THOUGHT
A JOURNAL OF PHILOSOPHY

Thought ISSN 2161-2234

ORIGINAL ARTICLE

# One-Person Moral Twin Earth Cases

Neil Sinhababu

National University of Singapore

This paper presents two cases demonstrating that theories allowing the environment to partially determine the content of moral concepts (such as the causal theory of reference) that provide incorrect truth-conditions for moral terms. While typical Moral Twin Earth cases seek to establish that these theories fail to account for moral disagreement, neither case here essentially involves interpersonal disagreement. Both involve a single person retaining moral beliefs despite recognizing actual or potential mismatches with the purportedly content-determining facts. This lets opponents of such theories grant objections that standard Moral Twin Earth cases fail to demonstrate disagreement, and argue more straightforwardly that they generate implausible truth-conditions for moral claims.

**Keywords** moral twin earth; causal theory; metaethics; semantics; disagreement

DOI:10.1002/tht3.400

## 1 Environmental content-determination, disagreement, and knowing the truth

Many naturalistic moral realists accept semantic theories that allow the natural, social, and historical environment to partly determine the content of moral concepts. The causal theory of reference is an example.[1] Objections to these theories involve cases where folk from Earth and Moral Twin Earth meet each other and seem to disagree about morality.[2] These theories entail that these folk talk past each other rather than disagreeing, since their different environments give their concepts different content and their terms different meanings. Recent responses to this objection employ sophisticated linguistic arguments that Moral Twin Earth cases don't demonstrate first-order moral disagreement.[3] Plunkett and Sundell (2013) argue that the disagreement is metalinguistic. Dowell (2016) offers methodological arguments that the cases aren't evidence of disagreement.

The problem with letting the environment partially determine moral content is more clearly illustrated in the consequences for moral truth and knowledge than disagreement. The causal theory, for example, entails that act-type A is right if A has the property that causally regulates the concept of rightness. In standard cases where the causal theory holds, knowing what causally regulates a concept lets us know what it applies to. By knowing that $H_2O$ causally regulates the concept of water, we can know that water is $H_2O$. But knowing what causally regulates moral concepts doesn't let us know the moral truth, suggesting that the causal theory is wrong for moral concepts. Other theories that give

*Correspondence to:* E-mail: neiladri@gmail.com

the environment a content-determining role face similar problems. This lets opponents of such theories concede that Moral Twin Earth cases do not demonstrate first-order disagreement. They can instead present cases showing that these theories make false predictions about the truth-conditions for moral claims and how we could know them.

I'll present two such cases. Both involve individual first-order moral judgments. Neither requires us to assess the presence of disagreement. Section 2 presents the first case, *Alien Nurse*. Section 3 explains why *Alien Nurse* is troublesome for the causal theory and others that give the environment a large role in content-determination. Section 4 presents and discusses the second case, *Grim Arc*.

## 2 Alien nurse

You wake up in a strange bed with a mild headache. An alien nurse says, "Greetings, astronaut! This is a historic moment. It's our first meeting with someone from another planet. We don't yet know whether you're from C-Earth or D-Earth, both of which we've observed with telescopes and drones. Both planets are nearly identical, and neither planet knows of the other. The biggest difference between them concerns the inhabitants' moral judgments, so we planned to test where visiting astronauts came from by giving them trolley problems. But it won't help much in your case. The impact from the crash might have scrambled your brain's moral judgment centers. If so, your answers could be misleading about which planet you're from. Still, I'm supposed to do it."

She reads from a script: "A bystander can save five people from being run over by a trolley by pushing a fat man off a bridge to block it. This is the only way to save the five, but it'll kill the fat man. What is your moral judgment of this action?"

You consider the scenario. While saving the most lives is the right thing to do in some cases, you can't just push an innocent bystander to his death! So it's clear that pushing the fat man off the bridge is wrong. That's what you tell the nurse.

She records your answer and says, "Philosophers and ordinary folk on D-Earth have agreed on that deontological answer. Treating others as ends in themselves causally regulates their moral concepts. Millennia of thoughtful debate shaped by causal regulation have led them to organize their society around a deontological theory. Philosophers and ordinary folk on C-Earth have agreed on the consequentialist answer that you should push the fat man. Aggregate happiness causally regulates their moral concepts. Millennia of thoughtful debate shaped by causal regulation have led them to organize their society around consequentialism. Of course, we don't know whether you're from D-Earth with intact moral judgments, or from C-Earth with moral judgments scrambled by your injury. We'll know as soon as the forensics team figures out which planet your spaceship came from. I'll check how their investigation is going." She leaves.

Soon she returns. "The forensics team says you're from C-Earth! I guess this means that the crash scrambled your moral judgment centers. Anyway, we'll care for you until you recover. Ring the bell if you need anything."

You thank the nurse as she leaves. Your thoughts return to the trolley problem. You think to yourself: "It really seems like pushing the fat man off the bridge is wrong! But

 

aggregate happiness causally regulates moral concepts in my linguistic community. My contrary intuitions merely result from brain damage. Do the facts about causal regulation settle things so that it's right to push the fat man?"[4]

## 3 Problems with letting the environment determine content

I take it that the natural answer here is: no, the facts about causal regulation don't settle the moral issue. Learning that I'm from the linguistic community where aggregate happiness causally regulates moral concepts intuitively doesn't settle moral questions in favor of consequentialism. If this is true, the causal theory is wrong for moral concepts. This section argues that *Alien Nurse* indeed demonstrates this, and extends the counterexample to other theories.

On the causal theory's account of concept-individuation, *Alien Nurse* doesn't let the moral concepts of the different planets fuse into one concept shared across linguistic communities. The planets are sufficiently isolated to prevent concepts from having previously fused. The nurse doesn't introduce any moral concepts distinctive to either planet. She uses "moral" only in a broad sense that applies to concepts on both planets, just as "scientific" applies to concepts of both $H_2O$ and XYZ. Her question about how you'd judge pushing the fat man is neutral between C-Earth and D-Earth moral concepts, and is designed to elicit responses in terms of either.

Cases like *Alien Nurse* are good for testing whether the causal theory is right for particular types of concepts. For a similar case with natural kind concepts, change the planets to Earth and Twin Earth from Putnam's (1975) classic example, and change the question from a trolley problem to the scientific question of whether water is $H_2O$ or XYZ. If I had initially said that water is $H_2O$, but then learned that brain damage had scrambled my scientific beliefs and XYZ causally regulated the use of "water" in my linguistic community, I'd accept that water is XYZ. Proper names, also widely regarded as a good case for the causal theory, behave similarly in *Alien Nurse* cases. Suppose the nurse shows me pictures of a tall person and a short person and asks me which one is my friend Ben. If I initially think Ben is the short person, but then learn that the tall person causally regulates my use of "Ben" and that brain damage has changed my memory of Ben's height, I'll accept that Ben is the tall person. *Alien Nurse* shows that learning the causal-regulatory facts does not similarly settle moral questions. Those who apply the causal theory to moral concepts face the challenge of explaining this difference.

*Alien Nurse* avoids difficulties in assessing disagreement. While standard Moral Twin Earth cases elicit metalinguistic judgments about disagreement and meaning, *Alien Nurse* follows Putnam by eliciting first-order judgments. Do these judgments have equal probative value? Dowell writes:

> No. An important difference between Putnam's thought experiment and the Moral Twin Earth thought experiment is that, while the former are to trigger first-order judgments that *deploy* the targeted term ("water") and are about its extension (about water), the latter are to trigger judgments about whether disagreement is possible

> when two speakers use different terms in different languages and about whether those terms differ in meaning. The crucial difference is that, while the former is expressed *using* the targeted term to characterize its referent, Judgment about Meaning is a semantic judgment *about* the targeted term. Likewise, Smith's Judgment about Disagreement is not a first-order judgment that deploys our moral terms—for example, a judgment about which acts are right. (10)

She argues that Moral Twin Earth judgments are less reliable because detecting disagreement with a "rival, hypothetical language" (11) requires more metasemantic competence than ordinary speakers can be expected to have. *Alien Nurse* lets opponents of the causal theory concede Dowell's point, and argue instead from first-order use of moral terms.

Slightly modifying *Alien Nurse* produces an objection to the stabilizing function account Dowell draws from Millikan (1987, 2010). This account treats the content of a concept as what it covaries with when serving the function (perhaps biological or cultural) that explains its persistence. So instead of explaining what causally regulates moral concepts on the two planets, the nurse might explain what they covaried with to serve the function explaining their persistence, and then reveal which planet you're from. But this historical knowledge doesn't settle whether pushing the fat man is right. You're free to think of your concept as delivered to you by a history of error, and the other planet as receiving its from a history of truth.

Similarities between *Alien Nurse* and Putnam's case prevent stabilizing function theorists from blocking this objection by denying that metasemantic intuitions are good evidence. As Dowell notes, the stabilizing function account has "water" referring only to $H_2O$ and not XYZ. Since this agrees with our metasemantic intuitions, the stabilizing function account treats intuition as right about Putnam's case.[5] Dowell thinks intuition goes awry in standard Moral Twin Earth cases because they require assessing disagreement with rival hypothetical speech communities. *Alien Nurse* and Putnam's case don't require assessing such disagreement, requiring only first-order judgments. If intuition is accurate about Putnam's case, it should also be accurate about *Alien Nurse*.[6] This is a problem for the strategy of debunking *Alien Nurse* intuitions by questioning our metasemantic competence. Any semantic theory providing the result that "water" refers only to $H_2O$ and not XYZ treats intuition as right about Putnam's case. Then it's hard to see why intuition would get the similarly first-order *Alien Nurse* wrong.

Dunaway and McPherson (2016) note the feature of the stabilizing function account and the causal theory that give them problems with *Alien Nurse*. They allow the environment too large a role in determining the content of moral concepts. Then if we have the minimal level of metasemantic competence that lets us get Putnam's case right, changing our beliefs about our environment should change our moral beliefs. *Alien Nurse* shows that changes in belief about these purportedly content-determining features don't in fact have this effect.

Distinguishing moral judgments from all-things-considered judgments of what to do might help with standard Moral Twin Earth cases, but it doesn't help with *Alien Nurse*.[7] This distinction helps causal theorists satisfy the intuition of disagreement in standard

Moral Twin Earth cases by saying that it's disagreement about what to do, not moral disagreement. In *Alien Nurse*, the distinction lets our astronaut think: "Given the facts about causal regulation, it's morally right to push the fat man. But then I guess I do not care that much about morality. I'm still against pushing the fat man." The problem with this thought is that it involves giving up on one's moral beliefs too quickly. The facts about causal regulation simply don't motivate accepting that it's morally right to push the fat man, even if one softens the blow by saying that one wouldn't actually push him. Merli and Copp's distinction is designed to accommodate intuitions that standard Moral Twin Earth cases involve disagreement of some kind. But it doesn't seem to change the answer to the nurse's question specifically eliciting moral judgment.

*Alien Nurse* can be modified to generate stronger intuitions. One might replace C-Earth with K-Earth where "right" is causally regulated so that it applies to killing strangers. Learning that you're from K-Earth doesn't settle moral questions in this theory's favor! If you didn't get the intuition initially (perhaps you're a utilitarian and it's easy for you to think that C-Earth has it right) considering an awful moral theory may help.

Problems with giving the environment such a big role in determining content can be illustrated in how easily this lets awful moral theories be true. *Grim Arc* is such an illustration.

## 4 Grim arc[8]

You are an educated person on a planet like ours. While studying philosophy, you learned about consequentialist and deontological theories, each of which seemed to get at part of the moral truth. While studying history, you learned that properties at the level of gender, class, and race had significantly influenced moral judgment over millennia. Other societies within the broad linguistic community of your planet had accepted hierarchical class and gender norms, and valorized conquest, enslavement, and genocide of other races. Remnants of these anti-egalitarian norms still lingered in your society's folk moral beliefs. You were optimistic that they would eventually be revised away. Future folk morality would then coincide with the philosophers' values: happiness for all creatures and respect for rational agents.

Your optimism was shaken by disturbing events. Politicians gained approval in your society and won election to its highest offices by proudly expressing sexist, classist, and racist values. Many of their influential supporters wanted to revise folk morality in favor of these values. They worked to entrench the old anti-egalitarian influences, even against values of happiness for all creatures and respect for rational agents. If their favored revisions succeeded, folk morality would favor the subjection of women, deference to the wealthy, and the glory of a master race.

You were forced to consider a grim future possibility. What if the long-run causal-regulatory influences on moral concepts were as your enemies hoped? What if the popularity of moral theories concerned with happiness for all creatures and respect for rational agency in recent centuries was merely a contingent historical aberration?

What if gender, class, and racial properties were the strongest causal regulators of moral concepts across all of time? Would sexism, classism, and racism then be right?

I hope you'll agree that the answer is no. The causal theory says yes. It entails that "water is XYZ" is true if XYZ causally regulates the concept of water, and that "sexism, classism, and racism are right" is true if sexism, classism, and racism causally regulate the concept of rightness in our linguistic community. I do not believe that either concept is causally regulated in such a way. If proven wrong on both counts, I will start believing that water is XYZ. I will not start believing that sexism, classism, and racism are right.

The stabilizing function account seems to have the same bad result. Moral concepts could have the cultural function of perpetuating the oppressive hierarchies in which they're used, through their covariance with gender, class, and racial properties. The concept of wrongness might have a cultural function of condemning those who resist these oppressive hierarchies. If it covaries with resisting oppression, the stabilizing function account will treat resisting oppression as wrong.

*Grim Arc* is troublesome for the connectedness model of Schroeter and Schroeter (2014), on which content is determined by the "whole set of attitudes, dispositions, social practices, and environmental feedback loops associated with the historically and socially extended representational tradition" (14). The scenario envisioned in *Grim Arc* would allow features of the social environment like the racist attitudes, classist dispositions, and sexist practices of others in our language community to determine the content of moral concepts. Early causal theorists were optimistic about the causal influence of moral properties and the arc of the moral universe.[9] They wrote in times of progress, when the fall of apartheid and communism made grim possibilities for the long-run causal regulation of moral concepts less salient.

The time has come to consider these grim possibilities. Doing so reveals that the causal theory, the stabilizing function account, and the connectedness model allow a dystopian future to shape the moral truth in its own image. Moral concepts must let us convey the horror of such a future, rather than falling under its control.[10]

## Notes

1 Brink (1989), Sayre-McCord (1988), and Boyd (1988) accept the causal theory, which helps them address Moore's (1903) Open Question Argument. Sections 3 and 4 discuss other theories facing similar problems.

2 Horgan and Timmons (1993), Rubin (2008).

3 See also Merli (2002).

4 "Wrong" here appears within one person's thoughts, disconnecting the uncertainty from the sort of metalinguistic negotiation invoked by Plunkett and Sundell (2013).

5 Dowell cites Millikan's explanation of why on her theory "as on Putnam's own, XYZ was not within the extension of our word 'water' even in 1750" (23). As Dowell notes, this explanation doesn't appeal to ordinary speakers' intuitions. But accurate intuitions need not explain extension — they merely need to correlate with extension. If Millikan's account says that XYZ isn't in the extension and intuition agrees, Millikan has intuition being accurate about this case.

6 Perhaps semantic intuitions about novel environments are less accurate. But if intuitions about Putnam's case are accurate, *Alien Nurse* intuitions should be as well.
7 Merli (2002) and Copp (2000) suggest this distinction.
8 The name is inspired by King's (1964) remark that "The arc of the moral universe is long, but it bends toward justice." I regard him as making a plausible prediction about contingent future events, not as stating a metaphysical necessity connecting justice with long-run political outcomes.
9 Sayre-McCord (1988) and Brink (1989) see the injustice of apartheid as causing observers to recognize its injustice, which then caused protest against the South African apartheid government that led to its collapse.
10 I thank Michael Rubin, Jon Keyzer, and my colleagues in the NUS Philosophy Reading Group for suggestions that improved this paper.

## References

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

Brink, David. *Moral Realism and the Foundations of Ethics*. Cambridge: Cambridge University Press, 1989.

Copp, David. "Milk, Honey, and the Good Life on Moral Twin Earth." *Synthese* **124** (2000): 113–137.

Dowell, Janice. "The Metaethical Insignificance of Moral Twin Earth," in *Oxford Studies in Metaethics*, edited by R. Shafer-Landau. Oxford: Oxford University Press, 2016, 1–27.

Dunaway, Billy and Tristram McPherson. "Reference Magnetism as a Solution to the Moral Twin Earth Problem." *Ergo* **3**.25 (2016): 639–679.

Horgan, Terry and Mark Timmons. "New Wave Moral Realism Meets Moral Twin Earth," in *Rationality, Morality, and Self-Interest*, edited by J. Heil. Rowman and Littlefield. Lanham, MD, 1993, 115–133.

King, Martin Luther (1964). "*Wesleyan Baccalaureate Address*".

Merli, David. "Return to Moral Twin Earth." *Canadian Journal of Philosophy* **32**.2 (2002): 207–240.

Millikan, Ruth. *Language, Thought, and Other Biological Categories*. Cambridge, MA: MIT Press, 1987.

Millikan, Ruth. "On Knowing the Meaning; with a Coda on Swampman." *Mind* **119** (2010): 43–81.

Moore, G. E. *Principia Ethica*. Cambridge: Cambridge University Press, 1903.

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

Putnam, Hilary. "The Meaning of 'Meaning'," in *Mind, Language and Reality*, edited by H. Putnam. Cambridge: Cambridge University Press, 1975.

Rubin, Michael. "Sound Intuitions on Moral Twin Earth." *Philosophical Studies* **139** (2008): 307–327.

Sayre-McCord, Geoffrey. "Moral Theory and Explanatory Impotence," in *Essays on Moral Realism*, edited by G. Sayre-McCord. Ithaca, NY: Cornell University Press, 1988, 256–281.

Schroeter, Laura and François Schroeter. "Normative Concepts: A Connectedness Model." *Philosophers Imprint* **14** (2014): 1–26.