Meta-Semantic Moral Encroachment: Some Experimental Evidence[1]

Alex Davies[a], Lauris Kaplinski[b], Maarja Lepamets[b],

[a] Department of Philosophy, University of Tartu
[b] Institute of Molecular and Cell Biology, University of Tartu
[d] Estonian Genome Center, University of Tartu

This paper presents experimental evidence in support of the existence of metalinguistic moral encroachment: the influence of the moral consequences of using a word with a given content upon the content of that word. The evidence collected implies that the effect of moral factors upon content is weak. The implications of this for Esa Díaz-León's recent attempt to show how Jennifer Saul can legitimately reject an empirical semantic hypothesis on political grounds are described. Directions for future research are also described.

Keywords: context-sensitivity, experimental philosophy, moral encroachment

1. The meta-semantics of context-sensitivity

On the one hand, there is compositional truth-conditional semantics: an attempt to identify the contribution made by a word to the truth-condition of the sentences of which it (the word) is a part. On the other hand, there is meta-semantics. Whereas compositional truth-conditional semantics is the study of those linguistic meanings that words have, meta-semantics, as we will use the term, is the study of the factors that determine the meanings that words have. For instance, suppose the meaning of a name is an object (its referent). Then the factors that determine the meaning of a particular name are those factors that determine which object the name denotes. These fac-

Corresponding author's address: Alex Davies, Department of Philosophy, Institute of Philosophy and Semiotics, University of Tartu, Ülikooli 18, 50090 Tartu, Estonia. Email: alex.stewart.davies@gmail.com.

[1] The pre-registration of the two experiments, the data collected and the R-code used to analyse the data are all available here: https://osf.io/z4rg2/?view_only=2a771bcbd27445f9824e90317fc9f403

Studia Philosophica Estonica (2019) **12**, 7–33
Published online: December 2019
Online ISSN: 1736–5899
www.spe.ut.ee
http://dx.doi.org/10.12697/spe.2019.12.1.02

tors could be the psychological states of the name's user, or a causal-historical relation to the environment of the word's use, or something else.

Some expressions have context-invariant meanings but context-sensitive contents. For instance, the word "tall" has a meaning that allows its content to vary with the context in which the word is used. For this reason, it is possible for the extension of the word to change even though the world described using the word does not change. In one context, a boy of 1.5 metres may fall within the extension of "tall", whereas in another context, he does not. One area of meta-semantic inquiry is the factors that determine the contents of context-sensitive expressions in context: which features of the context shape the contents of context-sensitive expressions, and in what way do they do this shaping? Examples of theories in this area include: the view that speaker intentions fix the content of context-sensitive expressions in context (e.g. Predelli 1998, Predelli 2011, Åkerman 2015); the view that agreement between relevant language users determines the content of context-sensitive expressions in context (e.g. King 2014, Michaelson 2014); and the view that the broader purposes of the user of an expression determine the content of the context-sensitive expression in context (e.g. Davies 2018, Dobler 2019, Lewis forthcoming and Schoubye and Stokke 2016).

This paper addresses itself to a recent proposal in this area of meta-semantic inquiry (the area that pertains to how the content of context-sensitive expressions is fixed by and in context). Díaz-León (2016) has proposed that the moral consequences of using a context-sensitive expression with a given content, in a given context, have a bearing upon the content the expression has in context: she proposes that there exists (what we may reasonably call) a form of *meta-linguistic moral encroachment* (a label that alludes to similar ideas in the epistemic contextualist literature on pragmatic (e.g. DeRose 2009, Fantl and McGrath 2009) and moral encroachment (Fritz 2017). Díaz-León puts forward a proposal which implies that if the moral consequences of using an expression in a particular context with a content c1 are worse than the moral consequences of using an expression in that context with a content c2, then the content of the expression in that context is more likely to be c2 than c1. However, Díaz-León does not herself present empirical evidence in favour of this proposal. The purpose of the current paper is to present empirical evidence that bears upon the truth of Díaz-León's proposal. Two experiments were conducted with the aim of examining the hypothesis that meta-linguistic moral encroachment is a feature of the meta-semantics of context-sensitive expressions. Although an effect was indeed found (the moral valence of a context influences the truth-value judgements of language users), this effect was weak. As we will see in what follows, the effect's weakness suggests that although Díaz-León's proposal

may well be correct, it may well not be correct to the degree required for Díaz-León's dialectical purposes.

The paper will proceed in the following stages. In section 2, we outline Díaz-León's proposal in greater detail, and explain why it implies the existence of meta-linguistic moral encroachment. In section 3, we describe the experimental arrangement common to the two experiments described in this paper. In section 4, we report experiment 1. In section 5, we report experiment 2. In section 6, we describe how the findings of the two experiments can be built upon in future research.

2. The context of Díaz-León's proposal

Díaz-León (2016) makes a proposal about the meta-semantics of the word "woman", given that the word has the contextualist semantics described by Saul (2012). Saul's contextualist theory of the meaning of the word "woman" is presented as a truth-condition for sentences of the form "*X* is a woman":

> *X* is a woman is true in a context *C* iff *X* is human and relevantly similar (according to the standards at work in *C*) to most of those possessing all of the biological markers of female sex. (Saul 2012, 201)

A meaning for the word "woman" is not provided by either Saul or Díaz-León. But it is easy enough to infer one from this truth-condition. Let us suppose that "ABMFS" denotes a function from objects to truth-values, such that it maps x to true iff x possesses all the biological markers of female sex. Let us also suppose that "SIMILAR" denotes a function from contexts c to functions from objects z to functions from objects y, to truth-values such that the final function maps to true iff z is similar to y by the standards of c—which is to say, that "SIMILAR" denotes a character in Kaplan's sense.[2] Then:

$$\text{«}woman\text{»} = \lambda z.\lambda c.|\{x : \text{ABMFS}(x)\} \cap \{y : \text{SIMILAR}(c)(z)(y))\}|$$
$$> |\{x : \text{ABMFS}(x)\} - \{y : \text{SIMILAR}(c)(z)(y)\}|$$

The denotation requires that the cardinality of the intersection of the set of those who possess all the biological markers of female sex and those who are similar to z by the standards of c is greater than the cardinality of the set of those who possess all the biological markers of female sex but who are not similar to z by the standards of c. Thus, because of the presence of SIMILAR in its denotation, "woman" denotes a character also—which is what makes this a contextualist analysis.

Saul adopts this contextualist analysis of the meaning of "woman" (in part) on the ground that it allows her to account for the fact that users of

[2] Here we use "«»" to specify the translation of "«*woman*»" into lambda calculus.

"woman" will in one context consider a trans woman a woman, and in another context, they will not. But Saul voices a concern about the analysis: her own political view is that one should respect trans women's self-identifications (presumably: in all contexts). And one does not do this if one allows that denials that a trans woman is a woman are true. For this reason, Saul wishes to reject her own contextualist analysis of the meaning of "woman." Thus, Saul seems to be led to reject a semantic analysis not on the basis of the standard empirical grounds used in the assessment of a semantic hypothesis (namely, the empirical adequacy of the hypothesis), but rather, on *political grounds*. Although Saul sketches several ways in which this might be a permissible thing to do, nonetheless, at first glance, deciding whether to maintain an empirical hypothesis on political grounds seems very much to be the doing of something impermissible.

Díaz-León aims to show how Saul can permissibly keep her analysis without having to grant that utterances of sentences of the form "*X* is not a woman", where *X* is a trans woman, are true. If Díaz-León is successful, then Saul would not face pressure to reject the contextualist analysis on political grounds. Díaz-León proceeds by modifying (or perhaps, more accurately: precisifying) the meta-semantics of Saul's analysis: namely, by, in effect, providing us with more information about the particular shape of the function SIMILAR. SIMILAR is a function from contexts to contents. Differences in context result in the return of different contents. But how content depends on context is not specified by Saul. Díaz-León makes a proposal about how we should specify this:

> My proposal, then, is that we should understand the relevant standards at issue in a context as those that are relevant for practical purposes (where these are broadly conceived to include theoretical, prudential, moral, political, and even aesthetic values). (Díaz-León 2016, 249)

Díaz-León is here putting forward a purpose-centred meta-semantics for "woman": a meta-semantics in the same vein as that defended by Davies (2018), Dobler (2019), Lewis (forthcoming) and Schoubye and Stokke (2016). However, distinctive of Díaz-León's proposal is the reference to moral and political values. Díaz-León is proposing that the content of "woman" in a context is constrained by which content would best serve *morally and politically* justified or permitted purposes. Hence if, in a given context, we use "woman", and content CON best serves the morally and politically justified or permitted purposes of the context, then CON is the content of "woman" in this context. In effect then, Díaz-León proposes that there is no gap between what content a word actually has and the content a word ought to have. It is common to posit such a gap in the conceptual ethics literature (e.g. Burgess

and Plunkett 2013 and Haslanger 2012). Díaz-León is proposing that what content "woman" has, and what content "woman" morally speaking ought to have, are the very same thing. For if they were the same thing, then Saul's political considerations would be straightforward evidence bearing upon the content of "woman" in context. These political considerations would be no different, for instance, from an appeal to the presence and direction of a hand gesture, when we are attempting to discern the content of "that" as used in a context, or, for instance, from an appeal to the chemical structure of water, when we are attempting to discern the content of "water" (given the truth of natural kind semantic externalism).

However, Díaz-León does not provide evidence in favour of this proposal. She does consider how her proposal can handle a pair of objections Saul considers against her own contextualist analysis. But no positive empirical support is given for Díaz-León's proposal. This is a dialectically awkward fact. For, although there is certainly room in philosophy for proposals *qua* explorations of logical space, whose truth can be explored further in subsequent work, in this particular dialectical context, Saul's concern is that illegitimate factors are influencing her study of semantics and Díaz-León presents herself as providing a way to avoid this influence. But if Díaz-León's proposal itself is put forward for political reasons (because Díaz-León too feels moral pressure not to recognize denials that a trans woman is a woman as true), then Díaz-León is merely moving the bump to another corner of the rug: from semantics to meta-semantics. If Díaz-León provides no evidence in favour of her proposal, the concern that she adheres to a meta-semantic proposal for moral and political reasons is very much in place.

Suppose we were interested in finding evidence that bears upon the correctness of Díaz-León's meta-semantic proposal. What would that evidence look like? In what follows, we are going to assume that if contextual moral factors play a role in fixing the content of "woman" in context, then they play a role in fixing the content of context-sensitive expressions in general i.e. that an ad hoc proposal is not being made about "woman", but rather that a general proposal is being made about context-sensitivity. Given this, and given that meta-semantics is as empirical a matter as semantics, if Díaz-León's proposal is correct then we should expect that speakers' judgements about the truth of a context-sensitive sentence are sensitive to the perceived moral consequences of using words in the sentence with given contents. We should expect speakers to modulate their use and interpretation of the context-sensitive sentence in light of the moral consequences of the sentence having a given interpretation. For instance, if a speaker of English, Fred, thinks that endangering the safety of trans women is morally problematic, and if a consequence of using "woman" so that trans women do

not fall within its extension has the consequence that trans women's safety is endangered, then we would expect, if Díaz-León's proposal is correct, that Fred will use "woman" with a content ensures that trans women are included within the word's extension.

3. Experimental setup common to experiment 1 and experiment 2

We conducted two experiments which were aimed at uncovering confirming evidence for Díaz-León's proposal. The two experiments each had the same structure. This structure followed the use of so-called "context-shifting thought experiments" from the philosophical literature on context-sensitivity but as deployed in an experimental arrangement previously deployed by Hansen and Chemla (2013). *Standard* (i.e. non-experimental) context-shifting thought experiments are presented within philosophical papers. The reader is shown a scenario which describes the use of a sentence *S* in a context. The author of the paper offers the reader their favoured judgement as to the truth-value of *S*. The reader is then presented with a second scenario in which *S* is used. The author then, once again, offers the reader their favoured judgement as to the truth-value of *S*. The author's favoured judgement will switch between the two scenarios: in one scenario the judgement will be that *S* is false, in the other, that it is true. However, within each of the two contexts in which *S* is used, the object to which the sentence is applied will not have changed any of its properties. From this it is concluded that what is required for *S* to be true must have changed between the two contexts in which it is used. Examples of the provision of such scenarios and the reasoning just described can be found summarized in chapter 1 of Cappelen and Lepore (2005). Here is one example.

> Story I: Smith is quite proud of the results of the rigorous diet he has followed. He has lost easily 15 kg. Stepping on the scales one morning, he notes with satisfaction that they register a thick hair or two below 80 kilos. At the office, he proudly announces, "I now weigh 80 kilos." But the tiresome Melvin replies, "What! In that heavy tweed suit? Not very likely." and, pulling a bathroom scale out of his bottom desk drawer and pushing Smith on to it, notes with satisfaction, "Look. 83 and a bit." (For good measure, let us suppose Smith not yet to have taken off his overcoat, so that the scale actually reads 86.) Of course, we would say, what Melvin has demonstrated does not count against what Smith said. (Contrasting) Story II: Smith, dressed in the last way, is about to step into a crowded elevator. "Wait a minute.", someone says, "This elevator is really very delicate. We can only take 80 more kilos." "Coincidentally, that's exactly what I weigh.", replies Smith. In

> he steps, and down they plummet. So it appears that what Smith said this time is false. (Travis 1985, 199–200)

It is true that, in this example, the sentence used is not exactly the same in the two contexts. But correction for this inadequacy does not dampen the effect.

The judgements used in support of postulating context-sensitivity have been disputed (e.g. Berg 2002). Partly because they have been disputed, some philosophers have deployed experimental methods in order to discern how widespread and stable the intuitions of Travis and others are. And they have indeed been found to be robust (more robust, in fact, than the intuition that the verb "know" is context-sensitive (cf. Hansen and Chemla 2013 and Grindrod et al. 2019, though for recent further investigation into this finding see Francis et al. 2019). An advantage of experimental context-shifting experiments over the thought experiments presented in philosophical papers is that one does not have to suggest judgements to the participants of the experiments. Instead of expressing any judgement about scenarios, judgements of experiment participants are simply requested, collected, and analysed.

If participants' judgements are sensitive to the perceived moral consequences of the use of a word with a given content, then if we cause differences in the perceived moral consequences of the use of a word between two conditions, we should thereby be able to cause differences in participants' judgements about the content of that word. There are various ways to cash out the idea that a use of a word has moral consequences. In the experiments reported in this paper, we will cash out the idea as follows. We are going to be interested in sentences whose truth would help justify an action, where this action is in one context, morally bad, and in another context, not. What we are going to be interested in is whether we can see differences in truth-value judgements across the two contexts: can the use of a sentence to justify a morally suspect act lead interpreters of the sentence to rescind from judging the sentence true? Our hypothesis is that the answer to this question is "yes": if a sentence is used to justify a morally suspect act, interpreters will judge the sentence to be less true.[3] Given that that which the sentence is being used to describe is identical across the two contexts, such a difference in truth-value judgement would *ceteris paribus* be evidence that the *content* of the sentence is being understood differently across the different contexts, which in turn would constitute evidence that the moral valence of the context is a factor that influences the content of a word in context.

[3] In the experiments reported in this paper, participants were asked to provide truth-value judgements along a scale that ranges between false and true.

Each of the two experiments described in this paper included the use of three scenario topics and to each scenario topic corresponded two moral valences: bad and OK. In a version of a scenario with a morally bad valence, the target sentence is being used to justify a morally questionable action. In a version of a scenario with a morally OK valence, the target sentence is being used to justify a morally OK action. In each experiment, two of the three scenarios were test conditions, and one was a control condition that was used to check whether participants were indeed attending to the task. Following the paradigm employed by Hansen and Chemla (2013), each experiment was designed to allow for both within- and between- subject comparisons. This was achieved in the following way. Each participant of an experiment saw all scenarios (all topics, and both valences of each topic). This allows for within-subject comparison. But in addition, comparisons were to be made between subjects for the first scenario that each participant saw: i.e. the scenario which a participant saw without seeing any others. These between-subject comparisons allow us to see whether differences of context can influence speakers' truth-value judgements before they become aware of what contextual factors might be affecting their truth-value judgements; and similarly, they allow us to see whether differences of context can influence speakers' truth-value judgements even if they are potentially aware of what contextual factors might be affecting their truth-value judgements.

Prior to both experiments, a power analysis was conducted using Monte Carlo simulations and data collected from a pilot study. The pilot study had a very small sample size (39 for the "woman" scenarios, and 35 for the a "green" scenarios). To achieve power of 0.8 in the "woman" scenario, it was found that a sample size of approximately 270 participants would be needed. Given that a between-subject design was intended as well as a within-subject design, and given that the between-subject design would divide the sample size by 4, it was decided to recruit as many participants as could reasonably be afforded: around 400 for each experiment. This would be more than sufficient for the within-subject design, but it was recognized even before the experiments were conducted, that such a sample size is less than sufficient to detect an effect for the between-subject design. Nonetheless, since the between-subject comparisons would be comparisons using a subset of the data collected for the within-subject comparisons, little is lost in making the between-subject comparisons.

4. Experiment 1: "woman", "yellow", and "strong enough"

4.1 Participants

411 participants were recruited using Amazon Mechanical Turk.[4] To increase the likelihood of recruiting native English speakers, only participants with IP addresses in Canada and the United States were allowed to participate. To reduce the likelihood of inadvertently recruiting bots (following the Mturk bot-scare of 2018 APS 2018), only persons who had completed more than 1000 prior tasks, and only those who had had 97% of those prior tasks approved, were permitted to participate in this experiment.

There was a check on whether participants' native language was English. 5 participants were excluded from data analysis because they reported not having English as their native language. This brought the sample size down to 406—prior to removing participants who failed the control.

4.2 Materials

4.2.1 Control check: "strong enough"

The control check consisted of a pair of scenarios. These both had a morally OK valence and in both the sentence "that stick is not strong enough" was used to make a statement about a stick that would take the weight of thousands of ants, but not of a human being. However, in one scenario the stick is about to be used to carry the weight of thousands of ants, whereas in the other scenario, the stick is about to be used to carry the weight of a human being. In the first scenario, the sentence should be judged false, whereas in the second scenario, it should be judged true. The control scenarios were as follows:

> *Morally OK / True Scenario*
> Tia and Tony have been lost in the jungle for days. Their food supply is running low. They have reached a deep ravine that they must cross if they are to get out of the jungle alive. They begin searching for a way to cross the ravine. Tony finds a stick which is about as thick as a matchbox but the length of a car, and so just about reaches across the ravine. Although the stick would take the weight of many thousands of ants, it would not take the weight of a human being. Tia and Tony are discussing whether to use the stick. Tia laughs and says "*that stick is not strong enough.*" Tony puts down the stick and they continue searching.

[4] For an assessment of the quality of data collected from Mturk see (Sprouse 2011).

> *Morally OK / False Scenario*
> Tia and Tony are building a large-scale landscape for ants. They want it to be possible for the ants to move across a ditch that is filled with water. These ants cannot swim. So it is important to build some kind of bridge. Tony finds a stick which is about as thick as a matchbox but the length of a car, and so just about reaches across the ditch. Although the stick would not take the weight of a human being, it would carry the weight of many thousands of ants. Tia and Tony are discussing whether to use the stick. Tia laughs and says "*that stick is not strong enough.*" Tony puts down the stick and they continue searching.

The target sentence includes the word "strong", a relative gradable adjective. Relative gradable adjectives are known to be context-sensitive (Kennedy 1997). Thus, the target sentence should allow variation in content. If Díaz-León's proposal is correct, we should expect moral valence to have an effect on interpretations of the target sentence.

4.2.2 "woman"

The contrast in moral valence between the two scenarios with the topic "woman" lay in whether a trans woman was about to be coerced into a men's public toilet in a world where she faced a high risk of violence therein (morally bad valence), or instead, into a men's public toilet in a world where she faced no such risk (morally OK valence). We realize that the ethics of the situation are more complex than matters of safety. But this nonetheless seemed like a facet of the matter which may arouse more consensus than others. Consensus is important to maintain if we are to retain a large sample size once we have removed those who saw no difference of moral valence between the scenarios which nominally differed in moral valence. The differently valenced woman topic scenarios were as follows:

> *"woman" — Morally bad*
> A trans woman is a person who was born with a male body but who identifies as, or perceives themselves to be, a woman. Imagine a world in which trans women face a very large risk of significant physical and potentially sexual violence in men's public toilets. Imagine that, in this world, a guard stands at the entrance to some public toilets to ensure that everyone pays. A trans woman approaches the toilets. The trans woman heads towards the women's toilets. The guard intervenes. The guard says: "Wait. *You're not a woman*. You have to use the men's toilets."

> *"woman" — Morally OK*
> A trans woman is a person who was born with a male body but who identifies as, or perceives themselves to be, a woman. Imagine a world in which trans women face absolutely no risk of physical or sexual violence in men's public toilets—they're quite safe there. Imagine that, in this world, a guard stands at the entrance to some public toilets to ensure that everyone pays. A trans woman approaches the toilets. The trans woman heads towards the women's toilets. The guard intervenes. The guard says: "Wait. *You're not a woman*. You have to use the men's toilets."

We use this target sentence because we are assuming Saul's contextualist analysis of the meaning of "woman."

4.2.3 "yellow"

The contrast between the morally bad and morally OK valences for the topic "yellow" lay in whether a CEO's decision not to buy golden delicious apples from a given county (Tudru) as a matter of personal whim would have devastating effects upon the local economy and well being of the county. In the morally bad valenced scenario, the consequences would be terrible. In the morally OK valenced scenario, there would be no negative consequences. Whilst many will think both morally OK or both morally bad, it was expected that a sizeable portion of participants would think there is a moral contrast between the two valences. The differently valenced scenarios were as follows:

> *"yellow" — morally bad*
> Suppose that Tudru County's economy would collapse if Big Supermarket Chain did not buy their golden delicious apples—there would be high unemployment, extensive poverty and a collapse of the community and tax-base. Tudru County's golden delicious apples all have 1 or 2 mild red splotches on their otherwise yellow skins. But the CEO of Big Supermarket Chain personally prefers golden delicious apples with pure yellow skins, even though his clients are completely indifferent to whether the apples have perfectly yellow skins. When examining golden delicious apples from Tudru County, the CEO says, "*This apple is not yellow*. We won't be buying apples from Tudru County."
>
> *"yellow" — morally OK*
> Suppose that Tudru County's economy would be entirely unaffected if Big Supermarket Chain did not buy their golden de-

> licious apples—other companies would buy the apples instead. Tudru County's golden delicious apples all have 1 or 2 mild red splotches on their otherwise yellow skins. But the CEO of Big Supermarket Chain personally prefers golden delicious apples with pure yellow skins, even though his clients are completely indifferent to whether the apples have perfectly yellow skins. When examining golden delicious apples from Tudru County, the CEO says, "*This apple is not yellow*. We won't be buying apples from Tudru County."

"Yellow" is again a gradable colour adjective. Although it is standard to assume that relative gradable adjectives are context-sensitive, it is currently a matter of debate whether colour adjectives are relative or absolute (see Clapp 2012, Hansen 2011, Hansen and Chemla 2017, Kennedy and McNally 2005, McNally 2011). However, there is compelling evidence that, regardless of this dispute, colour adjectives are context-sensitive (see Hansen and Chemla 2013). We expect then that, if Díaz-León's proposal is correct, then colour adjectives have a content that is sensitive to influence from the moral valence of the context.

4.3 Arrangement

The six scenarios (three topics, each with two valences) were presented in an almost random order: "almost" because the controls never appeared first. This was done to increase the sample size for the between-subject comparisons—which would divide participants according to the first scenario they saw. If controls were allowed to appear first (as well as the test scenarios), then we would needlessly reduce the size of the sample for each of the test scenarios. Notice that all target sentences are negative in polarity. This was done to ensure that whatever differences in response that are found between conditions, it cannot be said that a switch of polarity was the cause of that difference.

For each scenario (of each moral valence), two questions were asked. The first question checked the participant's interpretation of the moral valence of the scenario. The second question asked the participant to assess the truth of the target sentence. Both answers were to be given using a sliding scale: in the first case from "Disagree" to "Agree", and in the second case from "False" to "True". For example, in the Woman scenario, the following questions were asked:

> *Moral Valence Check*
> Please use the sliding scale to indicate to what extent you agree with the following statement: "It is ethically wrong for the guard

> to force the trans woman to use the men's public toilets when the trans woman faces a large risk of physical or sexual violence in those toilets."
> [SCALE LABELS: DISAGREE — AGREE]
>
> *Truth-value judgement*
> The scenario above is the same as before. But now look at the underlined utterance and its context. Bearing in mind its context, assess whether the underlined utterance is true (Your answer may be subtle. Please use the sliding scale to specify your answer).
> [SCALE LABELS: FALSE — TRUE]

Scales were used in order to reap ordinal variables. Categorical questions (yes/no questions) would have reaped only nominal variables. This facilitates the use of more informative analysis than would otherwise be possible.

4.4 Results

Participants were excluded on the basis of the control scenarios if they did not judge the true sentence true (placing the marker on the scale higher than halfway), and the false sentence false (placing the marker on the scale lower than halfway). 233 participants passed the control. Everyone else was excluded from subsequent analysis.

For each scenario ("woman" and "yellow"), participants were excluded if they did not interpret each scenario's nominally morally bad version as morally bad, and each scenario's morally OK version as morally OK. This exclusion was required because we are interested in whether the difference in the moral valence of the scenarios, as understood by the participants, can influence participants' interpretations of the truth-conditions of sentences used in those scenarios.

Unexpectedly, this check radically reduced the sample size: for the "woman" scenario, to 31, and for the "yellow" scenario to 93. This is well below the sample size required to detect any effect of the size expected, given our pilot studies. Unsurprisingly all differences in truth-value judgements between versions of scenarios with negative and OK moral valence (for both within- and between-subject comparisons) were not significant. To save space, these results are not reported in any further detail.

4.5 Discussion

The failure of this experimental design to maintain a reasonable sample size after control checks, and after reduction of the sample to those who adopted the expected switch of interpretation of the valence of same-topic scenarios,

led to attempts to discern ways to improve the design so as to sustain the sample size. Two ways to do this were identified. Firstly, the length of the scenarios should be reduced. Given that participants sourced from Mturk will be working for money, and given that time is money, a reduction in the amount of time required to fully understand a scenario (by reducing the scenarios' word count) should improve performance. Secondly, the differences in moral valence intended for both the woman and yellow scenarios were really quite controversial: this meant that a substantial reduction in sample size is highly likely when we attend only to those who exhibit a shift in valence interpretation in accordance with our expectations. The morality of forcing a trans woman to use men's toilets in countries such as the US and Canada (where unisex toilets are not the norm) is likely to be a divisive topic for participants from those countries, where participants are likely to have fixed moral views regardless of differences of safety. Similarly, whether a company CEO is doing anything wrong in guiding his company as he wishes (even if his wishes have strong negative consequences for a particular local economy) is likely to split participants into the more libertarian and the more egalitarian, who again, will not shift their moral assessment of the situation across the nominally different valences. We can sustain a larger sample size by deploying less controversial differences of moral valence, on which there is likely to be a large consensus that there is indeed a difference of moral valence across each version of each topic scenario.

5. Experiment 2: "a lot of cake", "old", and "short"

5.1 Participants

399 participants were recruited using Amazon Mechanical Turk. As with experiment 1, only participants with IP addresses in Canada and the United States were allowed to participate. Only persons who had completed more than 1000 prior tasks, and only those who had had 98% of those prior tasks approved, were permitted to participate in this experiment. Those who had participated in experiment 1 were also excluded.

3 participants were excluded from analysis because they reported not having English as their native language, bringing the sample size down to 396—prior to removing participants who failed the control.

5.2 Materials

Note that the scenarios employed in experiment 2 took half as many words to present as the scenarios in experiment 1.

5.2.1 Control check: "short"

The control check consisted of a pair of scenarios. These both had a morally OK valence and the sentence "Jake is small" was used in both to make a statement about an 8-year-old boy who is typical in size for an 8-year-old boy. However, in one scenario, what is at issue is whether Jake is small for his age, whereas in the other scenario, what is at issue is whether Jake is small enough (amongst a group of much older boys) to fit through a hole in a fence. We expect the sentence to be judged false in the first case and true in the second. The scenarios were as follows:

> *Control: Morally OK and False*
> Jake is 8 years old. He is typical in size for an 8-year-old boy. Sam wrongly thinks Jake is significantly smaller than the average 8-year-old boy. Sam asserts to Jake's father, "Jake is small. You should take him to a doctor."
>
> *Control: Morally OK and true*
> Jake is 8 years old. He is typical in size for an 8-year-old boy. Jake is playing soccer with a group of much older, and so larger, boys. The ball goes over the fence. Jake is the only one small enough to fit through a gap in the fence to retrieve the ball. Martin, an older boy, asserts, "Jake is small. He could get the ball."

"Short" is a relative gradable adjective, and so expected to exhibit a context-sensitivity which is in principle susceptible to influence from the moral valence of context.

5.2.2 "a lot of cake"

The contrast in moral valence between the two scenarios with the topic "a lot of cake" lay in whether the target sentence was being used to continue to cause an already bratty child to be bratty, or instead, to praise a child who is dying of cancer, and who is not at all bratty. In each case, the sentence "that's a lot of cake" is used to make a statement about two sponge cakes. The differently valenced cake topic scenarios were as follows:

> *Morally Bad*
> Because Chad's meek parents praise absolutely everything Chad makes, Chad is fast becoming an insufferable brat. Today Chad made two sponge cakes. Once again, keen to please his son, Chad's father asserts, "Wow Chad! That's a lot of cake." He thereby encourages Chad's bratty behaviour.

> *Morally OK*
> Andy is a young, good hearted boy who is dying of cancer. He's often very sad. One day he makes two sponge cakes. Ben sees the cakes and is impressed. Ben asserts, "Wow Andy! That's a lot of a cake." Ben thereby makes Andy very happy indeed.

"a lot of cake" is not a gradable adjective. But it plausibly is context-sensitive with respect to how much cake constitutes *a lot* of cake in pretty much the same ways that relative gradable adjectives are. Firstly, a given quantity of cake might be small, say, for a wedding, but large for grandma's desert. The threshold of quantity for constituting a lot of cake may thus vary with context. Secondly, the scale used to measure quantity may be different in different contexts. For one example, consider the exchange cited by Pomerantz (1984, 77) in her conversational analytic study of agreement and disagreement. B says, "that's not an awful lot of fruitcake" which is met with silence by the person who is selling the fruitcake (the insinuation being that the fruitcake is being sold at an expensive rate). The speaker then adopts a different scale for measuring the quantity of fruitcake (rather than focusing on physical mass, the speaker turns to how much you need in order to feel full): "Course it is. A little piece goes a long way." Thus we expect "a lot of cake" to exhibit the context-sensitivity required for contextual factors such as moral valence to shape the content of the expression in context.

5.2.3 "old"

The contrast in valence between the two "old" scenarios lay in whether the sentence "the hospital is old" was being used to justify knocking down a much-needed hospital in order to build a new, better equipped hospital, or instead to justify knocking down a much-needed hospital in order to make way for a casino run by gangsters. The hospital spoken of in each case was 40 years old.

> *Morally Bad*
> The much-needed children's hospital was built 40 years ago. The mob wants the children's hospital to be knocked down so they can build their casino. When arguing that the hospital should be knocked down, the mob asserts to the city planning committee, "The children's hospital is old."
>
> *Morally OK*
> The much-needed children's hospital was built 40 years ago. The mayor is going to replace the hospital with one that provides better treatments to more children. When arguing that the cur-

> rent hospital should be knocked down, the mayor asserts to the planning committee, "The children's hospital is old."

"old" is a relative gradable adjective and was for this reason used in the target sentence.

5.3 Arrangement

The six scenarios (three topics, each with two valences) were, as in experiment 1, presented in an almost random order: "almost" because controls never appeared first. Notice that all target sentences are positive in polarity. Again, as in experiment 1, for each scenario (of each moral valence), two questions were asked (though this time in the reverse order). The first question asked the participant to assess the truth of the target sentence. The second question checked the participant's interpretation of the moral valence of the scenario. Both answers were to be given using a sliding scale. For example, in the cake scenario, the following questions were asked:

> *Truth-value judgement*
> Is Ben's assertion "That's a lot of cake" true?
> [SCALE LABELS: FALSE — TRUE]
>
> *Moral Valence Check*
> Do you agree with the following statement? "It is morally problematic for Ben to make Andy happy by praising Andy's cake-making."
> [SCALE LABELS: DISAGREE — AGREE]

5.4 Results

Participants were excluded on the basis of the control scenarios if they did not judge the true sentence true (placing the marker on the scale higher than halfway), and the false sentence false (placing the marker on the scale lower than halfway). 250 participants passed the control. Everyone else was excluded from subsequent analysis.

5.4.1 "cake" scenarios

Participants were excluded if they did not rate the scenario with nominally bad moral valence as morally bad, and the scenario with nominally OK moral valence as morally OK. This further reduced the sample size to 212—substantially larger than the 31 and 93 who exhibited the expected judgements of moral valence in experiment 1.

Table 5 and Figure 5 present the distributions of truth-value judgements for the within-subject comparison for the cake scenarios for the two moral valences:

	Morally Bad	Morally OK
Maximum Value	100	100
Third Quartile	93.5	99.25
Median	75	81
First Quartile	50.75	67.75
Minimum Value	1	1

Table 5.

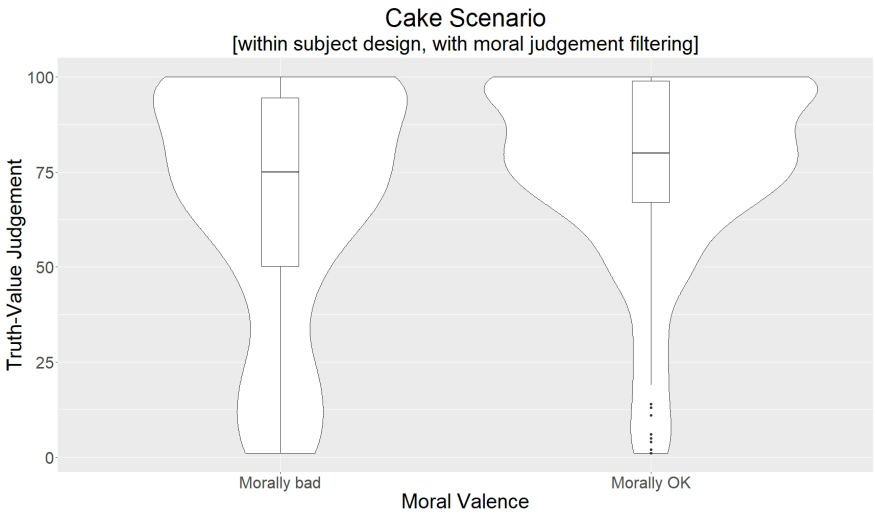

Figure 5.

A one-tailed Wilcoxon signed-rank test did reveal the truth-value judgements given in the morally bad condition to be significantly lower than the truth-value judgements given in the morally OK condition (n = 212, V = 3063.5, α = 0.05/4, p = 0.0000000003681). The effect size was small (r = 0.22). This is reflected in the violin plot (figure 5). In the morally OK condition, responses are more clustered toward the 100 (i.e. "true") end of the spectrum of possible answers than in the morally bad condition, where the density of answers is comparatively thicker towards the 0 (i.e. "false") end of the spectrum of possible answers.

Table 6 and Figure 6 present the distributions of truth-value judgements for between-subject comparison for the two moral valences for the cake scenario:

	Morally Bad	Morally OK
Maximum Value	100	100
Third Quartile	96	89.25
Median	75	78
First Quartile	47	67
Minimum Value	1	4

Table 6.

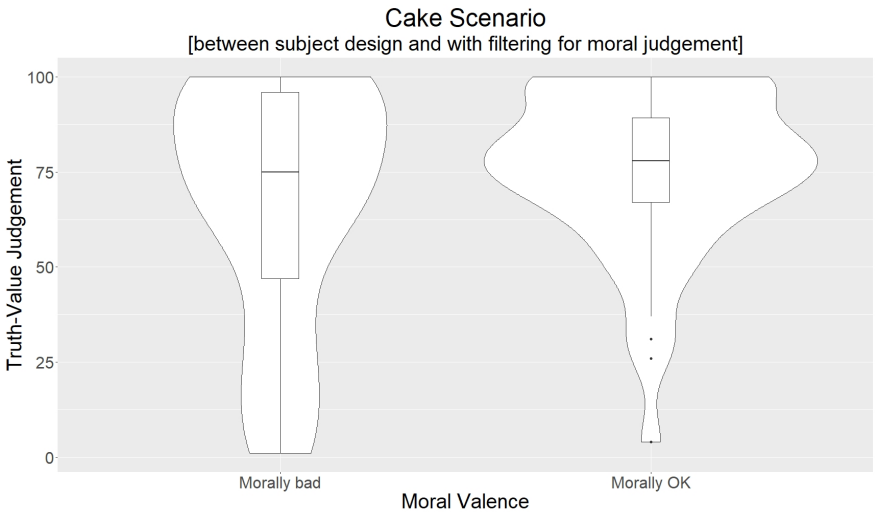

Figure 6.

A one-tailed Wilcoxon rank-sum test did not find the truth-value judgements given in response to the scenario with morally bad valence (prior to seeing any other scenario) to be significantly lower (i.e. closer to 0/"false") than the truth-value judgements given in response to the scenario with morally OK valence (prior to seeing any other scenario) ($n1$ = 48, $n2$ = 58, W = 1227.5, α = 0.05/4, p = 0.1477).

5.4.2 "old" scenarios

Participants were excluded if their interpretations of the moral valence of the old scenarios did not differ across the scenarios with nominally different moral valences. This reduced the sample size to 220—again substantially larger than the 31 and 93 who passed the controls in experiment 1.

Table 7 and Figure 7 present the distributions of truth-value judgements given in the old scenarios for the two moral valences for the within-subject comparisons:

	Morally Bad	Morally OK
Maximum Value	100	100
Third Quartile	93	100
Median	77	85
First Quartile	36	61.75
Minimum Value	1	1

Table 7.

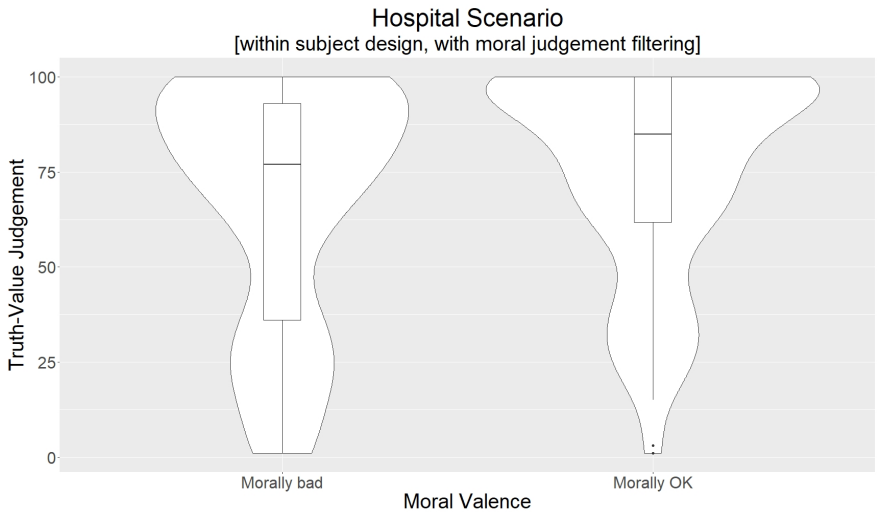

Figure 7.

A one-tailed Wilcoxon signed-rank test did find the truth-value judgements given in the morally bad condition to be significantly lower than the truth-value judgements given in the morally OK condition (n = 220, V = 2936, α = 0.05/4, p = 0.0000000000001707). The effect size was small (r = 0.21). Thus, as in figure 5, we can see again in figure 7, that the negative moral valence of the scenario pushes many of the responses that were packed around 100/"true" (where the density of responses is greatest in the morally OK condition) downwards, thickening the density of responses below 50 (i.e. "false" responses), and thinning the most extreme responses around the 100 mark (i.e. strong "true" responses).

Table 8 and figure 8 present the distributions of truth-value judgements given in the old scenarios for the two moral valences for the between-subject comparisons:

	Morally Bad	Morally OK
Maximum Value	100	100
Third Quartile	94	100
Median	78	88
First Quartile	31.75	65.5
Minimum Value	1	1

Table 8.

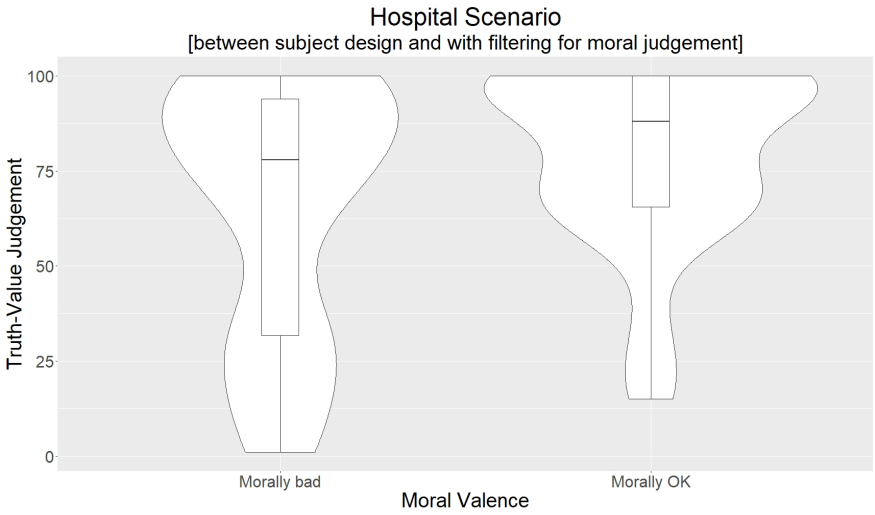

Figure 8.

A one-tailed Wilcoxon rank-sum test did not find the truth-value judgements given in response to the scenario with morally bad valence (prior to seeing any other scenarios) to be significantly lower than the truth-value judgements given in response to the scenario with morally OK valence (prior to seeing any other scenarios) ($n1 = 48$, $n2 = 58$, $W = 1096$, $\alpha = 0.05/4$, $p = 0.0416$).

5.5 Discussion

There appears then to be evidence consistent with Díaz-León's proposal. If a sentence is used to justify a morally suspect act, then people are less inclined to judge the sentence true. This is just what we would expect if the character of a context-sensitive word (e.g. the function SIMILAR in our rendering of Saul's contextualist analysis of the meaning of "woman") makes the content of the relevant word depend in part upon *moral features* of the context, as Díaz-León proposes.

No such effect was found for between-subject comparisons. It would however obviously be a mistake to conclude from this that when participants

were unaware that the differently valenced contexts were influencing their truth-value judgements, they were unaffected by the differences of valence (as if only the opportunity to explicitly contrast the differently valenced contexts was responsible for the differential responses). As indicated by the pre-experiment power analysis, the sample sizes for the between-subject comparisons were simply too small for the experiment to be powerful enough to detect an effect comparable to the size of the effect found in the within-subject comparisons.

Nonetheless, the within-subject finding is interesting. Within-subject designs allow participants to make comparisons between different scenarios, and that means, they are more likely to be aware of the differences between the scenarios with which they are presented. Participants are therefore more likely to be more aware of what features of the context (if any) are influencing their truth-value judgements (for helpful discussion of this issue see Hansen 2014). If we had found that effects arise only if participants were unaware of the effect context was having on their judgements, it would seem that participants do not think any such effect is legitimate; in short, it would look as though they considered it a form of bias—to allow their judgements to be influenced by context in the way it was being influenced. But this is not what happened. Even in the within-subject comparisons (wherein, it seems likely that participants will recognize what contextual differences are affecting their judgements), it is nonetheless the case that people have a tendency to disfavour ascription of a content to a sentence in context if the sentence's having that content has morally suspect consequences. The appearance of this effect in the within-subject comparison suggests that this influence is not considered a form of bias. Rather, even when participants are aware that moral considerations are influencing their judgements, they continue to allow such considerations to have that influence.

However, the effect is indeed small. Given that contexts of speech and interpretation are filled with other influential factors, one wonders what would happen to this weak effect when those other factors are modulated. Pursuit of this question would require a different kind of experimental design. Although this does not mean that we do not have evidence consistent with Díaz-León's proposal, one may wonder about its utility to her, given her dialectical position with respect to Saul. For what Díaz-León sought was not just some effect of moral valence upon content, but an effect that would show denials that trans women are women to be false. What we can see from our data is that although bad moral valence can drag truth-value judgements closer to absolute falsity, the shift is not large. Notice that in the violin plots which present statistically significant differences (viz. figures 5 and 7), we do not see a wholesale shift of the largest bulge from the top to the bottom

of the 50-point marker: we do not see a majority judging the target sentences true in the morally OK valenced scenario, and a majority judging the target sentences false in the morally bad valenced scenario. We instead see in the morally bad valanced scenarios a thinning of the bulge that we see in the morally OK valenced scenarios, and a slight thickening of the bulge of responses over the other side of the 50-point divide between true and false. Although some people are wholesale changing their judgements from true to false, a large proportion of respondents maintained the same truth-value judgement across the two moral valences. If moral meta-semantic encroachment is going to save Saul from impermissibly rejecting an empirical semantic hypothesis because of her political and ethical beliefs, we need the encroachment to be stronger than this. For it is only then that we should expect that when speakers are presented with what Saul and Díaz-León take to be the moral factors surrounding gender, speakers will agree that denials that trans women are women are false.

6. Limitations and future work

We close the paper by describing four limitations of the current study, and what these limitations imply about fruitful directions for future research.

One major limitation of this study is the possibility that other factors besides moral valence are systematically different between the two moral valences of each topic. If they are, they could be responsible for any detected effect. This is a problem for any attempt to support a moral encroachment hypothesis (for instance the same problem arises for Fritz 2017). The problem arises because it is difficult to identify and distinguish moral from non-moral contextual factors. A more sophisticated experimental design—one which facilitated multivariate analysis—would be required in order to better understand possible interactions and "screening-off" effects of different contextual factors operating in concert upon the truth-value judgements of language users.

A second limitation is that we did not manage to collect sufficient data to examine the occurrence (if any) of meta-linguistic moral encroachment for the specific word that interested Díaz-León viz. "woman". In general, future work that tests a much larger range of expressions for the influence of moral valence on truth-value judgements, would help to discern how systematic meta-linguistic moral encroachment might be.

A third limitation, which has already been noted, was the failure to generate samples sufficiently large to be in a position to detect meta-linguistic moral encroachment in between-subject comparisons. For this, either a more ingenious experiment design than we have mustered on this occasion is required, or else a much larger sample is required. The data collected for

the between-subject comparisons is suggestive of the possibility that the effect of moral valence on truth-value judgements is stronger when participants have had no opportunity to compare the moral valences of contrasting contexts. If this were indeed the case, we would need to re-examine the idea that participants consider the influence of moral valence on context to be legitimate: for such a difference in effect strength would be evidence against this.

Fourth, we have adopted an assumption which is typical in the study of linguistic context-sensitivity: that shifts of truth-value judgement across the use of the same sentence in two different contexts to describe a single object which is unchanged across the two contexts are evidence of a shift in the content of the sentence. But when dealing with the influence of moral considerations upon truth-value judgements, the possibility that participants are shifting their truth-value judgements without shifting their interpretations of the content of the sentences employed (they are simply being inconsistent in the use of a single content across the two contexts) becomes more salient. We know of two kinds of experiment which have been performed in the past which check for shifts of interpretation without inferring such shifts from shifts in truth-value judgement. Firstly, a range of experiments in social psychology have directly asked participants to select between interpretations of presented linguistic stimuli (cf. Asch 1940, Asch 1948, Hayes et al. 2018, Pool et al. 1998, and Wood et al. 1996). Secondly, Hansen and Chemla (2017) test for shifts in judgements about the entailments of a sentence, rather than judgements about a sentence's actual truth (relative to a described scenario). Díaz-León's proposal implies that a syllogism like "Only women are allowed in this bathroom. *X* is a trans woman. So, *X* is not allowed in this bathroom" should be judged invalid when the moral consequences of forcing trans women to use men's public bathrooms are significantly negative. Use of either or both of these methods would help to set aside the concern that the small effect in truth-value judgement detected in experiment 2 does not reflect a bona fide shift in truth-conditional content.

Acknowledgements

This paper has benefited immensely from advice and comments from Nat Hansen and Peeter Tinits. Thanks are also due to Julia Fieser and Pablo Veyrat, for helping us to uncover some bugs that arose during data collection and storage. This research has been supported by the programme Mobilitas Pluss project MOBTT45 and the Centre of Excellence in Estonian Studies (European Regional Development Fund) and is related to research project IUT20-5 (Estonian Ministry of Education and Research).

Bibliography

Åkerman, J. (2015). The communication desideratum and theories of indexical reference, *Mind and Language* **30**: 474–499.

APS (2018). Researchers investigate problems with MTurk data. Retrieved from Association for Psychological Science. https://www.psychologicalscience.org/publications/observer/obsonline/researchers-investigate-problems-with-mturk-data.html

Asch, S. (1940). Studies in the principles of judgments and attitudes: II. Determination of judgments by group and ego standards, *The Journal of Social Psychology* **12**: 433–465.

Asch, S. (1948). The doctrine of suggestion, prestige and imitation in social psychology, *Psychological Review* **55**: 250–276.

Berg, J. (2002). Is semantics still possible?, *Journal of Pragmatics* **34**: 349–359.

Burgess, A. and Plunkett, D. (2013). Conceptual ethics I, *Philosophy Compass* **8**: 1091–1101.

Cappelen, H. and Lepore, E. (2005). *Insensitive Semantics: A Defence of Semantic Minimalism and Speech Act Pluralism*, Oxford University Press, Oxford.

Clapp, L. (2012). Indexical color predicates: Truth conditional semantics vs. truth conditional pragmatics, *Canadian Journal of Philosophy* **42**: 71–100.

Davies, A. (2018). Communicating by doing something else, *in* T. Dobler and J. Collins (eds), *The Philosophy of Charles Travis: Language, Thought, and Perception*, Oxford University Press, Oxford, pp. 135–154.

DeRose, K. (2009). *The Case for Contextualism: Knowledge, Skepticism and Context*, Vol. 1, Oxford University Press, Oxford.

Díaz-León, E. (2016). Woman as a politically significant term: A solution to the puzzle, *Hypatia* **31**: 245–258.

Dobler, T. (2019). Occasion-sensitive semantics for objective predicates, *Linguistics and Philosophy* **42**: 451–474.

Fantl, J. and McGrath, M. (2009). *Knowledge in an Uncertain World*, Oxford University Press, Oxford.

Francis, K., Beaman, P. and Hansen, N. (2019). Stakes, scales and skepticism, *Ergo* **6**: 427–487.

Fritz, J. (2017). Pragmatic encroachment and moral encroachment, *Pacific Philosophical Quarterly* **98**: 643–661.

Grindrod, J., Andow, J. and Hansen, N. (2019). Third-person knowledge ascriptions: A crucial experiment for contextualism, *Mind and Language* **34**: 158–182.

Hansen, N. (2011). Color adjectives and radial contextualism, *Linguistics and Philosophy* **34**: 201–221.

Hansen, N. (2014). Contrasting cases, *in* J. R. Beebe (ed.), *Advances in Experimental Epistemology*, Bloomsbury, London, pp. 71–95.

Hansen, N. and Chemla, E. (2013). Experimenting on contextualism, *Mind and Language* **28**: 286–321.

Hansen, N. and Chemla, E. (2017). Color adjectives, standards, and thresholds: An experimental investigation, *Linguistics and Philosophy* **40**: 239–278.

Haslanger, S. (2012). *Resisting Reality: Social Construction and Social Critique*, Oxford University Press, Oxford.

Hayes, T., Lee, J. C. and Wood, W. (2018). Ideological group influence: Central role of message meaning, *Social Influence* **13**: 1–17.

Kennedy, C. (1997). *Projecting the Adjective: The Syntax and Semantics of Gradability and Comparison*, Gardland, New York.

Kennedy, C. and McNally, L. (2005). Scale structure, degree modification, and the semantics of gradable predicates, *Language* **81**: 345–381.

King, J. (2014). Speaker intentions in context, *Nôus* **48**: 219–237.

Lewis, K. (forthcoming). The speaker authority problem for context-sensitivity (or: You can't always mean what you want), *Erkentnis* .

McNally, L. (2011). Color terms: A case study in natural language ontology, *Workshop on the syntax and semantics of nounhood and adjectivehood*, Barcelona.

Michaelson, E. (2014). Shifty characters, *Philosophical Studies* **167**: 519–540.

Pomerantz, A. (1984). Agreeing and disagreeing with assessments, *in* J. M. Atkinson, J. Heritage and K. Oatley (eds), *Structures of Social Action*, Cambridge University Press, Cambridge, pp. 57–101.

Pool, G., Wood, W. and Leck, K. (1998). The self-esteem motive in social influence: Agreement with valued majorities and disagreement with derogated minorities, *The Journal of Personality and Social Psychology* **75**: 967–975.

Predelli, S. (1998). I am not here now, *Analysis* **58**: 107–115.

Predelli, S. (2011). I am still not here now, *Analysis* **74**: 289–303.

Saul, J. (2012). Politically significant terms and philosophy of language: Methodological issues, *in* S. Crasnow and A. Superson (eds), *Out from the Shadows: Analytic Feminist Contributions to Traditional Philosophy*, Oxford University Press, Oxford, pp. 195–216.

Schoubye, A. and Stokke, A. (2016). What is said?, *Nôus* **50**: 759–793.

Sprouse, J. (2011). A validation of amazon mechanical turk for the collection of acceptability judgments in linguistic theory, *Behavior Research Methods* **43**: 155–167.

Travis, C. (1985). On what is strictly speaking true, *Canadian Journal of Philosophy* **15**: 187–229.

Wood, W., Pool, G. L. and Purvis, D. (1996). Self-definition, defensive processing, and influence: The normative impact of majority and minority groups, *The Journal of Personality and Social Psychology* **71**: 1181–1193.