University of Tartu
Institute of Philosophy and Semiotics

THEORY-INDEXED MORAL CONTEXTUALISM

Master's Thesis in Philosophy

Piero Suarez

Supervisor: Patrick Shirreff

Tartu, 2021

Acknowledgements

To Nonoy, Ines, Lucho, and all my family.

To my classmates in Tartu: Youssef, Egle, Sophito, Gabriele, Jed, Nada, and everyone I discussed this work with during the seminars.

To Theorema and its great people: Edvard, Jaime, Nurit, Andrés, Gabi, and everyone else I discussed this work with.

To Beatriz, Reno, Deko, Niina, Iina, Sami, Kukka, Laura for the moral support.

To Francesco and Patrick for the amazing supervision.

To Pipo.

THEORY-INDEXED MORAL CONTEXTUALISM

TABLE OF CONTENTs

0. Introduction

Metaethical theories that are trying to account for moral disagreement face important challenges. On the one hand, if the semantic treatment of moral terms assigns a meaning too specifically related to a contextual parameter (like culture, religion, etc.) we might be ruling out the substantiality of moral disagreements, since disagreeing parties can be both correct under their own terms. On the other hand, if our treatment of moral terms ignores their relation to a contextual parameter, we might be unable to explain the nature of the very disagreement, as we ignored how parties ended up believing different things. This M.A thesis explores the theoretical room for a contextualist account of the meaning of moral terms that is able to model the substantiality of moral disagreements in a way both compatible with non-exclusionary disagreements and with standard externalist semantics.

In the following, I would like to account for what I call *Theory-Indexed Moral Contextualism.* That is: a realist, contextualist and externalist account for our moral speech. In Section 1, I present Horgan and Timmons' Moral Twin-Earth argument that purports to argue against any approach with fixed references for moral terms. In Section 2, I present the wide range of solutions philosophers have found to the challenge presented by the Moral Twin-Earth Argument. In Section 3, I elaborate on how a contextualist approach to moral terms might address the Moral Twin-Earth challenge: I will defend such an approach throughout the present work. In Section 4, I use a Hirsch-like argument to show why moral disagreements under my contextualist modeling of moral terms can still be substantive. In Section 5, I motivate a realist and externalist approach as the best ways to treat moral terms. And finally, in Section 6, I will propose a variation of the Moral Twin-Earth Argument to show how we can handle moral relativism.

The Moral Twin-Earth Argument appears in the literature on metaethics with the purpose of rejecting a naturalist approach to the meaning of moral terms. That argument revived the questions set by Moore's Open Question Argument, and it is inspired by the thought experiment of Putnam's Twin-Earth Argument. With that motivation, Horgan and Timmons, concerned about what kind of reference do moral terms have, observed that if we use the Putnamian story of reference fixation we would have problems later in the explanation of the nature of moral disagreements. During the first section of my thesis, I show their argument and the consequences

of it for future metaethical accounts. In the subsequent section (2), I go through an important number of theories facing the Moral Twin-Earth Argument. I show the different strategies that have been adopted in order to explain away the challenges posited. Philosophers like David, Copp, Pekka Vayrynen, Matti Eklund, Ralph Wedgewood will, so to speak, *bite the bullet* while accepting a big part of the consequences of the aforementioned argument. Other authors, like Khoo & Knobe and Timothy Williamson, on the other hand, will question some of the presuppositions that Horgan and Timmons seem to have. The purpose this section is to show why we should adopt the strategy followed by this second group. I do that by giving the reader a fair look at the possible outcomes that these theories might have. I put special attention to the role these authors assign to the *context* in order to determine the content of moral claims. In the following section (3), I draft the details of a contextualist understanding of the meaning of moral terms—the strategy I find most promising—by making use of the lessons learned from the authors of the previous section. In the next sections (4, 5, and 6), I confront a number of problems that could arise from the postulation of my contextualist account. The fourth section deals with the following accusation: under a contextualist account, moral disagreements look much like merely linguistic disagreements. I build an argument inspired by the work of Eli Hirsch to show why moral disagreements are not merely linguistic ones. The goal of the next section (5) is to motivate an externalist and realist understanding of my contextualist account by showing the problems that moral exceptionalism deals with when makes use of antirealist's commitments. The last section (6) is intended as a way to understand how my contextualist account can stand against relativism. I propose a variation of the Moral Twin-Earth Argument as a thought experiment whose theoretical consequences show that even within contextualism, where many theories might legitimate different moral claims, we still have the resources to compare within these theories. These are the same resources we appeal to when comparing theories from other bodies of knowledge.

1. On Moral Twin Earths

One of the canonical ways to think of the relation between a term and its reference is according to what we have learned from Putnam's Twin-Earth Argument (1975). This argument showed that even if two individuals have the same inner-goings in their minds and they have very much alike languages, they could be referring to different properties despite using the same

terms. The classic H2O-XYZ case consists of two agents using the term ‘water’ on different planets. On the planet of one of them, Earth, there’s a transparent, liquid entity composed of H2O that earthlings refer to with the term ‘water’; on the planet of the other agent, Twin-Earth, there’s a superficially identical entity that isn’t H2O but XYZ that twin-earthlings refer to with the term ‘water’. Even though these two agents have the same mental experiences, their uses of the term ‘water’ refer to different properties. Putnam concludes from this thought experiment that Semantic Externalism is true. That is, meaning is not in the head but outside of it. One particular meta-semantics, i.e. a story about how a specific reference ended up fixed by a linguistic term, is in a good position to explain this difference in reference: references causally regulate our uses of terms. This idea allows us to understand why the two agents of Putnam’s Twin-Earth Argument are using terms with different references fixed. It also allows us to understand why a sentence like ‘Water contains hydrogen’ would be intuitively true in Earth, but false on Twin-Earth.

Horgan and Timmons (1991, 1992a, 1992b) were interested in the references of moral terms, so they proposed the following variation of Putnam’s Twin-Earth thought experiment. Let us imagine that on Earth, earthlings behave, think, and argue morally according to Consequentialist principles. And whenever earthlings use moral terms, like ‘good’, to evaluate their actions, the references of those will be related to Consequentialist properties (e.g., what maximizes expected utility). Let us say that the extension of the predicate ‘good’ used by the earthlings is E_c. On the other hand, on another planet Twin-Earth, twin-earthlings behave, think and argue morally according to Deontological principles. And whenever twin-earthlings use moral terms, like ‘good’, to evaluate their actions, the reference of those will be around Deontological properties (e.g., what honors Kantian categorical imperative). Let’s say that the extension of the predicate ‘good’ used by the twin-earthlings is E_d. Furthermore, let us concede that *lying for saving a life* is considered as a good thing according to Consequentialist principles, but as something wrong by Deontologist principles. For exposition purposes, let’s name ‘Coco’ and ‘Dede’ our representatives of each respective planet. Let’s now imagine that Dede, a twin-earthling, manages to go to Earth and she argues with Coco, an earthling, about the truth of the sentence ‘Lying for saving a life is good’. As Horgan and Timmons (H&T) point out, there is an intuitive sense in which the earthling, Coco, and the twin-earthling, Dede, are having a moral

disagreement.[1] In that sense, they are arguing about whether *lying for saving a life* is morally good or not and not merely about something linguistic. If we are to take Coco-Dede's disagreement as a substantive moral disagreement, and not as a merely linguistic one, H&T argue that we would need to take their uses of '*(*morally*)* good' as referring to the same property. Otherwise, if we take Coco and Dede to be arguing about if *lying for saving a life* is morally *good*$_c$ or morally *wrong according to its coherence with so and so principles*, H&T point out that a strong moral relativism comes into place since we can take both Coco and Dede to hold something true at the same time. That is, it would be simultaneously true that lying for saving a life is good$_c$, and also that lying for saving a life is not good$_d$.

If we follow the metasemantics learned from Putnam's Twin-Earth Argument and we claim that meaning is determined by the properties causally regulating the uses of a term, we will have to admit that some sort of relativism seems to be in place. Insofar as earthlings' and twin-earthlings' moral claims are causally regulated by different properties, their moral terms will refer to different properties. Since *lying for saving a life* is part of the extension E_c but not of the extension E_d, then the earthling's 'good' and the twin-earthling's 'good' are not co-extensional. That drives H&T to argue that one of the following must be false:

(1) Moral properties causally regulate how we use moral terms. The extension of moral terms is determined by the moral principles ruling the user's community.
(2) The meaning of 'good' is the same in Coco-Dede's dispute.

H&T argue that for understanding Coco and Dede's exchange as a moral disagreement, we would have to take (2) to be the one that is true. It's intuitive, they claim, that in this kind of exchange the earthling and twin-earthling are having a moral disagreement while disagreeing about the truth of 'Lying for saving a life is good'. And both uses of 'good' have to refer to the same property. Denying (2), according to H&T, would rule out the possibility of substantive moral disagreements in cases like the one presented and we would be left only with a linguistic disagreement. Since the Putnamian metasemantics[2] seems to be predicting that two different properties are the ones behind the two uses of 'good' by the disagreeing parties, H&T argue that (1) is false, as it is inconsistent with (2). Any story like (1) about how moral terms fix their

[1] The names 'Coco' and 'Dede' weren't mentioned by H&T but by me for exposition purposes.
[2] We could also put it as the *Causal Theory of Reference* (Boyd 1988).

references has to be false, since it wouldn't capture the substantivity of the disagreement. That way, we could build a Moral Twin-Earth disagreement scenario with different properties regulating the use of moral terms. If a moral term has a fixed reference R, then a Moral Twin-Earth disagreement scenario can be built with different fixed reference R_2 for the moral terms of the twin-earthlings to show that the intuition about this being a moral disagreement is not compatible with any fixed reference for our moral terms; at least with our Putnamian story of reference fixation. Nevertheless, there are different attempts to escape from this problem. In the following section, I show the different ways in which philosophers have proposed to make moral disagreements compatible with other compelling stories about their references.

2. On the Semantics of Moral Disagreements

Similar to Moore's Open Question Argument, the thought experiment proposed by H&T was intended as a weapon against moral naturalism. That is, a weapon against the claim that the meaning of moral terms could be *defined, reduced* to, or *expressed in* (natural) non-normative terms. Moore's Argument invites us to propose a non-normative reduction for a moral term like 'good', call it *F*, in such a way that for something to *be good* is just for something to *be F*. Then, if we ask the question 'Is it true that *F* is good?' we will be facing a circular, meaningless, closed question. Since a question like 'Is it true that *F* is good?' is supposed to be open, relevant and meaningful, Moore concludes that moral terms like 'good' cannot be defined as the Argument works for whatever *F* we propose.

Similarly, if a moral term was definable in non-normative terms (like F), we could just imagine a Twin-Earth with a slightly different definition (like F_2) and, again, we wouldn't be able to explain any subsequent substantive disagreement between the user of these two terms. H&T argue that we have no other option than to accept that a moral term like 'good' can only refer to the property of *goodness* simpliciter, independently of the moral principles that rule upon a community. Thus, the Putnamian metasemantics is not a good candidate to explain how this reference is fixed. Once we accept that one irreducibly normative property (if any) is the one being referred to by moral terms both in Earth and in Twin-Earth, we can account for Coco-Dede's exchange as a substantive disagreement. That is, H&T argue that if there is a disagreement over whether lying for saving a life is morally good or not, both disagreeing parties has to be referring to the same property by their respective uses of the term 'good'.

Different authors have accommodated or rejected the Moral Twin-Earth Argument with varying strategies. David Copp (2007), for instance, holds that (1) and (2), from above, are not actually inconsistent, even though he accepts that the properties referred to by moral terms in the disagreement are different. Pekka Vayrynen (2018) also argues that (1) and (2) are compatible but at the same time he holds that the properties referred to in the disagreement are the same. Other authors like Matti Eklund (2017) and Richard Williams (2018) explore how the conceptual role that moral terms have within communities might rule out the possibility of more than one property being referred to by moral terms; taking (1) as false in the way. Khoo and Knobe (2018) and Timothy Williamson (2020), on the other hand, are willing to accept that we could have moral disagreements where none of the disagreeing parties is mistaken, in such a way that you could say that they are actually talking past each other in a sense; rejecting the truth of (2) in this way. In the following, I build up a bit more of the details of these strategies in such a way that we can compare them better.

David Copp disagrees with H&T's exclusive disjunction of either (1) or (2). He argues that we could accept as true both (1) and (2) and, at the same time, make that compatible with the satisfaction of the widespread intuition about the earthling and the twin-earthling disagreeing morally. That is, Copp holds that even if we grant that different properties are the ones regulating the use of moral terms in Coco-Dede's disagreement, we can still claim that both moral terms *mean* the same. Copp recognizes that both communities (earthlings and twin-earthlings) use those terms to guide their behavior. Earthlings and twin-earthlings will disagree in many practical implications of their respective uses of 'lying for saving a life is (isn't) good*'*. For example, earthlings will advise their children to lie when the life of a person is at risk and twin-earthling will advise their people to avoid or resist lying even when the life of a person is at risk. However, both uses of 'good' will still convey appraisal of actions. Copp also argues that it's plausible to believe that consequentialist's and deontologist's 'good' share an important part of the extension. Based on that, as Copp (2007: 214-215) puts it, we can think that earthling's 'good' is the best possible translation of twin-earthling's 'good'. In that sense, we can consider Coco's and Dede's 'good' to have the same *meaning*. In order to hold that, Copp reveals what seems to be an assumption from H&T that I will put in the following terms:

(3) Extensionalism is true for moral terms: the meaning of a moral term is determined by its extension only.

Since Copp accepts the Putnamian metasemantics, (1), the only way to also accept (2), to argue that the meaning of moral terms in Coco-Dede's scenario is the same, is by disassociating extension from meaning. That is, Copp thinks that (3) as false. That way, with a broader notion of *meaning*, Copp satisfies our intuition regarding earthlings and twin-earthlings disagreeing morally by holding that the *meaning* of their moral terms is the same.

Holding that the extension of a predicate isn't enough to know its meaning is not the only way to accept (1) and (2). According to Pekka Vayrynen (2018), we don't have to reject the Putnamian metasemantics in order to have earthlings and twin-earthlings meaning the same by their moral terms. We don't have to accept that different properties are the ones being referred to, as Copp does. Vayrynen suggests blaming the epistemic conditions in which the disagreement occurs, which leads us to believe that different properties might be the ones ruling the uses of Coco and Dede terms. Two communities might use different descriptive characterizations for their morality but that doesn't imply that they are referring to different properties. That is, the differences between earthlings' and twin-earthlings' uses might only be a consequence of the epistemic access we have to whatever objective component morality has rather than a difference in reference. We could have "competing methodologies for inquiring into the nature of the same property"[3]. Vayrynen invites us to think of the possible convergence that the extensions of the moral terms would have under ideal epistemic conditions. If we grant that convergence would happen, we can still accept the Putnamian metasemantics while accepting that the properties referred to by the disagreeing parties are the same; as it would be just a matter of time until these different methodologies bring subsequent closer outcomes. By making epistemic conditions explain the beliefs held by the disagreeing parties, Vayrynen's solution is able to keep consistently (1) and (2) without having to reject (3) as Copp does.

It has also been suggested that the best way to account for the reference fixation for moral terms isn't the ordinary Putnamian metasemantics that we use for natural kind terms,like H2O. A conceptual role for the use of moral terms has been proposed in different ways in order to

[3] Vayrynen (2018: 6)

show how this will grant that the same property is being referred to in disagreements like the one proposed in the Moral Twin Earth example. Matti Eklund (2017) has defended the plausibility of a version of Realism, *Ardent Realism*[4], that seems to be immune to Twin-Earth variations when supported by a conceptual role approach to normativity. Eklund bases his approach on Ralph Wedgewood's (2001) approach to the semantics of moral terms. According to Wedgewood, for an agent to be semantically competent with a normative term, the term must imply, for instance, formulations like the following: « *x* is better than *y*» iff one is «disposed to prefer *x* over *y* at *t*». A semantic description like the latter is not without precedent. Whenever we speak about the meaning of a term like 'and', we typically characterize it in terms of its conceptual role in logical and sentential operators. Such an understanding of the meaning of moral terms allows us to think that even if in Earth Coco *is disposed to prefer x over y*; and in Twin-Earth Dede *is disposed to prefer y over x*, since the conceptual role is the same, it's plausible to take both to be talking about the same property. Such a characterization of the Moral Twin-Earth disagreement gives us room to theorize about the objective truth conditions that the moral concepts might be associated with; this is what Eklund understands as *Ardent Realism.* A similar thesis is defended by J. Robert Williams (2018). He argues that if the conceptual role that two terms have is the same, the reference will stabilize. The conceptual role approach to the meaning of moral terms, then, is a way to accept (2) while proposing an alternative to the metasemantics proposed by (1). It's worth noticing that a conceptual role approach to the meaning of moral terms, similar to Copp's strategy, rejects (3), as something else other than extension seems to be necessary to understand the meaning of moral terms.

All the previous formulations have illustrated to us how can we model a moral disagreement in such a way that both disagreeing parties are talking *about the same thing.* That is, the focus has been on showing how we could avoid readings of the disagreeing parties as talking past each other, through a broad notion of meaning, a story about how both parties refer to the same property, or a shared conceptual role, for example. Nevertheless, these approaches have presupposed the following idea:

[4] Eklund draws just a rough picture of what an ardent realist pursues. It could be put as the claim that the failure to instantiate a normative concept might be characterized through objective truth-values: a thesis that implies that only one property is being referred to by disagreeing parties in the Moral Twin-Earth Argument.

(4) In a substantive moral disagreement, both disagreeing parties cannot be correct at the same time.

Only by the presupposition of (4), that was inherited from H&T's Moral Twin-Earth original formulation, can we understand the theoretical efforts of the previous authors as necessary in order to escape from views where a moral claim is true *just relative to* so and so standards. However, I claim that H&T ignored that in some moral disputes, backgrounds are far enough from each other that it could be necessary to think that one of the disputants is mistaken. On that issue, Khoo and Knobe (2018) have defended the plausibility of moral contextualism by taking empirically tested semantic intuitions as a guide. The following is an illustrative graph of their findings:

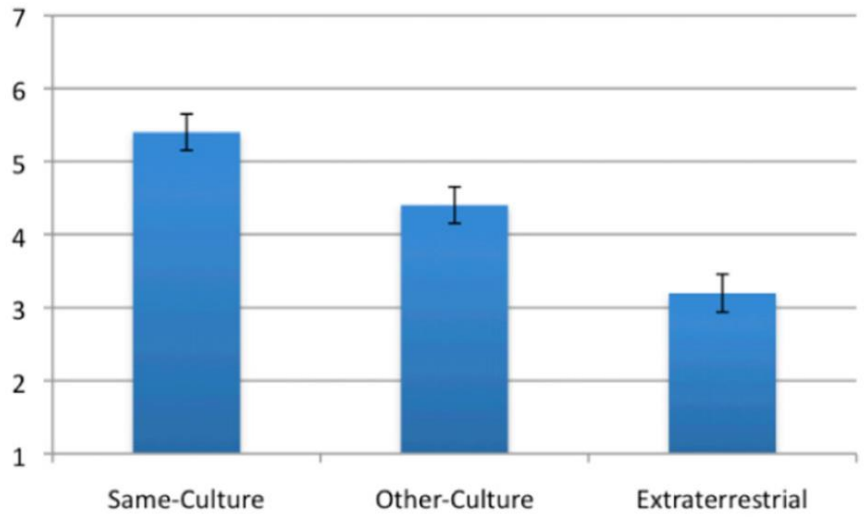

Figure 1. Mean agreement with the claim that "At least one must be wrong" by condition. Error bars show standard error of the mean.

The experiment tested the semantic intuition on the existence of *exclusionary* content in a moral disagreement between two fictional characters uttering opposite moral claims about an action[5]. That is, the experiment tests if necessarily one of the disagreeing parties has to be wrong—or if necessarily both disagreeing parties cannot be correct—given how we use moral terms in a disagreement. The intuition was tested in three different scenarios: when the fictional characters

[5] An individual that got a new knife and decided to test how sharp it was by stabbing the first person he encountered.

belonged to the same culture, when they were from far-away cultures within the same planet (an American student and a warrior from the Amazon), and with an inter-planetarian distance between the origin of the characters. The intuition regarding the truth of the claim 'At least one of them must be wrong' changed from closer-to-true to closer-to-false while the characters were from further backgrounds. Khoo and Knobe (K&K) argue that the experiment shows that our semantic intuitions regarding the use of moral terms allows room for *non-exclusionary* disagreements: that is, we can understand conflicting moral sentences as a disagreement even if both parties could be somehow correct in their own terms. And, furthermore, they go on to argue that the semantic theorizing for our moral terms should give room for non-exclusionary cases. That is, K&K reject (4) by arguing that even if an individual utters something of the form «*x* is wrong» and another one utters «*x* is not wrong», if their backgrounds are far enough away, our semantics should allow the possibility for both to be saying something true. Now, it's worth noticing that K&K are not just pointing out our recognition of moral conflicts where parties are just talking past each other. It's pointed out that people's ordinary intuitions take these non-exclusionary cases as disagreements[6].

A contextualist treatment of the meaning of moral terms has also been defended by Timothy Williamson (2020) in his *parochial*[7] approach for the scope of moral judgements. According to Williamson, the lack of universality aspiration in our judgements is not a threat for the mind-independence of their contents or for their truth-conditionality. Williamson invites us to think of what would 'good' mean for rival sides in a battle. We could take both sides to have conflicting ideas about a possible outcome[8] being good or bad. Victory is evaluated as good and defeat as bad, and both sides are aware that the event one considers good is precisely the one that the other considers bad. Nevertheless, it's not required to think of one necessarily mistaken side. That way, Williamson makes a case of what seems to be compatible with the context-

[6] K&K add: "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". (Khoo & Knobe 2018: 118).

[7] We could understand his 'parochialism' as the claim that the scope of moral judgements is not universal and –in that sense—the meaning of moral terms could be somehow related to a set of interests of a particular community.

[8] A similar idea was defended by Stevenson in his expressivist account of moral judgements. According to him, we could explain away our intuition of disagreement on content by appealing to a sort of "incompatibility of projects". Nevertheless—in opposition to Williamson—Stevenson held that there was not descriptive content in moral sentences, and that they were not truth-evaluable.

sensitivity of moral terms. As he puts it: "if the universalist aspiration is essential to morality, then moral evaluation may play a smaller role in human life than many philosophers, especially moral philosophers and metaethicists, assume. We are often content to make our decisions on parochial grounds" (Williamson 2020: 18). In that sense, moral conflicts might have parties asserting non-exclusionary contents through their claims as their claims are not related to one univocal normative property—like 'good' *simpliciter*. In other terms, both K&K and Williamson could perfectly accept (1), a Putnamiam story of how moral terms fix their meaning, and reject (2), that the meaning of the moral terms in the Moral Twin-Earth scenario is the same.[9]

It was only under the assumption of (4), an attempt to avoid two equally correct claims in a moral disagreement, that philosophers tried to build theoretical machinery such that the same meaning was given to disagreeing parties despite of the different moral backgrounds. Nevertheless, insofar as we could accept the idea of a faultless, or non-exclusionary, disagreement in our treatment of moral judgements, the doors for a contextualist account for the meaning of moral terms open. However, more has to be said about the characterization of whatever contextualist component moral terms are sensitive to—in a way compatible with non-exclusionary disagreements, but also, with room for exclusionary ones. In the following section, I work on how this contextual component might look like.

3. Contextualism and Non-Exclusionary Disagreements

A contextualist approach to the meaning of moral terms entails that there is something missing in the content of an *out of the nowhere* moral sentence. Somehow, in order to understand what someone is saying with her moral terms, contextualism suggests a standard or a parameter that complements the moral claim. Since it appears the possibility of more than one standard in which our moral terms are somehow *indexed*, it also appears the possibility of faultless disagreements. For instance, coming back to our original moral disagreement, Coco might be

[9] Another theory that fits this approach of accepting (1) while rejecting (2) is defended by Plunket & Sundell (2013). They argue that our intuition of disagreement between parties with different moral beliefs could be taken as a *metalinguistic negotiation*. That is, disagreeing parties in a moral discussion are "negotiating" what is the best meaning for the term 'good'—or whatever normative term in question. That way, without having to accept that the meaning of the moral term was originally the same, we can understand in what sense they are having a disagreement—that although metalinguistic—has substantive consequences.

indexing his claim of lying for saving a life being good to consequentialist parameters; and Dede doing the same with her claim of lying for saving a life being wrong to deontological parameters. In such a way that the truth of one claim does not rule the other's one out.

As we showed in the previous sections, such a phenomenon, faultless moral disagreements, seems to be compatible with our semantic intuitions regarding the use of moral terms, and with that, a semantic treatment that models that is motivated. In the following, I draft an application of the contextualist framework to the Moral Twin-Earth scenario.

In H&T's Moral Twin-Earth original formulation, earthlings and twin-earthlings speak languages very much alike. Let's consider that a *language* is a function from sentences to truth-conditions[10]. In that sense, we can take two languages to be alike if they assign the same truth conditions to the same sentences in a large enough number of cases. In such a way that if two speakers utter the same sentence but it has different truth-conditions, we could take the speakers to be speaking different languages. Let's say that an earthling speaks E-English and a twin-earthling, T-English. Let's say that the moral sentence *s* is true in E-English but false in T-English. That means that *s* is true in E-English and *not s* is true in T-English. The risk of a standard notion of relativism appears in formalizations like the previous one because, as H&T anticipated, we might be allowing paradigmatic moral disagreements to be such that each part is correct under their own terms, as they are saying true things in their respective languages. Nevertheless, as K&K's empirical test shows, an outcome like the previous one shouldn't be rejected *prima facie* since it is compatible with how our use of moral terms works. Moral disagreements, whose disagreeing parties are asserting non-exclusionary contents, are still disagreements[11]. But we still need to clarify the story behind the existence of exclusionary or non-exclusionary content in these disagreements.

The only logically consistent way to allow earthlings and twin-earthlings to utter the same moral sentence, but also allow the sentence to be true for one and false for the other, is to argue

[10] Since we are building a contextualist account, we could take languages to be functions from sentences to *characters.* A character –in the Kaplanian sense—is the part of the meaning of a declarative sentence that together with a context of emission are sufficient to know what the proposition is the one expressed. A proposition is the content expressed by a contextualized sentence of a specific language. A proposition is a function from circumstances of evaluation to truth-values. Nevertheless, for the purposes of my work, we can take a language to be a function directly from sentences to truth-conditions.

[11] Khoo and Knobe (2018), for instance, characterize moral disagreements as—even if non-exclusionary—relying on the proposals speakers do to update the common ground.

that that moral sentence has different truth conditions in the earthling utterance and in the twin-earthling one. A disagreement like the previous could only be understood consistently under a *contextualist account.* That is, a moral claim is indexed to a contextual parameter of use in order for it to have truth-conditions. At the same time, we don't want to reduce absolutely every moral disagreement to a non-exclusionary one based on differences in truth-conditions. If we are to follow K&K's test regarding our moral semantic intuitions, we should also allow room for exclusionary disagreements: that is, we should allow that people can express the same moral proposition from the utterance of the same sentence. One way to do it is to follow K&K's suggestion and try a contextualist approach. If the same sentence is indexed to the same local parameter, we obtain the same proposition (like in the same-culture-like cases) and distinct propositions in the human-extraterrestrial-like case. That way, we could satisfy the semantic intuitions regarding the use of moral terms in disagreements with exclusionary content in same-culture cases and non-exclusionary content in the extraterrestrial case.

As we observed, H&T ignored the possibility of substantives non-exclusionary disagreements—insofar they held (4), whatever seem to be disagreements where both parties could be simultaneously correct were not actually disagreements. With them, philosophers trying to address that challenge posited by the Moral Twin-Earth Argument also ignored the possibility of this kind of moral disagreement. The strategies that we observed in the previous section built by Copp, Williams, Eklund, Väyrynen, don't account for non-exclusionary disagreements since they were trying to avoid the existence of more than one property as reference for moral terms. For instance, a contextualist approach for the meaning of moral terms is in a position to explain scenarios like the following:

Let's imagine that in the Moral Twin-Earth disagreement, twin earthlings, instead of being human-like deontologists, would have been non-anthropomorphic entities that had a moral theory related to Nitrogen-Accumulation Principles. That is, their moral system will consider something to be *good* if the action accumulates a certain amount of nitrogen. In that case, earthlings and twin-earthlings could have been disagreeing about the truth of the moral sentence 'Killing lives in order to get nitrogen is morally wrong' and that disagreement would have been clearly non-exclusionary. The participants of the disagreement could have been defending something true (even though one claim appears to be the negation of the other)

because their claims are indexed to different contextual parameters. H&T would probably reply that if this new disagreement between these new earthling and twin-earthling were a moral one, it would have to be substantial otherwise a standard relativism would come into place. However, I argue that the standard relativism that H&T fear, in this case, is a desirable consequence for a semantic treatment that allows non-exclusionary disagreements. The earthling and the twin-earthling's utterances are relative to different contextual parameters and, in that sense, a fair semantic treatment, since different properties are being referred, has to allow the possibility of both utterances being true at the same time.

One way to understand the contextualist approach that we are proposing would be the following. Let *C* be a set of moral claims. And let a moral theory *M* be a finite set of sentences $\{s_1, s_2, s_3 \ldots s_n\}$. *M* works as a function from the set *C* to a set of truth-values that correspond to each moral claim. A Moral Theory sets the truth-conditions of a moral claim because a moral claim has truth-values in virtue of a moral theory. One way to explain why earthlings and twin-earthlings could have non-exclusionary moral disagreements is to index a different *M* to the earthling and the twin-earthling's utterance. That way, despite how exclusionary both utterances look, they would be related to different truth-conditions allowing the possibility of both moral sentences simultaneously true.

On the other hand, if the same moral sentence *s* uttered by an earthling and by a twin-earthling were indexed to the same moral theory, it would have the same truth conditions. Only with moral claims whose moral terms somehow indexed to the same moral theory, we would have the possibility of exclusionary disagreements. That way, the intuition presented by K&K in their test is being met. Since it's natural to think that two people being culturally closer makes it easier for them to share a moral theory, we can now explain why the exclusionary feature of the disagreements rises in culturally close cases. We take moral disagreements in culturally close cases as being exclusionary ones because it is assumable that they are under the same moral theory—unless, of course, there were clear reasons to interpret otherwise. Similarly, the

further the communities of the people disagreeing morally, the easier it is to imagine that they endorse different moral theories: so, non-exclusionary disagreement's intuition is explained[12].

H&T might still reply, nonetheless, that substantive moral disagreements are something different than merely linguistic disagreements and that our contextualist account is reducing some moral disagreements to that. In the following section, I would like to address that reply by presenting a Hirsch-like approach for differentiating substantive from merely verbal disagreements.

4. A Hirsch-like Argument for Substantive Disagreements

It might be argued that moral disagreements are not the kind of disagreement that can be solved just by adjusting our terminology or by making explicit which moral theory works as a contextual parameter. If an earthling typically indexes its moral claims to the moral theory M_1 and a twin-earthling to the moral theory M_2, then it will seem like their disagreement might be solved just by agreeing to the vocabulary. Let's take, for instance, the Coco-Dede disagreement (where Coco is a consequentialist and Dede a deontologist). When Coco is saying 'Lying for saving a life is not morally wrong' (according to our contextualist approach), he would be saying something like 'According to consequentialism, lying for saving a life is not morally wrong'. Similarly, Dede would be saying something like 'According to deontologism, lying for saving a life is morally wrong'. So, just by making the moral theory explicit, someone could think that the disagreement could be solved, as Coco wouldn't have any problem in accepting Dede's utterance. That might make us think that moral disagreements modeled in that way would be reduced to merely verbal disagreements, as H&T feared. This is not the case. On this issue, Eli Hirsch (2009) addresses the differences between disagreements that are in reality versus merely verbal ones. So, following a Hirsch-like line of argumentation, I present one reason to believe that moral disagreements, even non-exclusionary ones under our contextualist approach proposed above, cannot be reduced to merely verbal ones by comparing them to actual mere verbal disagreements.

[12] It might be argued that debates on normativity are such that even with different moral theories, the intuition of exclusionary content remains. This issue is addressed in Section 6 where I argue that moral disagreements, whose parties are indexing their claims to different moral theories, might still disagree on the quality of the theory.

I would like to present how a mere verbal disagreement would look like. Let's imagine that someone comes from a region X where they don't think that people «die» but that they «pass away». In a region Y, people think that people never «pass away» but they «die». If the inhabitant of X goes to the Y region and they are at the funeral of the baker of the town, the X-inhabitant will disagree when Y-people utter the sentence 'The baker died'. The X-inhabitant could reply that 'It is false that the baker died, he passed away'. Y-inhabitants could, of course, reaffirm themselves by saying 'You, X-inhabitant are totally mistaken, the baker didn't pass away. The baker died'. We are observing a highly shallow disagreement that has a clear solution. Let's imagine, for instance, that when the X-inhabitant arrived in the Y-region, he pretended to keep a low profile and to avoid being recognized as an X-inhabitant. The X-inhabitant decided to speak and behave like the Y-people but at the same time, he decided to keep all the beliefs he had from the past. So, the X-inhabitant decided to create a secret language: every time he would say or hear that someone «dies» he would think to himself that someone «passed away», that way his beliefs will be kept and he could accept what the Y-inhabitants tell him during the funeral. After this change in the meaning of those sentences that the X-inhabitant did in his mind is made, there will be no possible complication during his secret visit to the Y-region with that respect. No other possible dispute will arise from that sentence secret translation that would put at risk his secret visit and no ramifications of semantic complications are possible. The X-Y inhabitants' disagreement about the truth of 'The baker died' is, then, merely verbal.

Now, let us observe what would happen if the X-inhabitant didn't think that people «die» but that they «sleep for three days and resuscitate». In his travel to the Y-region, this time, the X-inhabitant decides to create the following secret language: every time he would say or hear that someone «dies» he would think to himself that someone «sleeps for three days and resuscitates». In this case, multiple possible disputes would arise after this arbitrary translation, even if he succeeds in interchanging a term for the other in his mind. Clearly, the X-inhabitant doesn't just have a merely verbal disagreement with the Y-inhabitants about if people just «die» or «sleep for three days and resuscitate». In that sense, the verbal translation fails.

Let us imagine now that Coco secretly travels to Twin-Earth and he wants to be there among the twin-earthlings without them noticing that he is an earthling. However, Coco doesn't want to forget his moral beliefs, so he creates a secret language that he only speaks while talking to

the twin-earthlings. Whenever Coco utters 'Lying for saving a life is morally wrong', he is going to mean (secretly in his mind) 'Lying for saving a life is deontologically wrong'. We could be tempted to conclude, like in the «die»-«pass away» example, that since the twin-earthling's sentence can be translated into a sentence that Coco is willing to accept as true in his secret language, the disagreement is merely linguistic too. But that would be a mistake. Let's imagine that during his trip Coco is caught *lying for saving a life*. What honest answers could Coco give to justify his behavior in front of the twin-earthlings? Notice that Coco accepts the truth of the sentence 'Lying for saving a life is deontologically wrong', a sentence both Coco and twin-earthlings agree with, yet he still disagrees with the twin-earthlings.[13] Coco cannot solve the disagreement just by virtue of hiding the verbal differences through a secret language. If Coco or Dede are guided by their moralities—as the original Moral Twin-Earth experiment requires—then non-linguistic behavior will put into evidence Coco's disagreement with twin-earthlings. This wouldn't happen if the disagreement were merely verbal. More than just vocabulary changes are needed by Coco to hide his disagreement. Now, it might seem that resources from outside contextualism are being imported to explain away the intuition of disagreement. Nevertheless, in general, it is only by virtue of resources from outside semantics that we are able to figure out if a disagreement is merely verbal or not. In that sense, the same conclusion holds for whatever piece of information from the world that is sufficient to show how making explicit the moral theory behind claims doesn't settle moral disagreements.

In Hirsch terms, "more is needed for an issue to degenerate into "merely a matter of choosing a language". It is required that each side ought to find it plausible to interpret the other side as speaking the truth in the other side's language." (Hirsch 2009: 238) The psychological reality of us endorsing our moral theories avoids the possibility of Coco, or Dede, finding plausible to interpret Dede's, or Coco's, belief as being true in the other side's language. It's highly implausible that a moral disagreement could be solved that way. That is, the following argument holds:

[13] It might be argued that a better characterization of the twin-earthling sentence 'Lying for saving a life is wrong' inside Coco's mind would be something like 'Lying for saving a life is wrong according to deontologism & deontologism is correct'. Nevertheless, since what is or isn't *morally wrong* in Twin-Earth is just—sort to speak—what is *deontologically wrong*, I believe that the chosen characterization proves the following point: even though there's sentence that Coco is willing to accept in replacement of the twin-earthlings' moral claim, that doesn't rule out a possible disagreement.

P1 A disagreement is merely verbal if it could be solved just by virtue of a change into a vocabulary the parties accept.

P2 Coco and Dede cannot solve their disagreement just by adjusting their vocabulary.

C Coco and Dede's disagreement is not merely verbal.

Even if we create a possible language in which a sentence we disagree with is true, the disagreement can still be substantive. Thus, it doesn't follow from H&T's Moral Twin-Earth argument that if there's room for Coco's and Dede's claims to be simultaneously true, then we wouldn't be able to explain how's that a disagreement. As we have seen, since mere verbal disagreements behave differently, we have reason to believe that non-exclusionary moral disagreements might not be reduced to purely verbal ones. Anyone arguing that moral disagreements under a contextualist approach are reduced to purely verbal ones should explain why moral—and other substantive—disagreements don't seem to be settled just by vocabulary adjustments.

5. The Exceptionalist Temptation

Moral exceptionalism comes in different forms. It could be formulated as the claim that a general semantic theory that works for non-moral terms, doesn't work for moral terms. Or as the claim that a general metaphysical theory, that works for non-moral properties, doesn't work for moral properties. In general, moral exceptionalism demands an exceptional theoretical treatment for the moral claims with respect to how non-moral claims are treated in semantics, epistemology, metaphysics, etc[14]. For the purposes of the present thesis, I focus on exceptionalist treatments that could somehow challenge a realist treatment of moral properties under a Putnamian metasemantics. In the following, I present complications that arise from holding moral motivational internalism, and from holding moral semantic internalism: two theses that could support an antirealist treatment of moral claims[15].

[14] Formulated as the more general thesis of exceptionalism of the normative, a similar exceptionalist approach is characterized by Williamson (2020).

[15] When arguing against moral semantic internalism (MSI), we address also realist versions of it. In that sense, the scope of the present section includes moral antirealism but is not limited to it.

Moral antirealism could be formulated as the claim that there's no truth in morality or that there's nothing in the world that moral sentences are describing. Since more has to be said by the antirealist about what moral claims are about, in this section, I draw what I think are the motivations behind an antirealist approach and I point out the reason why not to adopt it. One motivation for adopting an antirealist approach is ontological economy. If moral properties are irreducibly normative, and the rest of the properties in our ontology are not, we might very well prefer a simpler metaphysical theory and find a different story for the role of our moral terms rather than referring to moral properties. This idea has been presented as the *Argument from Queerness*[16]. I think of this antirealist move as a temptation since it would indirectly address the challenge posited by H&T's Moral Twin-Earth Argument. Let's remember that the Moral Twin-Earth Argument leaves us with no easy solution for our explanation of how moral terms get their references. An antirealist motivated by the Argument from Queerness wouldn't have to account for this explanation as she could directly hold that moral terms have no reference whatsoever[17]. The moral exceptionalist that endorses moral antirealism can appeal to something else to explain our intuition of disagreement: something internal to the speaker.

It could be said that one of the reasons why philosophers find moral claims interesting is because of their *magnetism*, or their connection to the speaker's motivation when it comes to behavior. This line of reasoning has motivated different versions of internalist approaches to moral judgements and the meaning of moral terms. In other terms, the internal world of moral speakers has been given a leading explanatory role over external features speakers might be referring to. Approaches in this direction might very well be part of the explanatory tools an antirealist could use to show in what moral disagreements rely on—by pointing out what moral claims imply or commit speakers to. The internalist approaches I'm interested to argue against are the versions of them that an antirealist could use to escape from the Moral Twin-Earth challenge. Either by explaining away our intuition of disagreement by appealing to a necessary

[16] See Mackie (1977), Olson (2014).
[17] Either by claiming that all moral claims are false, or by claiming that moral sentences are not truth-evaluable.

motivation in moral claims[18], or by claiming that nothing outside the speaker is being referred by moral terms.[19] Let's start with the following version of internalism:

(MMI) If Moral Motivational Internalism is true, for *x* to believe the moral proposition *'F is morally right'* implies a motivation-like attitude of *x* towards *F-ing.*

In the following, I present two problems that MMI has to face despite it being better suited to account for the aforementioned magnetism of moral claims.

The first problem for Moral Motivational Internalism (MMI) is the complications that arise while accounting for *akrasia*—or *lack of will*. The lack of will obtains when something seems to affect the motivation of an agent regardless of her moral beliefs. The akrasia element seems to require a complicated subdivision in beliefs that only apply to moral beliefs. For instance, it is plausible to think that CEOs of the biggest companies in the world are aware of the poverty and hunger in the world. It is plausible to believe that they think that a world with less hunger is a better world so it would be plausible to believe that the CEOs believe that decreasing hunger is morally correct. They also know that if they donate 10% of their salary every month, they will decrease the hunger in the world without affecting their quality of life. However, we can acknowledge that it's also plausible to believe that those CEOs don't have any motivation towards donating 10% of their salary every month. The moral internalist could propose two things, either that the CEOs are experiencing akrasia because they have a moral belief without the motivation, or that the CEO's are not honestly believing those moral claims. That would mean that either there are two kinds of moral beliefs—and, in that sense, two kinds of beliefs in general—, the ones vulnerable to akrasia (beliefs in which a specific mental state affects their functionality) and beliefs that are not vulnerable to it; or there are two kinds of moral believing: honest believing and not honest believing. In any case, new complications to

[18] One of the ways an antirealist could start an explanation of what moral disagreements rely on could appeal to *motivation to incompatible projects*, for instance. That way, no story of how moral terms get their references is needed—since moral terms might very well not have one. Nevertheless, I intend argue against the necessitation of a motivation-like attitude in the assertion of moral claims rather than against a particular strategy regarding how this might be used to explain away disagreements.

[19] Similarly, it could be argued that the idea that the references of moral terms are not external to the speaker doesn't automatically set the internalist free of the H&T's challenge. The internalist should still find a metasemantics for moral terms that is compatible with substantive moral disagreements. Nevertheless, the generation of a story for the fixation of references for moral terms is particularly hard to compatibilize with substantive moral disagreements when we assume that what is expressed doesn't rely on the speaker only.

our theories of belief appear that would have to be taken into consideration because of the akrasia cases, and they would seem to be *ad hoc* in order to save the internalist's commitments.

The second problem that I would like to present for MMI is the complications that arise when we want to theorize about moral terms in more complex environments such as the terms being embedded within a counterfactual, under the scope of doxastic operators, or within impossible, or fictional, scenarios. MMI is easily and smoothly understood when moral terms appear in simple (unembedded) sentences predicating a moral predicate from an event; like 'Killing is wrong' or 'Abortion shouldn't be forbidden'. Nevertheless, the way motivation is tractable from sentences with moral terms in embedded environments is not, by any means, transparent. For example, we can have moral sentences of the kind 'If Hitler hadn't killed anyone, he wouldn't have been an immoral person' or 'Torturing unicorns would never be a good thing' or 'Going faster than the speed of light to stop a meteorite from crashing the earth would be a moral thing to do' or 'If humans were immune to acid, throwing acid to other humans wouldn't be an immoral thing to do'. If we would have to propose a motivation-like attitude, in whatever form, from agents that honestly believe any of these sentences to be true, we would have to tell a non-obvious story about how this motivation attitude is endorsed in non-simple sentences like the ones mentioned. If MMI proposes a necessary connection with motivation, since these are cases where the akrasia element doesn't seem theoretically relevant, we would require an explanation of how such a mental state could be ascribed to these possible events. And that's an explanation that we don't have to give if we deny MMI. In that sense, if the moral exceptionalist that endorses moral antirealism, takes use of MMI to explain away our intuition of disagreement, the aforementioned complications arise.

The moral exceptionalist could also argue that, unlike the general semantic treatment of most non-moral terms, moral terms' semantics shouldn't be understood in terms of external properties. Either by the effects of holding a fully-fledged moral antirealism, or by holding that whatever reference moral terms might have are internal to the speaker, I present the complications that these versions of moral exceptionalism might face. In the following, I characterize what I call Moral Semantic Internalism. Then I show how the possible implications of this theory make problems with truth-conditional treatment arise. Let's accept the following definition:

(MSI) If Moral Semantic Internalism is true, moral properties referred to by moral terms are not external to the speaker.

A moral exceptionalist might have to accept a Moral Semantic Internalism[20] (MSI). Either because she claims that there are no moral properties whatsoever, or that moral properties are not external to the speaker[21]. If MSI is endorsed, then, one of the following three theoretical implications will have to be held:

A) Moral claims are not truth-conditional.

B) Moral claims are truth-conditional, and all moral sentences are false.

C) Moral claims are truth-conditional, but moral properties are not external to the speaker.

In the following, I show the complications that arise when we accept any of the previous theoretical commitments. That way, I intend to argue against MSI, one of the assumptions a moral exceptionalist might have in her attempt to theorize on moral claims on grounds that are internal to the speakers.

Regarding (A): truth in morality is a desirable feature, as it would directly solve the challenges of the Frege-Geach problem. The Frege-Geach problem could be characterized as the posing of the following question: How is it possible that we can handle moral sentences as other descriptive claims if they aren't describing anything? We are able to make logical inferences from and within moral sentences that we need to account for in order to keep rationality in our treatment of moral contents. From 'It is wrong to kill people' it follows that 'It is wrong to kill Latin-American people'. Any expressivist theory that implies that (A) is true will not be in an easy position to explain that inference[22].

[20] It could be better put as 'Moral Semantic Anti-externalism' but for explanatory purposes I think that—due to its possible consequences—we could understand it as some sort of internalism too.

[21] Perhaps a theorist self-identified as 'antirealist' could argue a sort of moral fictionalism to hold that moral properties are *non-real* in a sense but external to speaker—as fictions are external. Nevertheless, even if such an account is consistent I wouldn't consider her as a moral antirealist or as an moral exceptionalist. In any case, the arguments presented are no directed to such a theorist.

[22] It's important to notice that I'm only arguing against expressivist accounts that imply (A). That is, theories about the meaning of moral terms that claim that sentences with moral terms are not truth-evaluable. See Ayer, Stevenson. For a version of expressivism that doesn't imply (A), see Gibbard (1990), (2003), (2012).

Regarding (B): logical inferences in moral sentences can be explained only by appealing to truth-conditionality, not necessarily to truth. A version of error-theory can be postulated by affirming (B) to preserve the intuition about the existence of valid arguments made of moral sentences without commitments on the existence of external facts that make those sentences true.[23] The problem of any metaethical theory that implies (B) is the following: we can give moral information with negated sentences. Sentences like 'The abortion is not wrong' or 'To avoid paying taxes is not a good thing' are legitimate moral claims. If (B) is true, then a moral claim like 'Abortion is wrong' must be false. So, it would follow that 'It is false that abortion is wrong' is true. But since we can have moral claims using negated sentences, claims like 'Abortion is not wrong' are at the same time true and false. Any theory that endorses (B) will have similar undesirable consequences[24].

Lastly, any theory that implies (C) will have Putnam's Twin-Earth-style counterexamples. Let's imagine that an agent *x*, after observing the behavior and success of different generations under a system of parental punishments where everybody is and seems happy, comes to believe and claim that 'physical punishment to children is morally good'. Since, given the assumption (C), whatever makes that moral sentence true is up to properties of the speaker[25] we might think that whatever truth-conditions that sentence has, a counterpart of *x* claiming the same shares them. Let's assume that *x* is an inhabitant of the Earth and that the sentence is true. Let's imagine now a Twin-Earth with exactly the same phenomenal experience for the corresponding twin-*x*, with the same experience of evidence gathered regarding successful education with parental punishment, the same evidence regarding everybody looking happy; but in Twin-Earth, everybody just seems happy while in the inside, everybody is miserable and profoundly unhappy. Our twin-earthling, twin-*x*, again, claims 'Physical punishment to children is morally good'. If the truth of the sentence 'Physical punishment to children is morally good' depends on the inner-goings of the speaker, twin-*x* would be saying also something true in the twin-earth scenario if she utters that sentence, independently of the facts in the new planet that intuitively

[23] See Mackie (1977), Olson (2014).
[24] Now, a defender of moral error-theory could argue the following: all moral claims are false, 'x is wrong' is false, and 'x isn't wrong' is false. Nevertheless, 'x isn't wrong' gets two readings: the normative reading—that is false—and the non-normative reading that is just the negation of the 'x is wrong'. In that sense, this version of error-theory could be defended by positing an ambiguity in all negative moral claims. However, even if this move saves error-theory from the inconsistence, it doesn't help with its simplicity.
[25] E.g. whatever disgusts the speaker, whatever sounds correct to the speaker, etc.

make the moral sentence false. Any moral exceptionalism that endorses (C) will be incapable of explaining the difference in truth-value of the two moral claims on these Earth and Twin-Earth scenarios. For similar reasons, any moral internalism that endorses (C) won't be in a position of modeling moral mistakes, which is a desirable consequence for any realist approach[26].

We have seen in the problems presented that if moral exceptionalism endorses any of the different assumptions of MSI, it ends up arriving at undesirable theoretical consequences. However, we might still ask the anti-exceptionalism theorist[27], the one that treats moral terms as referring to external moral properties as with most non-normative terms, how to escape the risk of triviality. That is, how to account for the normative nature of moral claims just by virtue of truth-conditional descriptions—that is, without trivializing them or stripping them from their functions. I think that one of the ways this theorist can explain how to escape from the risk of triviality is by pointing out the existence of a contingent psychological reality: we want to do what is right. An approach like the one used by Shafer-Landau (2000) could be used, for example. Shafer-Landau (2000) argues that ethical behavior is the consequence of moral beliefs plus a practical ingredient: the *motive of duty*. So, to claim properties external to the speaker make moral sentences true does not deny the existence of a tendency to feel motivated towards acting morally, it just denies that is something that has to be explained by the semantic content of moral beliefs or moral terms. The existence of a widespread psychological tendency to do what we think is moral is why philosophers have considered morality something interesting to investigate. However, that doesn't mean that the meaning of moral terms needs to include information on our widespread psychological reality.

So far, in Section 3, I have motivated a contextualist account of the content expressed by moral claims. In Section 4, I have shown why it does not follow that this contextualist account would reduce moral disagreements to linguistic ones. In the present section (5), I have motivated an anti-exceptionalist approach. That paves our way towards a contextualist, truth-conditional, motivational externalist, and semantical externalist account for the meaning of moral terms: roughly, a moral claim determines its content by being sensitive to the context of use, and this

[26] A simpler explanation would be just to point out that endorsing (C) makes moral claims self-validating. And that would be incompatible with how we do moral claims.
[27] See Williamson (2020).

claim might be true or false, and this doesn't leave us with merely verbal disagreements. In the following section, I argue that not all moral theories, what moral sentences are sensitive to, are equally valid. I show this through a variation of the Moral Twin-Earth Argument that we have strong intuitions and theoretical resources to compare, differentiate, and value different moral theories.

6. Diachronic Moral Twin-Earth and Choosing Within Moral Theories

The things that we have considered as morally wrong or morally permissible have changed through time. Western societies used to consider slavery as a morally permissible activity, more recently, homosexuality was considered morally wrong. Ancient Greeks considered pedophilia as something morally permissible. We could always say that our morality just *changed* simpliciter. Our contextualist account might, indeed, be accused of such modeling; that the only thing that differentiates our contemporary western morality from the morality of our slavers ancestors is just a *different theory* that makes our claims true. That would leave us with an irrational modeling of the contents of morality through time. There is an intuitive sense in which moral theories got better since we wouldn't want to change back to theories that consider slavery or pedophilia as morally permissible. Or, at least, there is an intuitive sense in which we think that a moral theory could be better than others. But, how can we theorize about progress, or *betterness*, in moral theories? I think that it would be an important feature of any theory of the content of our moral terms to keep space for such an explanation, contextualist or not.

If we take individuals (or societies) to be capable of moral learning, we can theorize about the content of the moral beliefs these individuals have at t_2, which they lack at t_1. Then we can theorize within a variation of the Moral Twin-Earth thought experiment on a disagreement between an earthling from t_2 and a twin-earthling, which moral claims are ruled by the same moral principles that ruled upon earthlings at t_1. Our contextualist account must treat the truth-conditions, partially given by moral theories, of these disagreeing claims as standing in a relation *R*. Whatever the way we would like to treat a hypothetical moral disagreement between ourselves and an individual from the past of our society, e.g., moral disagreement with a slaver about the truth of 'slavery is morally right', we must keep the same relation *R* between the truth-conditions of the moral beliefs. I propose that we can theorize about progress in moral theories, by observing the nature of the *R* relation between the truth conditions

of the disagreeing claims. Progress will be given the same way we think there is progress in other theories: objective metatheoretical criteria. Moral theories, or whatever gives truth-conditions to moral claims, are sensitive to evaluation. We can prefer moral theories for being more parsimonious, for having compatibility with other theories, or for being sensitive to new data. For instance, an ancient Greek moral theory that says that *it is wrong to torture children unless it is done for pedophilic reasons* is a worse theory than one that says that *it is wrong to torture children even for pedophilic reasons*: it would have fewer exceptions. Similarly, maybe new information was revealed about children's psychology during the forthcoming centuries, so moral theories could have adapted to new bodies of knowledge, like psychology, by changing some of the things it considered morally permissible. Moral theories adapting to new bodies of knowledge is one reason to think that moral theories can get better. In the following, I build the variation of the Moral Twin-Earth Argument from which I work on.

Let's say that on the planet Earth, in the year 2021, earthlings behave, think and argue morally following a set of principles M_1 that constitute the moral theory earthlings index their moral sentences to. From M_1, one can easily infer that the sentence 'Under any circumstance, slaving black people is wrong' is true. That is, similar to what we would expect from our best contemporary moral theories, 'Under any circumstance, slaving black people is wrong' belongs to the set of moral claims *C* and M_1 is a function that assigns that sentence the truth value *true.* Let *s* be 'Under any circumstance, slaving black people is wrong' and let:

M_1 (*s*) = T

Similarly, let be that on the planet Twin-Earth, things are pretty much the same as things were on Earth two hundred years ago. That is, slavery was still in place, and it was socially accepted, even among philosophers of the time. Again, twin-earthlings behave, think and argue morally following a set of principles. Let's call 'M_2' the moral theory that assigns the truth-values to the set of moral claims twin-earthlings believe in. Without many surprises, let's grant that according to twin-earthlings' morality the sentence 'Under certain circumstances, slaving black people isn't wrong' is true –or, in other words, *s* is false. So, it follows that:

M_2 (s) = F

An earthling, Neo, and a twin-earthling, Morpheus, will now be our characters. H&T's Moral Twin-Earth has shown us that we will have troubles in order to model the disagreement between Neo and Morpheus. After all, since the content of their utterances of '*s*' and 'not *s*', respectively, is partly determined by a contextual parameter, M_1 and M_2, respectively, the proposition expressed by Morpheus is not quite the negation of the one expressed by Neo. In that sense, there's no immediate way to take Neo and Morpheus to be disagreeing as one is expressing *p* and the other is expressing *not q* (and not *not p*). Nevertheless, as we have seen in section 2, faultless disagreements are not without motivation. In that sense, there is no theoretical obligation to model moral disagreements as the acceptance and the denial of the same proposition—we would if there were reasons to think that the disagreement has speakers endorsing the same moral theory. But in this non-exclusionary case, so far, we don't have any theoretical angle to compare M_1 with M_2. From what has been shown so far, any theory is in a position to obtain and there's no ground to complain about other theories. But this might strike us as counterintuitive: the following variation will make that clearer, however.

Let's take a variation of the thought experiment we have just formulated. Let's imagine everything happening just on the Earth of the year 2021. Neo hears his neighbor, Morpheus, a neo-Nazi, claiming that *not s*; that is, Morpheus says that slaving black people is not always wrong. Let's say that Morpheus' moral claims can be derived from the set of sentences M_2. If we were in Neo's position, it's clearly not usual that we just consider that we are both correct in our own terms. There are properties and relational properties that theories have that have explanatory roles regarding why a particular society changed from one Moral Theory to another. We typically categorize and compare theories. But, on what grounds do we compare within theories? What criteria could we use in order to argue that this evaluation is legitimate? I argue that we could use what other bodies of knowledge use for comparisons within theories: objective metatheoretical criteria. In order to illustrate these criteria, I will show how these criteria have been used in the comparison of theories of natural sciences.

We could all understand the sense in which Newton's theory of gravity was incorrect and was *replaced* by Einstein's theory of gravity. Let's accept that both Newton and Einstein shared an object of study. Nevertheless, insofar as we accept physics as proposing idealized models of fundamental features of reality, Newton and Einstein had different sets of axioms in

their models. In that sense, when we think of the content of the assertions Newtonian and Einsteinian physicist did, we face, similar to H&T's challenge, complications to model the disagreement. Take the sentence 'Gravity is a constant force'. Since it's a sentence a Newtonian physicist would affirm and an Einsteinian one would deny, we can take them to be disagreeing. However, given that the constructions of gravity both physicists do are determined by different sets of axioms, there's a sense in which they are talking about different things. The Newtonian physicist is affirming something like 'Newtonian gravity is a constant force', and the Einsteinian something like 'Einsteinian gravity is not a constant force'[28]. Again, we can understand the situation as an exchange in which both are correct under their own terms. But we usually don't think of that as the end of the story. We compare and adopt theories in physics by evaluating their internal consistency, parsimony, elegance, compatibility with other bodies of knowledge, explanatory capacity, natural *joint carving*, etc. To have a disagreement where disagreeing parties are both correct in their own terms is no threat to think of one as a better theory than the other. This kind of disagreement doesn't imply any sort of relativism either; it doesn't create relativism in physics. So we shouldn't think that non-exclusionary disagreements imply relativism in morality as H&T suggest. Coming back to our original Diachronic Moral Twin-Earth, Neo and Morpheus disagreement, even though we can theorize on their disagreement as considering both to be saying something true in their own terms, it doesn't follow that any moral claim is equally valid as we can still avoid that relativism by comparing the moral theories behind: the same way we do with theories in other bodies of knowledge.

To think of moral theories as sets of sentences allows us to individuate the contextual parameter moral claims are sensitive to. To think of the contextual parameter moral claims are sensitive to as moral theories allows us to understand what changes when morality changes through the time. Modeling the moral change in terms of change in moral theories allows us to understand it from a kind of change we are already familiar with: adopting of theories in virtue of better explanatory capacity, consistency, parsimony, compatibility with other theories, as we do with scientific theories[29]. It also opens a door to think of *moral progress* the same way we

[28] We cannot take both Newtonians and Einsteinians to be disagreeing about the truth of a sentence 'whatever makes things fall is a constant force' or something like that. Because both think that *something different* is what makes things fall. In that sense, the problem doesn't just rely in the term 'gravity'.

[29] There is an immediate reply to this idea. The way we explain moral terms without making use of irreducible normative properties ends up using normative properties (high order normativity) to escape from a first order

think of progress in science, since we aren't with a merely irrational story of how morality changes through time. It also allows us to understand in what sense moral is learnable: we require moral education as we can learn moral theories.

7. Conclusion and Final Remarks

Theory-Indexed Moral Contextualism is intended as a way to face the challenges posit by H&T's Moral Twin-Earth Argument. After presenting the challenge, the presuppositions of the challenge, and a group of strategies that faces the challenge (Section 1 and 2), I sketched my own version of what I think is the most promising strategy to face H&T's challenge: moral contextualism (Section 3). Once the contextual parameter was characterized, I have shown (in Section 4) how a contextualist approach to the meaning of moral terms avoids the reading of moral disagreements as merely linguistic disagreements: moral disagreements aren't just solved the way linguistic disagreements are. I have motivated a realist and externalist truth-conditional treatment (in Section 5) of the meaning of moral terms by presenting the problems that moral exceptionalism deals with if it's held either that (1) moral claims entail a motivation-like mental state from the speakers, or that (2) there are no moral properties external to the speaker. Finally, (in Section 6) I have shown the tools that my contextualist approach, as a deflationary approach to moral semantics, has to face accusations of relativism: objective metatheoretical criteria. Theory-Indexed Moral Contextualism allows us to draft an answer to the following questions. First, how do contextualism and faultless disagreements address the Moral Twin-Earth Problem? Second, in virtue of what do moral claims get their truth-value? Third, how does contextualism might address moral relativism?

Theory-Indexed Moral Contextualism paves the way for us to account for the following theoretical desiderata. We can characterize moral truth-makers (indexed moral theories). We can characterize moral learning (by virtue of synthetic knowledge of moral theories). We can characterize moral progress (through metatheoretical criteria used for other bodies of

moral relativism. In that sense, it could be argued that my solution doesn't help with moral relativism, it just moves to a higher level. Afterall, metatheoretical criteria might also be context sensitive. I don't have a great response to this inquiry. But I see two possible ways to build a way out. (1) I argue that metatheoretical criteria are non-normative, not context sensitive properties; or (2) I argue that my theory is a way out moral relativism but not to a more general high-order normative relativism.

knowledge). We can characterize the nature of a moral dilemma (thinking of the same moral sentence being indexed to different moral theories).

There is a sense in which moral theories are just like other scientific theories. Of course, the objects of study might be quite apart. Nevertheless, if the way theories are individuated, the way disagreements could be modeled, the ways we can interpret their changes through time, are alike, we shouldn't be surprised that the same metaphysics, semantics, and reference fixation work apply too. There are still questions that are to be answered, For instance, it still isn't clear the theoretical tolerance we should have in order to consider a set of sentences a moral theory. If the set of sentences are too disconnected from what we consider *morality* questions might arise with respect to the minimum requirements a set of sentences should have in order to qualify as a moral theory. We don't have clear rules to differentiate a bad moral theory from something that is not a moral theory at all. Nevertheless, these questions are not exclusive of our contextualist approach, and in that way, just as there is no easy way to differentiate an incredibly terrible theory of gravity from a theory of something else than gravity, we can inherit their problems and accept a more or less vague individuation of what counts as a moral theory. It's worth it in the end for all that we obtain.

Abstract

Metaethical theories that are trying to account for moral disagreement face important challenges. On the one hand, if the semantic treatment of moral terms assigns a meaning too specifically related to a contextual parameter (like culture, religion, etc.) we might be ruling out the substantiality of moral disagreements, since disagreeing parties can be both correct under their own terms. On the other hand, if our treatment of moral terms ignores their relation to a contextual parameter, we might be unable to explain the nature of the very disagreement, as we ignored how parties ended up believing different things. This M.A thesis explores the theoretical room for one particular contextualist account of the meaning of moral terms: Theory-Indexed Moral Contextualism; in such a way that is able to model the substantiality of moral disagreements in a way both compatible with non-exclusionary disagreements and with standard externalist semantics.

References:

Boyd, R. "How to Be a Moral Realist," in *Essays on Moral Realism*, edited by G. Sayre-McCord. Ithaca, NY: Cornell University Press, 1988, 181-228.

Copp, D. "Milk, Honey, and the Good Life on Moral Twin Earth." In *Morality in a Natural World: Selected essays on Metaethics*, edited by D. Copp. Cambridge: Cambridge University press, (2007): 203-229.

Eklund, M. "Carnap and Ontological Pluralism" in *Metametaphysics: New Essays of the Foundations of Ontology*. (2009): 130-166.

Eklund, M. *Choosing Normative Concepts.* (2017): Oxford: Oxford University Press

Gibbard, A. *Wise Choices, Apt Feelings: A Theory of Normative Judgment*. (1990) Cambridge: Harvard University Press.

Gibbard, A. *Thinking How to Live*. (2003): Cambridge: Harvard University Press.

Gibbard, A. *Meaning and Normativity*. (2012); Oxford: Oxford University Press

Hawthorne, J., and Yli-Vakkuri J. *Narrow Content*. (2018): Oxford: Oxford University Press

Hirsch, E. "Ontology and Alternative Languages" in *Metametaphysics: New Essays of the Foundations of Ontology*. (2009): 231-259.

Horgan, T. and Timmons, M. "New Wave Moral Realism Meets Moral Twin Earth." *Journal of Philosophical Research* 16 (1991): 447-65.

Horgan, T. and Timmons, M. "Troubles for New Wave Moral Semantics: The 'Open Question Argument' Revived." *Philosophical Papers* 21 (1992a): 153-75.

Horgan, T. and Timmons, M. "Troubles on Moral Twin Earth: Moral Queerness Revived." *Synthese* 92 (1992b): 221-60.

Horgan, T. and Timmons, M. "Analytic Moral Functionalism Meets Moral Twin Earth," in *Minds, Ethics, and Conditionals*, edited by I. Ravenscroft. Oxford: Oxford University

Press, 2009, 221-36.

Horgan T, Timmons M. “Copping out on Moral Twin Earth”. *Synthese 124(1)* (2000):139–152

Kaplan, D. “Demonstratives: an essay on the semantics, logic, metaphysics, and epistemology of demonstratives and other indexicals”, in Joseph Almog, John Perry, and Howard Wettstein (eds.), *Themes from Kaplan*, (1989): 481-564. New York: Oxford University Press.

Khoo, J. Knobe, J. “Moral Disagreements and Moral Semantics”. *NOUS* 52:1 (2018) 109–143 doi: 10.1111/nous.12151

Mackie, J. L.. *Ethics: Inventing Right and Wrong*. (1977) Harmondsworth, England: Penguin.

Olson, J. *Moral Error Theory: History, Critique, Defence*. (2014) Oxford: Oxford University Press.

Plunkett, D. and Sundell, T. “Disagreement and the Semantics of Normative and Evaluative Terms.” *Philosophers’ Imprint* 13.23 (2013): 1-37.

Putnam, H. “The Meaning of ‘Meaning’.” *Midwest Studies in the Philosophy of Science* 7 (1975): 131–93.

Shafer-Landau, Russ. "A Defense of Motivational Externalism." *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 97, no. 3 (2000): 267-91.

Vayrynen, P. “A Simple Escape from Moral Twin Earth”., *Thought* (pre-print) (2018): 1-14.

Williams, J. R. G. “Normative Reference Magnets.” *Philosophical Review* 127 (2018): 41-71.

Williamson, T. “Moral Anti-exceptionalism”. (to appear in) *The Oxford Handbook of Moral Realism,* edited by Paul Bloomfield and David Copp. (version of 2020)

Non-exclusive licence to reproduce thesis and make thesis public

I, Piero Luis Orlando Suarez Caro

(*author's name*)

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to

reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright,

Theory-Indexed Moral Contextualism_____,

(*title of thesis*)

supervised by Patrick Shirreff.

(*supervisor's name*)

2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.

4. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Piero Suarez

author's name

15/05/2021