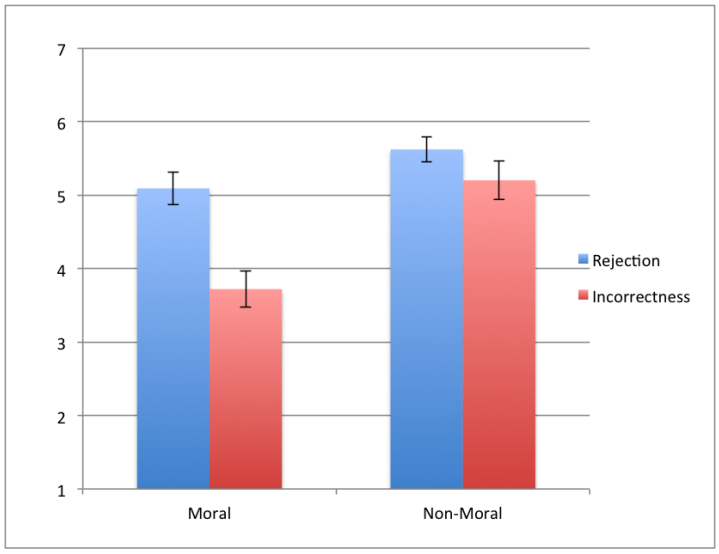


Figure 4. Mean responses by vignette and question in Experiment 3. Error bars show standard error of the mean.

As the figure shows, the responses obtained for the non-moral vignette were very different from those obtained for the moral vignette.[18] For the moral vignette, we replicated our previous finding: people gave higher ratings on the rejection question than they did on the incorrectness question.[19] By contrast, for the non-moral vignette, people gave high ratings on both questions, and there was no significant difference between them.[20]

### 2.4.3 Discussion

In our third and final experiment, we again found a divergence between intuitions about exclusionary content and intuitions about disagreement for the moral claim, but we did not find any such divergence for the non-moral claim. This result indicates that the divergence observed here depends in some way on specific features that can be found in certain moral conflicts but not in all cases of disagreement.

Taken together, these three experiments provide evidence for a claim about people's ordinary understanding of moral conflict cases. Specifically, they suggest that people un-

[18] Results were subjected to a 2 (vignette: moral vs. non-moral) x 2 (question: incorrectness vs. rejection) ANOVA. There was a main effect of vignette, $F(1, 295) = 19.5, p < .001$, a main effect of question, $F(1, 295) = 15.5, p < .001$, and a significant interaction, $F(1, 295) = 4.4, p < .05$.

[19] Rejection: $M = 5.1, SD = 1.7$. Incorrectness: $M = 3.7, SD = 2.2, t(144) = 4.1, p < .001$.

[20] Rejection: $M = 5.6, SD = 1.6$. Incorrectness: $M = 5.2, SD = 2.2, t(151) = 1.4, p = .18$.

derstand some moral conflict cases as being cases of non-exclusionary moral disagreement. Our aim in the remainder of this paper will be to work out the broader implications of this key finding.

Before moving onward, however, we pause briefly to ask whether there is any way of resisting the basic claim that these experimental results are best understood in terms of people seeing certain cases as involving disagreement without exclusionary content. As far as we can see, the most plausible objection is that people are adopting some interpretation of the word "incorrect" as it appears in our experimental stimuli such that these stimuli do not adequately tap into people's intuitions about exclusion. This objection might take either of two basic forms.

The first is what one might call the 'epistemic interpretation.' It might be thought that participants did not interpret our incorrectness question as being about whether certain claims were false but instead interpreted it as being about something more epistemic (e.g., whether the speaker was justified in saying what she did). This objection is certainly plausible on its face, but existing studies using the same method employed here provide some evidence against it. When participants were given a case in which a speaker said something that was justified but false, they tended to disagree with the claim that the speaker had "no good reason," but they nonetheless agreed with the claim that what the speaker said was incorrect (Sarkissian *et al.* 2011: 498-500). Thus, existing experimental evidence suggests that this method is not merely tapping into something about participants' judgments concerning the epistemic status of certain claims.

The second is what one might call the 'pragmatic interpretation.' It might be thought that participants did not interpret our incorrectness question as being about the literal semantic content of what the speaker said but instead interpreted it as being about something that the speaker pragmatically implicated. We certainly acknowledge that this is, in general, a serious worry about the method employed here, but we do not think that it is very plausible as an alternative explanation of these specific findings. It seems highly plausible that there might be cases in which people regard the literal content of a sentence as true but, because the sentence has a false implicature, still declare it to be "incorrect." This, however, is not what one would have to say to generate an alternative explanation of the present findings. Instead, what one would have to say is that (a) people regard the claim that a particular act of murder is morally wrong as false but (b) because this claim has a true implicature, they are reluctant to declare it "incorrect," and yet (c) despite all that, they still think that this claim should be rejected. Although it is in principle possible

that all this will turn out to be right, it does not appear to be an especially promising approach to explaining our present findings, and we will not be pursuing it further here.

This concludes our examination of the experimental evidence. In what follows, we take it that the studies reported here provide evidence in favor of the view that people regard some moral conflict cases as cases of non-exclusionary moral disagreement, and we ask what implications this finding might have for moral semantics.

## 3 Making sense of non-exclusionary disagreement

The results of our three experiments are evidence that in some moral conflict cases, the disagreeing parties do not make exclusionary claims. This leaves us with a puzzle: what could such people be disagreeing about, if not the truth of their respective claims? In this section, we propose a hypothesis about conversational disagreement which allows for the possibility of moral conflict cases in which the disagreeing parties do not make exclusionary claims.[21]

Here is the basic idea. Claims have content—what is being claimed—and contents are the kinds of things that may be true or false (the content of a claim is the content of the sentence used to make the claim in the context in which it is made). But to make a claim by uttering a sentence is to do something, and our strategy will be to understand what it is to disagree by making certain claims in terms of what we do when we make claims.

Within contemporary semantics (following Stalnaker 1978, 1999, 2002, Lewis 1979), it is common to understand what people are doing with their speech acts in terms of

[21] We pause to consider briefly the relationship between the state and activity notions of disagreement. Before learning the results of our experiment, one might have thought it plausible to ground activity disagreements in states of disagreement, and think that the latter requires endorsing exclusionary contents. The results of our experiment suggest that at least one of these claims is false. But then what is the relationship between activity and state disagreements? One interesting idea in line with our experimental results is to invert the order of explanation and explain state disagreements in terms of activity disagreements. A version of this strategy would be to understand what it is for A to be in a state of disagreement with B in terms of something about A's disposition to reject certain claims B has made or is disposed to make. One interesting upshot of this proposal that it predicts state disagreements may be asymmetric, which may turn out to be right. Suppose Sue has a mistaken belief about Jim's last name—so, their beliefs about Jim's last name are exclusionary. Nonetheless, Sue is not disposed to reject any claim of Jim's about his last name—on the contrary, her belief on this matter is deferential to Jim. Suppose Jim and Sue are not in any conversation at the moment. The above proto-theory of state disagreement would predict about such a case that Sue doesn't disagree with Jim (about what his last name is), even though it predicts that Jim disagrees with Sue (about this). This seems correct to us, but we leave arguing for such a view aside for another day.

proposals to *update* some parameter(s) of their speech context. In particular, the speech act of assertion is often understood as a proposal to update the common ground of the conversational context, where the common ground is represented by a set of propositions that are commonly accepted by everyone in the context (from this we can derive the *context set* of a context as the set of worlds compatible with the common ground of that context).[22] Thus, suppose that you utter the sentence, "The capital of Delaware is Dover." You are making an assertion that has a particular content, and your sentence can therefore be judged true or false. But to make an assertion is to propose a particular update to the context. Specifically, in making this assertion, you are proposing to update the context in such a way that the common ground will include the proposition that the capital of Delaware is Dover.[23]

One broad approach to understanding conversational disagreement is in terms of a certain kind of conflict at the level of content. (Our notion of exclusionary contents is a way of spelling out this broad approach in more detail.) This may well have appeared to be a promising strategy, but in light of our experimental results, it might be time to start looking elsewhere. Perhaps the notion of conversational disagreement is better understood in terms of a conflict at the level of proposed updates to the context. Consider a conversation in which one speaker proposes to update the context one way, and the other speaker rejects this proposal and proposes updating the context in an incompatible way. Regardless of whether the claims made by these speakers are exclusionary, it may be felt that there is a clear sense in which the two speakers disagree.

This approach provides an importantly different picture of what is going on in ordinary cases of disagreement about straightforwardly descriptive facts.

(6) a. Nate: The capital of Delaware is Dover.
b. Halley: No, the capital of Delaware is not Dover.

In this dialogue, it seems clear that the two speakers disagree. One might think at first that this disagreement is to be understood in terms of a conflict between the contents of what the speakers are saying. We are suggesting a different approach. Broadly speaking,

[22] Cf. Stalnaker 1978. We'll follow Stalnaker (2014) in assuming that $p$ is common ground among a group $G$ iff all members of $G$ accept $p$ and accept that all members of $G$ accept $p$, and so on.

[23] All this talk of update proposals might suggest that we are allied with the tradition of dynamic semantics, in which the content of a sentence is identified with its characteristic update proposal (cf. Heim 1982, Veltman 1996). However, we intend our discussion here to be neutral on the question of dynamic vs. static semantics.

the idea is that the speakers disagree because there is a conflict between the updates they are proposing. Nate is proposing to make it common ground that the capital of Delaware is Dover, while Halley is proposing to make it common ground that the capital of Delaware is not Dover. In some important sense, these proposals seem to be in conflict, and for that reason, the two speakers disagree.

In this specific case, the contents of their claims are exclusionary, but notice that this approach opens the door, at least in principle, to the possibility that speakers disagree even while making non-exclusionary claims. After all, the picture is that speakers disagree whenever there is a conflict between the proposals they are making. Thus, if two speakers can make conflicting proposals even while making non-exclusionary claims, we would have a case of disagreement without exclusion.

### 3.1 Conflict and incompatibility

We have been introducing a general picture of the nature of conversational disagreement. A key task now is to spell out this general picture in real detail and say more precisely what it means for two proposals to be in conflict. This turns out to be quite a difficult task. For example, consider the following case:

(7) [Context: It is common ground that the keys are either in the drawer or in the desk.]

a. Josh: The keys are not in the drawer.

b. Justin: Wait, I just checked, and the keys are not in the desk.

Here, Josh proposes to make it common ground that the keys are not in the drawer, while Justin proposes to make it common ground that the keys are not in the desk. As long as the conversational participants hold on to the assumption that the keys are either in the drawer or in the desk, there seems to be some important sense in which they cannot update with both of these proposals. Yet the obvious solution would be just to give up on this assumption. Hence, it seems that these two proposals do not conflict, and the speakers therefore do not disagree.

But now consider a slightly different case:

(8) a. Josh: No person has ever directed more than 40 movies.

b. Justin: Wait, I just checked, and Woody Allen directed 47 movies.

On some purely formal level, this second case might seem complete parallel to first. As long as we hold on to the assumption that Woody Allen is a person, we cannot update with both proposals, but we can update with both proposals if we give up that assumption (concluding, e.g., that Woody Allen is actually a highly intelligent robot). Nonetheless, to the extent that we want an account of conflict that captures people's intuitive notion of conversational disagreement, it seems that we should say in this case that the two proposals conflict.

We are not quite sure how to develop a perfectly general account that can handle cases like these and all others. For that reason, we will adopt a somewhat different strategy. We define a very stringent condition according to which certain proposals are actually *incompatible*. Then we suggest that being incompatible in this sense is sufficient (but not necessary) for two proposals to be in conflict.

We will offer a more formal definition in section §4.3, but for the moment, we offer an informal account that aims to remain neutral between formal frameworks. The basic idea is that two proposals are incompatible when no possible context is the way both proposals propose the context to be. Say that a context *supports* a proposal iff that context is the way that proposal aims for the context to be. Then, our first pass at an account of what it is for two updates to be incompatible may be stated as follows:

INCOMPATIBILITY (FIRST PASS)
Two proposals are incompatible iff no context supports both.

This first pass is not quite right, for there is a way for a context to support two intuitively conflicting proposals. For instance, in the traditional Stalnakerian framework, the context is a set of possible worlds, and assertion is understood as a proposal to remove all worlds from the context at which what is asserted is false. Then, there will be a context (set of worlds) that supports the proposal made by asserting $p$ and the proposal made by asserting $\neg p$: namely, the empty set.

Therefore, we refine our above first pass account of incompatible proposals by appealing to a notion of a *trivial context*. In the Stalnakerian framework, the empty set is the trivial context. If we adopt a more complex account of context, the notion of a trivial context will have to become correspondingly more complex. Abstracting away from these details, we can then define the idea of incompatibility as follows:

INCOMPATIBILITY

Two proposals are incompatible iff no non-trivial context supports both.

Note that this is a very strong condition. It says that there is no non-trivial context that supports both proposals. Hence, no matter how one changes the context, there will still be no way to update with both proposals. (This very strong condition does not capture the disagreement in cases like the one described above in which it would be possible, at least in principle, to shift to a non-trivial context that supports both proposals). The claim now is that if two proposals are incompatible in this strong sense, they are in conflict; hence, we claim that if two speakers propose incompatible updates, they disagree.

Consider again the dialogue between Nate and Halley about about whether Dover is the capital of Delaware. In this dialogue, the two proposals are incompatible. Nate's proposal is to make it common ground that the capital of Delaware is Dover, while Halley's proposal is to make it common ground that the capital of Delaware is not Dover. There is simply no non-trivial context that supports both proposals.

### 3.2 Disagreement without exclusion

In this one example, the two sentences have exclusionary content. The key thing to notice about this account, however, is that it opens the door to the possibility of cases in which speakers do not make claims with exclusionary content but nonetheless propose incompatible updates. For example, consider again our case of a disagreement involving implicatures.

(5) a. Cody: John ate some of the cookies.
b. Sally: No, John ate all of the cookies.

On the present account, we do not need to introduce any further principles to explain why these two speakers disagree. Their disagreement is of exactly the same type one finds in the case involving factual disagreement. By uttering the sentence he does, Cody says something with a particular content (determined by the literal meaning of *some*), but he also proposes an update (determined in part by the scalar implicature that John didn't eat all of the cookies). Cody's proposal is incompatible with Sally's, and the two speakers therefore disagree.

In this last case, the difference between the content and the proposed update is to be explained in terms of an implicature, but it is not the case that implicatures explain all such differences. Consider an example involving epistemic modals.

> John and his friend are looking for the keys. John hasn't yet checked the drawer, so she says to his friend, "The keys might be in the drawer." Just then, Sabrina overhears John's claim. Sabrina has the keys in her pocket, and knows this, so she jumps in to say, "No, I have the keys here with me."

In this example, there is some intuitive pull to think that John's claim is true—after all, John seemed perfectly justified in making it, given what he knew. Yet, Sabrina's claim is also clearly true. Nonetheless, John and Sabrina seem to disagree by making their claims—after all, Sabrina's rejection of John's claim is appropriate. Making the exclusion inference, we might infer that John's claim here must be false.[24] However, the present account allows for another possibility, which respects the intuition that John's claim is true as well as the intuition that John and Sabrina thereby disagree (cf. Montminy 2012, Huvenes 2015, Khoo 2015). According to Khoo (2015), what might be going on in cases like these is that in making his claim, John proposes (perhaps among other things) that it be compatible with what's common ground between him and his friend that the keys are in the drawer; Sabrina rejects John's proposal and proposes instead that it be common ground between John, his friend, and Sabrina that the keys are *not* in the drawer but are rather with Sabrina. Thus, the update proposals associated with John's and Sabrina's claims are incompatible.[25]

Let us return now to the sort of moral conflict cases we explored in the experiments.

(9) A: Dylan didn't do anything morally wrong
B: No, Dylan did do something morally wrong.

We will be developing a view according to which in these cases the speakers do not make

[24] Relativists and speech act pluralists about epistemic modals have appealed to such examples to bolster their views, since they are designed to predict that (i) John was completely warranted in making his claim, (ii) Sabrina's claim is also completely warranted, and (iii) that at least one of John or his friend's claims must be false (cf. MacFarlane 2011, 2014, von Fintel & Gillies 2011).

[25] Cases like the following complicate this story:

(i) A: It might be that p, and it might be that q. There are no other possibilities.
B: Okay, I just checked and not p, so it must be that q.

Here, it seems clear to us that B does not disagree with A, even though prima facie they propose incompatible updates. Intuitively, the reason is that A cares about ruling out the $\neg p \wedge \neg q$-worlds, and B proposes a refinement of A's proposal. We are not sure what the right solution to this puzzle is, although we note two options. One is to refine the notion of incompatible updates, and the other is to hold that A and B's proposals are actually compatible, even though they may appear not to be. The second strategy would involve further complicating how we get from the sentence uttered to the proposal one makes for the context, and we leave for other work sorting out how to do this plausibly and systematically for examples like these.

exclusionary claims but nonetheless disagree. In particular, on the view we will develop, the speakers make non-exclusionary claims but actually propose incompatible updates.[26]

# 4 Moral semantics after the exclusion inference

Thus far, we have been arguing that it is at least conceptually possible for two speakers to disagree even if neither of them is saying anything incorrect. We have seen that thinking of two speakers as disagreeing with one another in terms of their proposals to update the context allows for the possibility of non-exclusionary disagreement. However, we haven't yet seen how this might work in the kinds of cases discussed in §2. We thus turn now to applying our general framework of disagreement to the case of moral language in particular. Our aim is to develop a theory of the semantics and pragmatics of moral sentences that explains how there can be moral conflict cases in which the two speakers disagree but in which neither of them says anything incorrect.

To begin, we need to introduce a few notions that will play a key role in our semantic theory. First, there is the notion of *norms*, which we'll assume are the sorts of things relative to which actions may be assessed. Take for example Dylan's act of stabbing a passerby. This action may be forbidden by some norms and not forbidden (perhaps even demanded) by others.[27]

The next important idea is that when a speaker makes a moral claim, he or she is some way affirming certain moral norms and opposing others. Thus, if a speaker says, "What Dylan did was morally wrong," the speaker thereby affirms moral norms that forbid Dylan's act of stabbing a passerby and opposes moral norms that do not forbid Dylan's act. This

[26]Note that this framework is compatible with various explanations of our experimental findings. For instance, it is compatible with Björnsson & Finlay (2010)'s proposal that such disagreements are grounded in the attitudes of the speakers towards a common propositional content, and with Plunkett & Sundell (2013)'s proposal that such disagreements are about the meanings of the expressions involved in the uttered sentence, and also with expressivist (and even possibly relativist) explanations. Below, we offer a more precise characterization of what is happening in these cases, but we ultimately want to remain open to alternative implementations of the basic strategy of distinguishing incompatibility in what is claimed (or said) from incompatibility in the proposal to update the context.

[27]A bit more formally, a set of norms is a set of propositions $P$. Let $\Phi$ be a variable over English expressions that denote particular actions (e.g., "What Dylan did", or "Dylan's stabbing of the passerby"), and $\phi$ be a variable over actions—in what follows, we'll assume that $\phi$ is the action denoted by $\Phi$. Let $p$ be the proposition which states that some particular action $\phi$ occurs. We can then say that $P$ forbids $\phi$ iff $P$ entails $\neg p$; and that $P$ requires $\phi$ iff $P$ entails $p$. These definitions assume that $P$ is consistent. If not, we'll need more complex definitions. None of these complications will matter for our purposes. See Lewis (1981), Kratzer (1981, 1991) for discussion.

affirming and opposing of moral norms in speech is very important. When speakers affirm certain moral norms, they seem to be putting those norms forward as guides for living: thus, if some action $\phi$ is forbidden by those norms, the speaker is communicating perhaps (among other things) that we should not perform it, that we should feel guilt if we do, that we should encourage others to avoid doing it, and so on. Whatever norms an individual endorses will be those by which he or she guides her life.[28]

The question now is how a theory about the semantics and pragmatics of moral sentences should explain this affirming and opposing of norms. More specifically, we will be asking whether it is possible to develop a theory that explains how there can be cases of moral conflict in which people conclude that two speakers disagree but in which people also conclude that neither speaker has said anything incorrect.

### 4.1 Factualist invariantism

One very standard theory of the semantics and pragmatics of moral sentences is the view we will call *factualist invariantism*. According to this view, the contents of moral sentences are propositions describing the moral norms that govern the world one is in (factualism), and that every unembedded moral sentence (containing no other context-sensitive vocabulary) expresses the same proposition in every context of that world (invariantism).[29] According to the factualist invariantist, when a speaker makes a moral claim, she proposes affirming those moral norms which she thinks actually govern. So, according to this view, when someone says "What Dylan did was morally wrong" he or she proposes eliminating those worlds from the context set at which the governing norms do not forbid Dylan's act.

Factualist invariantism as we understand it is specifically a thesis about the semantics of moral sentences. In this sense, it should be distinguished from certain metaphysical views in the vicinity. For example, consider the metaphysical view according to which there are objective moral facts (which might be understood as facts about the unique set of moral norms which determine the truth of all moral claims). This metaphysical view should not be confused with the purely semantic thesis under discussion here. As we will argue in

[28] We take no stand on whether endorsing some norms is a cognitive or conative state or some mixture of both.

[29] If moral necessitarianism and factualism are both true, then every metaphysically possible world is governed by the same moral norms, which are determined by that world. However, this still allows for some epistemically possible worlds to be governed by different moral norms—this will be how the realist models moral ignorance (of course any worlds governed by moral norms different from the actual world will be metaphysically impossible).

more detail below, it is perfectly possible to hold on to the metaphysical view that there are objective moral facts while abandoning the semantic thesis of factualist invariantism (cf. Silk 2013).

Given that factualist invariantism is a semantic thesis, it can be evaluated in part by looking to the intuitions of competent speakers. Prior to seeing the experimental results, one might have thought that the evidence from speakers' intuitions provided strong support for this thesis. In particular, one might have thought that factualist invariantism provides a straightforward explanation of people's intuitions about disagreement in moral conflict cases, whereas non-invariantist views would have to engage in fancy footwork of one sort or another to explain those intuitions. The experimental results from §2 turn this dialectic on its head. We find that people do attribute disagreement in cases of moral conflict but that they do not always conclude that at least one of the speakers must be saying something incorrect. In other words, it is actually the factualist invariantist view that has trouble predicting people's ordinary intuitions.

We now argue that these results provide evidence against factualist invariantism. Factualist invariantism says that, as a matter of the meaning of "wrong," whenever two speakers assertively utter ⌜$\Phi$ is wrong⌝ and ⌜$\Phi$ is not wrong⌝ (respectively) they make exclusionary claims. Thus, if factualist invariantism is true, it follows trivially from certain facts about our language that when speakers assertively utter sentences of this form, at least one of their claims must be incorrect. Now, we have good reason to think that our experimental participants know English and that they would correctly make judgments that follow trivially from facts about our language. Thus, factualist invariantism predicts that the participants should conclude in all of our vignettes that at least one of the speakers must be incorrect. This is certainly a plausible prediction, but as we have seen, it is not what is actually observed. The pattern of data therefore provides evidence against factualist invariantism.

To clarify the status of this argument, let us make three quick additional remarks.

First, note that the argument does not rely in any way on the assumption that the responses given by our participants are *true*. Our participants judged in certain cases of moral conflict that it was possible for neither speaker to be saying anything incorrect. Many philosophers would say that this judgment is false. For example, many philosophers hold views about the metaphysics of morality that are clearly incompatible with our participants' responses (see, e.g., Smith 1994, Shafer-Landau 2003). Suppose we now assume for the sake of argument that these philosophers are completely right and that the responses given

by our experimental participants are indeed false. Even then, the argument goes through. The key point is simply that our participants know how to speak English, and we therefore have some reason to reject any semantic theory according to which it follows from linguistic facts alone that the claims they are making are false.

Second, note that the target of our argument is quite narrow. Our claim is that these results provide evidence against the semantic thesis of factualist invariantism. We are not claiming that these results also provide evidence against all of the many meta-ethical theories according to which our experimental participants' responses are false. To give just one example, Smith (1994) provides arguments for such an account, but his arguments do not rely solely on claims about moral language. Instead, they rely on the metaphysical claim that if moral thinkers were idealized in a particular way (to be perfectly rational, fully informed, etc.), they would all converge on the same values. Notice that this account is deeply different from the semantic thesis of factualist invariantism. We have good reason to assume that our experimental participants correctly know how to speak English, but there is no particular reason to assume that our participants have correct metaphysical beliefs. Thus, it is no constraint on a theory of the metaphysics of morality that it predict the responses of our participants. Philosophers endorsing such theories may simply say that our experimental participants are giving false responses because they hold false metaphysical beliefs.

Third, note that we do not mean to suggest that the argument is completely decisive. Clearly, there are various ways a factualist invariantist might respond. She could say that the English word 'incorrect' fails to capture the notion of falsity that is most relevant to semantics, or that our experimental participants' tacit understanding of semantics is being overridden by their explicit metaethical beliefs, or that they actually fail to correctly grasp certain aspects of the semantics of moral sentences. Many other responses are surely possible here. The point is simply that these results provide some evidence against the factualist invariantist thesis.

In sum, factualist invariantism seems at first to be a plausible and promising account of the semantics of moral sentences, but the experimental results provide some evidence against it. The question now is how to develop an alternative view that makes sense of the most salient features of moral discourse without running afoul of these results.

Before moving onward, we should note that the objection we are raising is not directed only at factualist invariantism but rather at any semantic theory that makes the same predictions in the cases presented in our experiments. As one example, consider 'expressivist'

accounts according to which moral statements serve to express non-cognitive attitudes (Blackburn 1984, 1993, Gibbard 1990, 2003, Yalcin 2012). These expressivist account differ from factualist invariantism in numerous respects, but researchers have sometimes suggested that they be supplemented with a minimal or deflationary theory of truth that make it possible to derive exactly the same predictions as factualist invariantism in the cases presented in our experiments (e.g., Blackburn 1984, 1993). In other words, when expressivism is combined with certain further assumptions, the result is a semantic theory that runs into exactly the problem we diagnosed for factualist invariantism. Thus, our results provide just as much reason to reject this theory as they do to reject factualist invariantism.

What we need, then, is an account that differs not only in its theoretical foundations but also in the predictions it makes regarding these cases. Specifically, we need an account that can make sense of the idea that there might be moral disagreements in which neither speaker is saying anything incorrect.

### 4.2 Contextualism

One might try to accomplish this by drawing on any of a number of different semantic frameworks (see for instance, Egan 2012, Yalcin 2011, MacFarlane 2014). In the present paper, however, we will focus on predicting non-exclusionary disagreements within a *contextualist* semantic/pragmatic framework (see for example Lewis 1989, Dreier 1990, Brogaard 2008, Finlay 2004, 2009, 2014, Björnsson & Finlay 2010). This framework combines insights from the formal semantics literature (Kratzer 1977, 1981, 1991) with a longstanding tradition within meta-ethics (Harman 1975, Dreier 1990, Finlay 2009), and recent research has explored the ways in which it might help to resolve a variety of important puzzles about moral language (Björnsson & Finlay 2010, Plunkett & Sundell 2013, Finlay 2014, Pittard & Worsnip 2015). By no means do we think that moral contextualism is the *only* or even best way to predict our data—indeed, we encourage authors favoring other semantic frameworks to offer alternative explanations. Still, if we are able within this framework to make sense of disagreement without exclusionary claims, we will have more reason to believe that our broad approach is on the right track.

Broadly speaking, moral contextualism is the view that moral sentences are only true or false relative to a parameter that is fixed by the context in which they are uttered. Within the context of the present framework, this view is most naturally spelled out as

the idea that the context of utterance determines a set of norms. Thus, when a speaker says ⌜Φ is morally wrong⌝, her claim will be true if and only if $\phi$ is forbidden by the moral norms picked out by her conversational context.

Note that contextualism as we understand it is not itself committed to the conclusion that the very same sentence will sometimes turn out to be true in one conversational context but false in another. Rather, contextualism is a purely semantic theory. It allows one to see how this conclusion could be correct, but whether the conclusion actually is correct will inevitably depend on further questions that go beyond anything in semantics narrowly construed. (Recall the example discussed above of the expression "the speed of light here." Whether that expression picks out different speeds when uttered at different locations depends not only on facts about semantics but also on facts about physics.)

Our aim now is to ask whether moral contextualism has the resources to explain the puzzling patterns of intuitions observed in our studies. That is, we want to know whether the semantic framework of contextualism, conjoined with certain plausible hypotheses about people's substantive beliefs, can yield an explanation of people's judgments about exclusion and disagreement.

To begin with, if we are to make sense of the responses given by our experimental participants, we will need to assume that they understand the norm parameter in such a way that it does vary from one context to the next. We will not be developing a complete theory here about how people ordinarily understand the setting of this parameter. However, we do need to introduce two specific assumptions that will play an important role in our explanation.

The first is that ordinary speakers allow that the state of the norm parameter can be changed by the things that speakers say within a conversation. If one speaker makes an assertion and the other accepts it, people may conclude that the setting of the norm parameter has indeed changed.[30] Thus, certain assertions can be regarded as proposals to update the norm parameter.

Second, we assume that the value of the norm parameter in some context is something about which speakers of that context might care deeply. For example, suppose that a speaker makes an assertion that is best understood as a proposal to update the norm

[30] We want to remain neutral for now about what 'conclude' means here. It may mean that they come to believe that the norm parameter in their context has changed. However, it may also just mean that they *accept* that the norm parameter has changed, where to accept something is to endorse it for the purposes of conversation, but not necessarily believe it. See Stalnaker (2002) for discussion.

parameter. On the view we have been developing, this fact about the conversational context will have implications for whether certain sentences come out true or false. However, we assume that the conversational participants will also care deeply about this aspect of the conversational context in a way that goes beyond any concern they might have about purely linguistic matters. In particular, we assume that this update proposal is a way of *affirming* these norms in the sense we introduced above (§4). It is a way of putting these norms forward as guides for living. (For valuable discussion of these issues, see Plunkett & Sundell 2013.)

Given these assumptions, people should believe that there can be cases in which two speakers engaged in a moral conflict (one saying ⌜Φ is wrong⌝ and the other ⌜Φ is not wrong⌝) may be such that neither says anything incorrect. The core idea here is that each of these sentences may be true relative to certain sets of norms but false relative to others. Thus, if the context fixes the norm parameter in the right sorts of ways, it can happen that the two sentences appear to be contradictory but neither of them is false.

It was traditionally assumed that contextualist views that have these consequences face a serious problem, since they do not seem to be able to make sense of disagreement. However, in light of our experimental results and of our hypothesis INCOMPATIBILITY, we have reason to think that contextualism may have the resources to predict disagreement even in such cases. The basic idea is that the disagreement we find in such cases is not a matter of the speakers making exclusionary claims. Rather, it is a matter of the speakers making incompatible proposals to update the norm parameter.

We provide a formal analysis below (§4.3), but before we turn to the details, we explain the core idea here. Suppose that a speaker utters the sentence "What Dylan did was morally wrong." Prior to this utterance, the conversational context included two key elements. Specifically, it included both a representation of the descriptive facts accepted as common ground in the conversation (the context set) and a representation of the moral norms picked out by the conversation (the norm parameter, which in many cases may be indeterminate). When the speaker now utters this sentence, she is proposing to shift the context to one in which these two elements stand in a particular kind of relation. Roughly speaking, she is proposing to shift to a context in which the norms picked out by the norm parameter forbid what Dylan did in all the worlds picked out by the context set. In some cases, this update will involve changing the context set; in other cases, it will involve changing the norm parameter; and in still other cases, it will involve changing both.

With all that in the background, consider again the dialogue from our studies:

(10) A: Dylan didn't do anything morally wrong.
B: No, Dylan did do something morally wrong.

The two speakers in this dialogue appear to disagree, but how are we to understand their disagreement? On the view we have been developing, there are actually a number of different possibilities. In some cases, it may be that they are making exclusionary claims. In other cases, however, it may be that the disagreement here arises in the absence of any exclusionary claims. Our theory predicts that the latter cases will arise whenever the value of the norm parameter is indeterminate in the context. In those cases, we predict that neither of the speakers' claims are false, and yet nonetheless they disagree in virtue of making incompatible proposals for how to resolve this indeterminacy in the value of the norm parameter for their context.[31]

In other words, moral contextualism can give us the resources we need to understand the pattern of people's ordinary intuitions. As we explain in further detail below, it can help us see how one might think that, in cases of this form, the two speakers disagree without either of them saying anything incorrect.

### 4.3 Contextualism and non-exclusionary disagreement

In the previous section, we sketched an outline of how a contextualist theory can predict non-exclusionary moral disagreements. In this section, we turn to the details. Much of these details will involve assumptions which we won't be arguing for. Our aim is a proof of concept, not an argument that this is the best or only way to predict non-exclusionary moral disagreements.[32]

The guiding idea of our theory is that an assertive utterance of $\ulcorner \Phi$ is wrong$\urcorner$ is a proposal to update both the norms of the context and what is common ground in that context. We will state our theory within a two-dimensional semantic framework familiar from the work of Kaplan (1989). Let $[\![\ ]\!]$ be an interpretation function which maps sentences

[31] Similar ideas are proposed by Barker (2002, 2013), DeRose (2004) for resolving the contextual cutoff for "tall" and the contextual standards for what it takes to "know" (respectively). There are some important differences between our theories, however. Since Barker's is stated within a dynamic semantic framework, the theory does not make explicit predictions about the truth value of the various uttered sentences. And since DeRose's theory is stated at a higher level of abstraction, it is compatible with various formal implementations, one of which is the one we endorse below.

[32] In particular, there may be other contextualist theories which generate the same result. In an earlier version of this paper, we worked out a non-indexical contextualist theory (cf. MacFarlane 2009) which does this, but leave out that discussion now for the sake of space.

to truth values (1 or 0) relative to a Kaplanian context, and an index. For now, we'll just assume for simplicity that a Kaplanian context $c$ is a triple of a world, context set, and set of norms, $\langle w, X, N\rangle$; we will assume that an index is just a possible world $w$. A pair of a context and index is a *point of evaluation*; we define the truth of a sentence at a point of evaluation as follows:

(11) TRUTH AT A POINT OF EVALUATION:
$[\![S]\!]^{c,w} = 1$ iff $S$ is true relative to $c, w$.

We can now state a simple version of moral contextualism for "wrong" as follows (where $N_c$ are the norms initialized by context $c$):[33]

(12) CONTEXTUALISM:
$[\![\ulcorner \Phi \text{ is wrong} \urcorner]\!]^{c,w} = 1$ iff $N_c$ forbids $\phi$ at $w$.

Endorsing CONTEXTUALISM raises a new question: what norms are initialized by the context? We do not here want to take a definite stand on what facts determine the norms initialized by some context. Nonetheless, we think a plausible starting hypothesis is that ordinary conversational contexts are often not fully determinate—in particular, that they often do not initialize a unique norm parameter that decides the normative status of every possible action. As such, we will assume that there is often a great deal of indeterminacy in precisely which norms are picked out by the context. We want to emphasize that the kind of proposal we make about moral norms in a context is not limited just to this contextual parameter, but is in fact a general proposal about how to handle indeterminacy in any contextual parameter. We think it is plausible that many contextual parameters are indeterminate in this way (cf. Perry 1997 on "intentional" indexicals; King 2013b,a, 2014 on supplementives; Dowell 2011 on epistemic modal domains); we thus think the proposal below has much to offer beyond the semantics and pragmatics of moral expressions.

We propose to model this and other kinds of contextual indeterminacy by appealing to an innovation explored by von Fintel & Gillies (2011). The proposal is that a *speech situation* $\mathcal{S}$ (in their terminology, a "cloudy context") is best modeled as a set of Kaplanian contexts $c_1, \ldots, c_n$. When it is indeterminate which set of norms are *the* norms of some

[33] We here assume (without explicitly representing as such) that "wrong" has a covert variable over sets of norms in its logical form. This is a simplifying but non-essential assumption. We could instead treat $N$ as a parameter of the index of evaluation that gets initialized by the context of utterance.

speech situation, we say that it contains contexts which initialize different norms.[34] We can now define Kaplan's original notion of sentential truth in a context, as well as the standard notion of a sentence's intension relative to a context as follows:

(13) TRUTH IN A CONTEXT:
$p$ is true in $c$ iff $[\![p]\!]^{c,w_c} = 1$.

(14) INTENSION IN A CONTEXT:
The intension of $p$ in $c$ is $[\![p]\!]^c = \{w : [\![p]\!]^{c,w} = 1\}$.

Finally, we can now define our crucial notion of sentential truth in a speech situation as follows (where $w_c$ is the world of $c$):

(15) TRUTH IN A SPEECH SITUATION:

a. $p$ is true in $\mathcal{S}$ if $\forall c \in \mathcal{S}$: $[\![p]\!]^{c,w_c} = 1$.
b. $p$ is false in $\mathcal{S}$ if $\forall c \in \mathcal{S}$ $[\![p]\!]^{c,w_c} = 0$.
c. $p$ is neither true nor false in $\mathcal{S}$ otherwise.

Since claims are intuitively token utterances of sentences, we identify a claim with a pair of a sentence (the one uttered) and a speech situation (the situation in which it is uttered). Thus, we say that a claim is true or false just if the sentence uttered in making that claim is true or false in the speech situation in which it is uttered. Given CONTEXTUALISM, it follows that $\ulcorner\Phi$ is wrong$\urcorner$ is true in $\mathcal{S}$ if all $c \in \mathcal{S}$ are such that $N_c$ forbids $\phi$ in $w_c$; $\ulcorner\Phi$ is wrong$\urcorner$ is false in $\mathcal{S}$ if all $c \in \mathcal{S}$ are such that $N_c$ does not forbid $\phi$ in $w_c$; and $\ulcorner\Phi$ is wrong$\urcorner$/$\ulcorner\Phi$ is not wrong$\urcorner$ are neither true nor false in $\mathcal{S}$ otherwise.

We turn now to formalizing the update proposal made by uttering a normative sentence. On Stalnaker's theory, to assertively utter a sentence $p$ is to propose updating the context so that it is common ground that $p$ is true in that context; alternatively, it is to propose that the context set entail that $p$ is true in that context. Since we have introduced indeterminacy in speech situations, we must allow for a similar indeterminacy in the context set of a speech situation.[35] Thus, for each context $c \in \mathcal{S}$, let $X_c$ be its context set (as in Stalnaker 1978). We then define what it is for a sentence to be determinately common ground in a context

[34] Similar ideas are pursued in Willer (2013), Cariani (2015).

[35] We also allow for indeterminacy in the world of the speech situation (cf. Akiba 2004, Williams 2008, Barnes & Williams 2011, Barnes & Cameron 2011). We would model this by letting there be two contexts in some speech situation which initialize different worlds. In what follows, we will assume that such metaphysical indeterminacy doesn't arise.

as follows:

(16) COMMON GROUND:

a. $p$ is determinately common ground in $\mathcal{S}$ if $\forall c \in \mathcal{S} : \forall w \in X_c$: $[\![p]\!]^{c,w} = 1$.

b. $p$ is determinately not common ground in $\mathcal{S}$ if $\forall c \in \mathcal{S} : \neg\forall w \in X_c$: $[\![p]\!]^{c,w} = 1$.

c. It is indeterminate whether $p$ is common ground in $\mathcal{S}$ otherwise.

The most natural analog of Stalnaker's proposal about assertion in our indeterministic framework is that assertively uttering a sentence $p$ is to propose updating the context so that it is determinately common ground that $p$ is true in that context. We can model this formally by first defining the Kaplanian context update proposal of a sentence $p$ as follows:

(17) CONTEXT UPDATE:
$|p| = \lambda c \,.\, \langle w_c, N_c, X_c \cap [\![p]\!]^c \rangle$

The context update proposal of a sentence $p$ is $|p|$, and it is a function from Kaplanian contexts to Kaplanian contexts such that the output context is just like the input context except that we intersect its context set with the intension of $p$ at the input context. Then, we define the speech situation update proposal of a sentence as follows:

(18) SPEECH SITUATION UPDATE:
$[p] = \lambda \mathcal{S} \,.\, \{c : X_c \neq \emptyset \wedge \exists c' \in \mathcal{S} : c = c'|p|\}$

Basically, the update proposal of a sentence is a function from speech situations to speech situations that outputs the set of Kaplanian contexts with non-empty context sets that are the result of applying $|p|$ to some member of the input speech situation $\mathcal{S}$. Finally, our analysis of what it is to assertively utter a sentence:[36]

(19) ASSERTIVE UTTERANCE:
To assertively utter $p$ is to propose applying its update proposal, $[p]$, to your speech situation.

If successful, such a proposal will result in updating the speech situation $\mathcal{S}$ so that it consists only of contexts $c$ whose context set $X_c$ entails the intension of $p$ (at $c$). Hence,

[36] This is idealized in the sense that it leaves out the effects of implicatures and other pragmatic effects uttering a sentence will have on the context. We think such effects should be handled by the "total update proposal" made by assertively uttering a sentence in a context, where the total update proposal will often involve more than simply the proposal to apply the update of the uttered sentence to the speech situation.

assertively uttering a sentence $p$ is to propose updating the speech situation so that it is determinately common ground that $p$ is true in that context. For instance, assertively uttering a normative sentence $\ulcorner\Phi$ is wrong$\urcorner$ is to propose updating the speech situation so that it consists only of contexts $c$ whose norm parameter $N_c$ forbids $\phi$ at every world in $X_c$.[37]

Finally, to connect up our formal apparatus here with INCOMPATIBILITY and our discussion of disagreement from §3, we will show how to formally define a notion of *support* and *trivial* speech situation. Recall that the motivating idea from that discussion was that two context update proposals are incompatible iff no non-trivial context c supports both. We define the notion of a speech situation supporting an update proposal as follows:

(20) $\mathcal{S}$ supports $[p]$ iff $\mathcal{S}[p] = \mathcal{S}$.

The basic idea is that a speech situation supports an update proposal just if applying that proposal to the speech situation yields no change. Next, we define a trivial speech situation as the empty set. Then, in our formal framework, INCOMPATIBILITY yields that:

(21) Two proposals, $[p]$ and $[q]$, are incompatible iff there is no speech situation $\mathcal{S} \neq \emptyset$ such that $\mathcal{S}[p] = \mathcal{S}$ and $\mathcal{S}[q] = \mathcal{S}$.

Before we see how our theory handles the moral disagreement cases from before, let us first consider how it handles cases of factual disagreement, as in:

(6) a. Nate: The capital of Delaware is Dover.
b. Halley: No, the capital of Delaware is not Dover.

Our theory predicts that, in uttering his sentence (call it $d$), Nate proposes updating $\mathcal{S}$ with $[d]$, and this would yield the set of contexts $c$ that are the result of applying $|d|$ to some member of $\mathcal{S}$. By the definition of CONTEXT UPDATE, each of these will have a context set $X_c$ which entails $[\![d]\!]^c$. Halley, by contrast, proposes updating $\mathcal{S}$ with $[\neg d]$. Applying her update to $\mathcal{S}$ would yield the set of contexts $c$ that are the result of applying

[37] It is possible to extend this strategy to handle imperatives by adding a "to-do list" parameter for each conversational participant (Portner 2004, 2007), and holding that the update proposal of an imperative sentence is to add a property to certain participants' to-do lists. This is compatible with our semantics following Portner and assigning properties as the intensions (or semantic contents, we are not distinguishing these for simplicity) of imperative sentences. This would then allow us to predict that speakers may disagree by uttering conflicting imperatives ("Let's go to the movies!" "No, let's stay home!") even though neither asserts anything false.

$|\neg d|$ to some member of $\mathcal{S}$, and each of these will have a context set $X_c$ which entails $[\![\neg d]\!]^c$. However, only $\emptyset$ can entail both $[\![d]\!]^c$ and $[\![\neg d]\!]^c$. Thus, since contexts with empty context sets are discarded (by SPEECH SITUATION UPDATE), only $\emptyset$ will support both $[d]$ and $[\neg d]$. Therefore, these updates are incompatible, and hence, we predict that Nate and Halley thereby disagree in virtue of proposing them.

Of course, we know that $[\![d]\!]^c$ and $[\![\neg d]\!]^c$ are contradictory, so it is not surprising that Nate and Halley disagree here. We turn next to see how the theory handles cases of moral disagreement without contradictory claims. To see the theory in action, we will go through two examples.

**Intracontextual non-exclusionary disagreement**

Suppose that A and B are in a conversation in speech situation $\mathcal{S} = \{c_1, c_2\}$. Suppose that $N_{c_1}$ forbids action $Z$ in $w_{c_1}$ and $N_{c_2}$ does not forbid $Z$ in $w_{c_2}$, and that $w_{c_1} = w_{c_2}$. Finally, suppose also that both of these facts are (determinately) common ground. Hence, for each $c \in \mathcal{C}$: every world $w \in X_c$ is such that $N_{c_1}$ forbids $Z$ in $w$ and $N_{c_2}$ does not forbid $Z$ in $w$. Now consider the following exchange:

(22) A: $Z$ is wrong.
B: No, $Z$ isn't wrong.

Given ASSERTIVE UTTERANCE, in making her utterance, A proposes applying [“$Z$ is wrong”] to $\mathcal{S}$. Since at each world $w \in X_{c_2}$: $[\![\text{“}Z \text{ is wrong”}]\!]^{c_2,w} = 0$, by CONTEXT UPDATE, $c_2|\text{“}Z \text{ is wrong”}| = \langle w_{c_2}, N_{c_2}, \emptyset\rangle$. By contrast, since at all worlds $w \in X_{c_1}$: $[\![\text{“}Z \text{ is wrong”}]\!]^{c_1,w} = 1$, by CONTEXT UPDATE, $c_1|\text{“}Z \text{ is wrong”}| = \langle w_{c_1}, N_{c_1}, X_{c_1}\rangle$. Hence, by SPEECH SITUATION UPDATE, $\mathcal{S}[\text{“}Z \text{ is wrong”}] = \{c_1\}$. Therefore, by uttering “$Z$ is wrong,” A is proposing to eliminate $c_2$ from $\mathcal{S}$.

B rejects A's claim by saying “No” and instead proposes applying [“$Z$ is not wrong”] to $\mathcal{S}$. For the same reason as before, $\forall w \in X_{c_2}$: $[\![\text{“}Z \text{ is not wrong”}]\!]^{c_2,w} = 1$, and $\forall w \in X_{c_1}$: $[\![\text{“}Z \text{ is not wrong”}]\!]^{c_1,w} = 0$. Thus, by SPEECH SITUATION UPDATE, $\mathcal{S}[\text{“}Z \text{ is not wrong”}] = \{c_2\}$. Therefore, by uttering “$Z$ is not wrong,” A is proposing to eliminate $c_1$ from $\mathcal{S}$. Hence, we predict that in making their utterances, A and B propose incompatible updates to the context. Thus, we predict that they disagree.

Now, suppose that neither accepts the update proposed by the other. Then $\mathcal{S} = \{c_1, c_2\}$, in which case both sentences uttered by A and B are neither true nor false in $\mathcal{S}$. Hence,

both A and B's claims are neither true nor false, and thus neither is false. Therefore, our theory can predict non-exclusionary moral disagreements. As such, if the Jim/Mamilon and Jim/Pentar examples above are cases in which the two speakers are in a speech situation in which it is indeterminate what norms are initialized, our theory predicts that in those cases their disagreement is non-exclusionary.

Admittedly, though, it is not obvious that in those scenarios the two claims are made in the same speech situation. Take the Jim/Mamilon case, for instance. Perhaps the Mamilon's claim (made by uttering "What Dylan did was not morally wrong") occurs in one speech situation (comprised just of the two Mamilons talking) and Jim's claim (made by uttering "No, what Dylan did was morally wrong") occurs in another speech situation comprised of Jim and the two Mamilons. It is worth exploring what our theory predicts about that scenario, assuming that this is the right way of describing it.

**Intercontextual non-exclusionary disagreement**

Assume in what follows that the Mamilon's claim occurs in $\mathcal{S}_1$ (corresponding to the conversation between just the two Mamilons), and that Jim's claim occurs in $\mathcal{S}_2$ (corresponding to the conversation between Jim and the two Mamilons). On our theory, by making his utterance, the Mamilon proposes applying ["What Dylan did was not wrong"] to $\mathcal{S}_1$. Suppose that the second Mamilon accepts the first's claim; then their speech situation $\mathcal{S}_1$ will comprise only contexts $c$ in which $\forall w \in X_c : N_c$ does not forbid what Dylan did at $w$. Supposing as well that for all $c \in \mathcal{S}_1$: $N_c$ does not forbid what Dylan did at $w_c$ (since we're supposing each such context shares the same world; hence, $\forall c \in \mathcal{S}_1 : w_c = w^*$), we predict that the sentence "What Dylan did was not morally wrong" is true at $\mathcal{S}_1$.

Jim overhears the Mamilon's claim, and steps in to reject it, saying, "No, what Dylan did was morally wrong." Remember that we are assuming that Jim's utterance takes place in a distinct speech situation $\mathcal{S}_2$. On the contextualist theory above, Jim proposes applying ["What Dylan did was wrong"] to $\mathcal{S}_2$. We assume that the world of the speech situation $w^*$ is such that there is some $c \in \mathcal{S}_2$ such that $N_c$ forbids what Dylan did at $w^*$. This is a plausible assumption because (i) without it, Jim's assertion would only be able to succeed trivially (by reducing $\mathcal{S}_2$ to the empty set), and (ii) we are assuming that Jim's assertion is sensible (and hence not one that would only be able to succeed trivially).[38] But then

[38]Technically, what we need to avoid triviality is the assumption that there is some $c \in \mathcal{S}_2$ such that for some $w \in X_c$: $N_c$ forbids Dylan's action at $w$. But given the fact that it's common ground (and also actually true) that Dylan stabbed a random passerby on the street, this weaker assumption entails the

given that assumption, the sentence "What Dylan did was morally wrong" is not false at $\mathcal{S}_2$.

Next, we suppose that the Mamilon's proposal carries over to $\mathcal{S}_2$—this assumes that it was not made only for $\mathcal{S}_1$ but for any nearby speech situation the Mamilon might find himself a part of. Hence, neither Jim nor the Mamilon's claims are false as made in their respective speech situations. However, Jim and the Mamilon still propose incompatible updates to the context $\mathcal{S}_2$, and hence they disagree.

To be clear, the reason this semantic/pragmatic theory can predict these results is because it distinguishes (i) the truth value of the claim made by assertively uttering a sentence from (ii) the proposal to update the context set that is made by that utterance. In cases where two speakers disagree by making non-exclusionary claims, neither of their claims are false (since both are neither true nor false), but they disagree nonetheless because each makes a proposal to update the context that is incompatible with the proposal made by the other. Again, we want to emphasize that our aim here is merely a possibility proof that there is a plausible contextualist theory which can predict the data from §2—we are not defending this theory against its competitors at this time.

## 5 Concluding remarks

The notion of disagreement has played a central role in research in moral semantics, and rightly so. The most influential disagreement arguments in meta-ethics involved making an explanatory generalization that in all moral conflict cases, the two speakers make exclusionary claims. More recent theoretical work has questioned the force of this type of argument by exploring the possibility that there might be cases of moral disagreement even in the absence of exclusionary content (cf. Björnsson & Finlay 2010, Plunkett & Sundell 2013). We have argued for a stronger conclusion here—that in some cases of moral conflicts, people specifically have the intuition that the disagreeing parties are *not* making exclusionary claims. This completely upends the previous dialectic—now it is theories predicting that in every moral conflict case the disagreeing parties make exclusionary claims that are challenged by the disagreement data. If we want to respect ordinary intuitions about disagreement in moral conflicts, a viable semantic theory must allow for the possibility in

stronger one.

some cases of non-exclusionary disagreements.[39]

## References


Akiba, Ken. 2004. Vagueness in the World. *Nous*, **38**(3), 407–429.

Barker, Chris. 2002. The Dynamics of Vagueness. *Linguistics and Philosophy*, **25**, 1–36.

Barker, Chris. 2013. Negotiating Taste. *Inquiry*, **56**, 240–257.

Barnes, Elizabeth, & Cameron, Ross P. 2011. Back to the Open Future. *Philosophical Perspectives*, **25**, 1–26.

Barnes, Elizabeth, & Williams, J. Robert G. 2011. A Theory of Metaphysical Indeterminacy. *In: Oxford Studies in Metaphysics*. Oxford: Oxford University Press.

Beebe, James, & Sackris, David. 2014. *Moral Objectivism Across the Lifespan*. ms.

Björnsson, Gunnar, & Finlay, Stephen. 2010. Metaethical Contextualism Defended. *Ethics*, **121**, 7–36.

Blackburn, Simon. 1984. *Spreading the Word*. Oxford: Oxford University Press.

Blackburn, Simon. 1993. *Essays in Quasi-Realism*. Oxford: Oxford University Press.

Braun, David. 2012. An Invariantist Theory of 'Might' Might be Right. *Linguistics and Philosophy*, **35**, 461–489.

Brogaard, Berit. 2008. Moral Contextualism and Moral Relativism. *Philosophical Quarterly*, **58**, 385–409.

Cappelen, Herman, & Hawthorne, John. 2009. *Relativism and Monadic Truth*. Oxford: Oxford University Press.

Cariani, Fabrizio. 2015. Deontic Modals and Probabilities: One Theory to Rule Them All? *In:* Charlow, Nate, & Chrisman, Matthew (eds), *Deontic Modality*. Oxford: Oxford University Press.



[39]We would like to thank audiences at the Buffalo Annual Experimental Philosophy Conference and the MIT Conceptual Engineering Workshop for constructive feedback on earlier versions of this paper. We'd like to thank Gunnar Björnsson, Alexis Burgess, Fabrizio Cariani, Brendan Dill, Kevin Dorst, Janice Dowell, Jamie Dreier, Andy Egan, Steve Finlay, Sally Haslanger, Matt Mandelkern, Jack Marley-Payne, Sofia Ortiz-Hinojosa, David Plunkett, Mark Richard, Bernhard Salow, Alex Silk, Tim Sundell, Steve Yablo, and two anonymous reviewers for Noûs for many helpful comments and suggestions, which greatly improved the paper.

DeRose, Keith. 2004. Single Scoreboard Semantics. *Philosophical Studies*, **119**, 1–21.

Dowell, Janice. 2011. A Flexible Contextualist Account of Epistemic Modals. *Philosophers' Imprint*, **11**(14), 1–25.

Dreier, Jaime. 1990. Internalsim and Speaker Relativism. *Ethics*, **101**, 6–26.

Dreier, James. 2009. Relativism (and Expressivism) and the Problem of Disagreement. *Philosophical Perspectives*, **23**, 79–110.

Egan, Andy. 2007. Epistemic Modals, Relativism, and Assertion. *Philosophical Studies*, **133**(1), 1–22.

Egan, Andy. 2012. Relativist Dispositional Theories of Value. *The Southern Journal of Philosophy*, **50**(4), 557–582.

Finlay, Stephen. 2004. The Conversational Practicality of Value Judgment. *Journal of Ethics*, **8**, 205–223.

Finlay, Stephen. 2009. Oughts and Ends. *Philosophical Studies*, **143**(3), 315–340.

Finlay, Stephen. 2014. *Confusion of Tongues: A Theory of Normative Language*. Oxford: Oxford University Press.

von Fintel, Kai, & Gillies, Anthony. 2008. CIA Leaks. *Philosophical Review*, **117**(1), 77–98.

von Fintel, Kai, & Gillies, Anthony. 2011. Might Made Right. *Pages 108–130 of:* Egan, Andy, & Weatherson, Brian (eds), *Epistemic Modality*. Oxford: Oxford University Press.

Fisher, Matthew, Knobe, Joshua, Strickland, Brent and Keil, Frank (forthcoming). The Influence of Social Interaction on Epistemic Intuitions. *Cognitive Science*.

Gibbard, Allan. 1990. *Wise Choices, Apt Feelings*. Cambridge: Harvard University Press.

Gibbard, Allan. 2003. *Thinking How to Live*. Cambridge: Harvard University Press.

Goodwin, Geoffrey, & Darley, John. 2008. The Psychology of Meta-Ethics: Exploring Objectivism. *Cognition*, **106**(3), 1339–1366.

Goodwin, Geoffrey, & Darley, John. 2012. Why Are Some Moral Beliefs Perceived to be More Objective than Others? *Journal of Experimental Social Psychology*, **48**, 250–256.

Hare, R. M. 1952. *The Language of Morals*. Oxford: Oxford University Press.

Harman, Gilbert. 1975. Moral Relativism Defended. *The Philosophical Review*, **84**(1), 3–22.

Heim, Irene. 1982. *The Semantics of Definite and Indefinite Noun Phrases*. Ph.D. thesis, University of Massachusetts at Amherst, Amherst.

Horgan, Terence, & Timmons, Mark. 1990. New Wave Moral Realism Meets Moral Twin Earth. *Journal of Philosophical Research*, **16**.

Horgan, Terence, & Timmons, Mark. 1992. Troubles for New Wave Moral Semantics: the 'Open-Question' Argument. *Philosophical Papers*, **21**.

Horn, Laurence. 1985. Metalinguistic Negation and Pragmatic Ambiguity. *Language*, **61**(1), 121–174.

Horn, Laurence. 1989. *A Natural History of Negation*. CSLI Publications.

Huemer, Michael. 2005. *Ethical Intuitionism*. New York: Palgrave Macmillan.

Huvenes, Torfinn Thomesen. 2015. Epistemic Modals and Credal Disagreement. *Philosophical Studies*, **172**(4), 987–1011.

Kaplan, David. 1989. Demonstratives. *Pages 481–563 of:* Almog, Joseph, Perry, John, & Wettstein, Howard (eds), *Themes from Kaplan*. Oxford: Oxford University Press.

Khoo, Justin. 2015. Modal Disagreements. *Inquiry*, **58**(5), 511–534.

King, Jeffrey C. 2013a. The Metasemantics of Contextual Sensitivity. *In:* Burgess, A., & Sherman, B. (eds), *New Essays in Metasemantics*. Oxford University Press.

King, Jeffrey C. 2013b. Supplementives, the Coordination Account, and Conflicting Intentions. *Philosophical Perspectives*, **27**, 288–311.

King, Jeffrey C. 2014. Speaker Intentions in Context. *Nous*, **48**(2), 219–237.

Kölbel, Max. 2004. Faultless Disagreement. *Proceedings of the Aristotelian Society*, **104**, 53–73.

Kratzer, Angelika. 1977. What 'Must' and 'Can' Must and Can Mean. *Linguistics and Philosophy*, **1**, 337–355.

Kratzer, Angelika. 1981. The Notional Category of Modality. *Pages 38–74 of:* Eikmeyer, H. J., & Rieser, H. (eds), *Words, Worlds, and Contexts. New Approaches in Words Semantics*. Berlin: de Gruyter.

Kratzer, Angelika. 1991. Modality. *Chap. 23, pages 639–650 of:* von Stechow, Arnim, & Wunderlich, Dieter (eds), *Handbuch Semantik*. Berlin and New York: de Gruyter.

Lasersohn, Peter. 2005. Context Dependence, Disagreement, and Predicates of Personal Taste. *Linguistics and Philosophy*, **28**(6), 643–686.

Lewis, David. 1979. Scorekeeping in a Language Game. *Journal of Philosophical Logic*, **8**, 339–59.

Lewis, David. 1981. Ordering Semantics and Premise Semantics for Counterfactuals. *Journal of Philosophical Logic*, **10**, 217–234.

Lewis, David. 1989. Dispositional Theories of Value. *Proceedings of the Aristotelian Society, Supplementary Volumes*, **63**, 113–137.

Lyons, David. 1976. Ethical Relativism and the Problem of Incoherence. *Ethics*, **86**(2), 107–121.

MacFarlane, John. 2005. The Assessment Sensitivity of Knowledge Attributions. *Pages 197–233 of: Oxford Studies in Epistemology*, vol. 1. Oxford: Oxford University Press.

MacFarlane, John. 2007. Relativism and Disagreement. *Philosophical Studies*, **132**, 17–31.

MacFarlane, John. 2009. Nonindexical Contextualism. *Synthese*, **166**(2), 231–250.

MacFarlane, John. 2011. Epistemic Modals are Assessment-Sensitive. *Pages 144–178 of:* Egan, Andy, & Weatherson, Brian (eds), *Epistemic Modality*. Oxford: Oxford University Press.

MacFarlane, John. 2014. *Assessment Sensitivity: Relative Truth and its Applications*. Oxford: Oxford University Press.

MacFarlane, John, & Kolodny, Niko. 2010. Ifs and Oughts. *Journal of Philosophy*, **107**, 115–143.

MacFarlane, John, & Kolodny, Niko. 2014. *Ought: Between Objective and Subjective*. ms.

Montminy, Martin. 2012. Epistemic Modals and Indirect Weak Suggestives. *Dialectica*, **66**(4), 583–606.

Moore, G.E. 1922. *Philosophical Studies*. New York: Harcourt, Brace and Co. Inc.

Nichols, Shaun. 2004. After Objectivity: An Empirical Study of Moral Judgment. *Philosophical Psychology*, **17**(1), 3–26.

Perry, John. 1997. Indexicals and Demonstratives. *Pages 586–612 of:* Hale, Bob, & Wright, Crispin (eds), *A Companion to Philosophy of Language*. Oxford: Blackwell.

Pittard, John, & Worsnip, Alex. 2015. *Metanormative Contextualism and Normative Uncertainty*. ms.

Plunkett, David, & Sundell, Tim. 2013. Disagreement and the Semantics of Normative and Evaluative Terms. *Philosophers' Imprint*, **13**(23), 1–37.

Portner, Paul. 2004. The Semantics of Imperatives Within a Theory of Clause Types. *In:* Watanabe, Kazuha, & Young, Robert B. (eds), *Proceedings of SALT 14*. Ithaca: CLC Publications.

Portner, Paul. 2007. Imperatives and Modals. *Natural Language Semantics*, **15**, 351–383.

Richard, Mark. 2004. Contextualism and Relativism. *Philosophical Studies*, **119**, 215–242.

Sarkissian, Hagop, Park, John, Tien, David, Wright, Jennifer Cole, & Knobe, Joshua. 2011. Folk Moral Relativism. *Mind & Language*, **26**(4), 482–505.

Shafer-Landau, Russ. 2003. *Moral Realism: a Defense*. Oxford: Oxford University Press.

Silk, Alex. 2013. Truth-Conditions and the Meanings of Ethical Terms. *In:* Shafer-Landau, Russ (ed), *Oxford Studies in Metaethics*, vol. 8. Oxford: Oxford University Press.

Smith, Michael. 1989. Dispositional Theories of Value. *Proceedings of the Aristotelian Society, Supplementary Volumes*, **63**, 89–111.

Smith, Michael. 1994. *The Moral Problem*. Blackwell.

Stalnaker, Robert. 1978. Assertion. *Pages 315–332 of:* Cole, P. (ed), *Syntax and Semantics 9: Pragmatics*. New York: Academic Press.

Stalnaker, Robert. 1999. *Context and Content*. Oxford: Oxford University Press.

Stalnaker, Robert. 2002. Common Ground. *Linguistics and Philosophy*, **25**, 701–721.

Stalnaker, Robert. 2014. *Context*. Oxford: Oxford University Press.

Stephenson, Tamina. 2007. Judge Dependence, Epistemic Modals, and Predicates of Personal Taste. *Linguistics and Philosophy*, **30**(4), 487–525.

Stevenson, Charles Leslie. 1937. The Emotive Meaning of Ethical Terms. *Mind*, **46**(181), 14–31.

Streiffer, Robert. 2003. *Moral Relativism and Reasons for Action*. Routledge.

Veltman, Frank. 1996. Defaults in Update Semantics. *Journal of Philosophical Logic*, **25**(3), 221–261.

Willer, Malte. 2013. New Dynamics for Epistemic Modality. *Philosophical Review*, **122**(1), 45–92.

Williams, Bernard. 1986. *Ethics and the Limits of Philosophy*. Cambridge: Harvard University Press.

Williams, J. Robert G. 2008. Multiple Actualities and Ontically Vague Identity. *Philosophical Quarterly*, **58**, 134–54.

Wright, Crispin. 2001. On Being in a Quandry. *Mind*, **110**, 45–98.

Yalcin, Seth. 2011. Nonfactualism About Epistemic Modality. *Pages 295–332 of:* Egan, Andy, & Weatherson, Brian (eds), *Epistemic Modals*. Oxford: Oxford University Press.

Yalcin, Seth. 2012. Bayesian Expressivism. *Pages 123–160 of: Proceedings of the Aristotelian Society*, vol. 133.