1 Preliminaries

This paper motivates and explores a new epistemological argument against realism about normativity.

Moral realists—and, by extension, realists about normativity more broadly—are often accused of lacking the resources to provide an acceptable epistemology. These allegations take different forms, but examples include the following:

- Harman’s (1986) claim that moral facts do not causally explain why we form moral judgments;
- Street’s (2006) argument that it is compatible with a naturalistic evolutionary process that we make moral judgments according to any of a wide variety of mutually incompatible moral systems;
- Mackie’s (1977) “argument from queerness” which (in one form) claims that moral properties are massively different in kind from any other property we know about.

We could easily add additional examples to the list. But doing so would only make a longer list with a common theme: none of them are very compelling.

In summary the reason is that each argument relies on a characterization of moral realism, or an epistemological principle concerning what would constitute a successful realist epistemology, that the realist can easily reject. For example contra Harman, mathematical knowledge can provide a model of a domain in which causal interaction is not necessary for knowledge. And contra Street, epistemologists are generally in agreement that the bare possibility of forming false beliefs in your situation does not impugn the epistemic status of your actual beliefs.

The aim of this paper is not to repeat these criticisms and replies here. Rather it is to explore the prospects for new ways of developing epistemological problems.

---

1 See Dworkin’s (1996) discussion of the realist as assuming that a quasi-physical quantity called “morons” explain our capacity to form moral judgments.
2 Clarke-Doane (2012)
3 Dunaway (forthcomingb)
for realism from secure epistemological and meta-ethical starting points. The basic contours for such an argument should run as follows: it should begin from a central commitment of realism about the normative, which I take to be a commitment to a high degree of fundamentality in normative properties.

A compelling epistemological argument against realism should show that this aspect of realism about normative properties entails that an uncontroversial necessary condition on knowledge (and other epistemically desirable states) is violated. I will take this to involve normative beliefs being at risk of error, in a suitably refined sense.

The argument I will explore claims that the relevant anti-risk principles are violated because realism is committed to a meta-semantics about fundamental normative terms that gives rise to the possibility of false normative beliefs. That is: a basic story about how normative terms refer such as ‘ought’ refer to normative properties will imply, in conjunction with the realist’s metaphysics, that normative terms are what I will call *moderately semantically stable*. ‘Ought’ and similar terms are stable, because small changes in how a community uses the term does not change what the term refers to. But stability only goes so far: sometimes a community might change how they use a term, and the change *does* produce a reference-shift. Moreover when speakers are at risk of such a reference-shift, they aren’t guaranteed to know that they are. And in such cases they will also have beliefs that put actual true beliefs at risk of error in an epistemically relevant way.

I will call this the **Argument from Risk**.

The Argument from Risk is supposed to be a perfectly general argument, showing that every normative belief is subject to epistemically problematic risk. For illustration here is one application of the argument. Suppose as a matter of fact, what we ought to do is give 10% of our income to charity. Speakers in a community might apply ‘ought’ to only actions of giving slightly less—say, 8 or 9%—but they would not speak truly by denying that one ought not to give 10% with the expression ‘it is not the case that one ought to give 10% of one’s income to charity’. This is the semantic stability of ‘ought’. But other possible communities use ‘ought’ *very* differently, and it is highly plausible that they are talking about something other than obligation. For instance, they might systematically use their ‘ought’ to apply only to actions which are psychologically feasible for typical humans to perform. For these communities, ‘ought’ refers to a property that is instantiated by the action of giving 2 or 3% of one’s income to charity, but not giving 10%. They do not refer to obligation.

In such a circumstance, one would not be aware of the semantic shift in the reference of ‘ought’. So would have some false normative beliefs. The Argument from Risk claims that the risk of being in such a circumstance impugns the epistemic status of one’s actual normative beliefs.

Since this argument that targets realism on the basis of the definitive features of realism, it represents a powerful challenge. Realists cannot resist it by rejecting the metaphysical or epistemic premises it rests on, since these are premises that
realists are universally committed to. Existing epistemological arguments against realism cannot claim these virtues.

The Argument from Risk is a much more formidable challenge to realism than existing epistemological arguments. But in the closing sections I will show that it is not a knock-down argument against realism. A decisive epistemological argument against realism would show that realism entails normative skepticism—the conclusion that we cannot have knowledge normative propositions. It is doubtful, however, that the Argument from Risk is sound as an argument for across-the-board normative skepticism. The true lesson from the Argument from Risk, I will argue, is that realists must accept substantial and non-trivial limitations on normative knowledge, but are not committed to skepticism.

2 Moderate Semantic Stability

What is realism about the normative? If the Argument from Risk is going to present a successful epistemic challenge for realism, it must derive epistemic difficulties from the distinctive aspects to realism—and not from features it shares with its anti-realist competitors. So it is important to isolate an essential, rather than accidental, feature of realism about the normative.

I will tackle this question in two stages. First I will outline what I take to be some plausible constraints semantics of normative terms for realism. Then I will spell out the most plausible metaphysical commitments for realism that explain the semantic profile of normative terms for realists.

2.1 Disagreement and Stability

First: what are the semantic features that a realist view of the normative will need to accommodate?

There are two ways to read this question. One is a question about the semantic value of ‘ought’ in English. Candidate answers include: it is the property of maximizing happiness, it is the property of respecting the autonomy of rational agents, or it is a distinct and sui generis non-natural property that has in intension that matches one of these natural properties. The semantics of English might also relativize the semantic value of ‘ought’ to features of the context in which it is used.4

Questions about semantic value will not be the primary issue in this paper. Instead there is a second, more fundamental issue: the meta-semantic question. This is the question of what explains why ‘ought’ has the semantic value it does at a world. ‘Ought’ might, as a matter of fact, refer to the property of happiness maximization. What explains why ‘ought’ refers to happiness maximization, rather than some other property, is a meta-semantic question that a bare semantics for ‘ought’ does not answer. Moreover the semantic account will not answer

4Cf. Kratzer (1977)
questions about what ‘ought’ would refer to in counterfactual scenarios where facts about linguistic usage or environment are different. I will focus below on non-normative descriptive terms like ‘red’. These are “semantically plastic” in the sense that the term can easily refer to different properties when used differently by a community. In counterfactual scenarios where reference has shifted, the semantics for the plastic terms will be different. But the meta-semantics will remain the same: the mechanisms that determine reference will be the same in the counterfactual world and actual world, but determine different semantic values on account of the difference in linguistic usage.

I will be focused on the meta-semantic project of outlining what it is in virtue of which ‘ought’ has the semantic value that it does (and I will not be much concerned with what this semantic value in fact is). Here I will briefly summarize some more-or-less conventional wisdom on what the meta-semantics for ‘ought’ has to look like for the realist. (Much of this explored in more detail in Dunaway and McPherson (2016).) The upshot is that realists need a meta-semantic account with some very distinctive features, and it is not trivial to develop a non-ad-hoc theory that accomplishes this.

Realism cannot explain the semantic profile of ‘ought’ on the basis of analytic connections to terms designating natural properties. This is the lesson of variations on Moore’s (1903) “Open Question Argument”: since questions of the form

I know a has N, but ought I to do a?

where N is any natural property, such as the property of maximizing happiness, and a is an action one can perform.

Questions of this form make sense. They do not have the air of questioning obvious logical or analytic truths. What this (plausibly) shows is that it is not analytic of ‘ought’ that it refers to any particular natural property. Competent speakers can know all of the natural properties of an action, but not know its normative status. If ‘ought’ were analytically connected to terms standing for some natural property, speakers who granted the cogency of these questions would not be competent. So the reference of ‘ought’ is not determined by the conditions for being competent with ‘ought’.5

A meta-semantics for ‘ought’ should instead be a theory of how the reference of ‘ought’ is determined, without relying entirely on analytic connections with other terms. It is clear that for many non-normative terms there must be some correct meta-semantic story of this kind. But the realist cannot simply adopt any off-the-shelf descriptivist account of what ‘ought’ refers to. She needs an account that explains some apparently distinctive semantic features of ‘ought’ and many proposals for the meta-semantics of descriptive vocabulary will fail if they are recruited as meta-semantic explanations for the normative realist.

5More OQ references
Among the features a realist meta-semantics needs to explain, the literature has largely focused on one: disagreement. Roughly, communities that use their normative terms differently will be capable of disagreeing with one another, using their own normative vocabulary, about a contested case.

Richard Hare (1952), and later in a series of papers Terence Horgan and Mark Timmons (1991; 1992a; 1992b; 1993; 1996), highlighted the extent to which communities can use their normative language very differently, and yet still intuitively disagree with each other. The range of cases include a community of missionaries encountering a cannibal community (Hare), or a community of consequentialists conversing with a community of deontologists (Horgan and Timmons). Normative disagreement in these contexts is cited as a meta-semantic phenomenon that the realist cannot explain. I will not rehearse the reasons for doubting this ambitious conclusion here. But it is worth bearing down on what, exactly, the realist meta-semantics needs to explain.

Discussions of cross-community disagreement sometimes appeal to what would happen if the two communities were to come into contact with each other (cf. the missionaries and cannibals example). If speakers from the two communities were to come into each other and have a conversation about a particular action, it is plausible that they would disagree with each other. This requires that they are both talking about the same thing, namely about whether they ought to do the action in question. Hence the term ‘ought’ in each speaker’s mouth refers to the same property, obligation. But this doesn’t show that the two communities refer to the same property in their own languages. Once speakers come into contact with each other, they plausibly intend to speak about the same thing the other is speaking about. Each community on its own does not intend to coordinate with the other in this way. But once speakers come into contact with each other, new intentions are present, and it is hard to describe the case in a way where each speaker determinately refers to the same property as her original community, and does not co-refer with her interlocutor in virtue of her new intentions.

Some ways of motivating cross-community disagreement without encounters of this kind fail for other reasons. Horgan and Timmons point to hypothetical communities of consequentialists and deontologists, and suggest that there is a disagreement between speakers in each community. They then suggest (but do not explicitly argue) that similar cases can be constructed with other communities of speakers. What is important here is that the motivating thought for the claim that the communities refer to the same property relies on a notion of cross-world disagreement. That is, the consequentialist community says of an action \( a \) that is optimific but requires coercion of an autonomous agent, ‘one ought to do \( a \)’, while the deontologist community says ‘one ought not to do \( a \)’. We are told that these utterances, made in different circumstances, disagree with each other.

---

6See Dunaway and McPherson (2016).
7See Dunaway and McPherson (2016) for a reconstruction of this argument.
There are several problems with using this notion of cross-world disagreement to motivate the claim that both communities refer to the same property. First, the relations the two hypothetical communities bear to each other are easily confused with relations consequentialists and deontologists bear to each other in the actual world. We all know consequentialists and deontologists that are both a part of our own linguistic community. They speak a common language (English), and disagree with each other. But this isn’t cross-world disagreement. And it is easy to confuse the relationships these actual speakers bear to each other with the relations entirely separate hypothetical communities would bear to each other.

Moreover it is not clear how to make sense out of the notion of cross-world disagreement. Some natural ways of understanding the notion are flawed. For instance one might think that the cross-world disagreement here amounts to the fact that if speakers from the two communities were to encounter one another, they would disagree with one another. But this won’t show that there is a property which is such that each community in its original environment refers to it, since we have seen that once speakers are put in context with one another they might acquire intentions which shift the reference of one (or both) of their terms to ensure that they are not talking past one another. Nor can we understand cross-world disagreement purely in terms of inconsistency between what one community says with their sentence ‘one ought to do a’ and what the other community says with ‘one ought not to do a’. There are many inconsistent thoughts that these communities accept which do not count as cross-world disagreement. For instance the first community might accept the sentence ‘there are consequentialists’ (it would be reasonable for them to accept this, since ex hypopthesi they are consequentialists). The second community might accept the sentence ‘there are no consequentialists’ (again this is reasonable in some worlds since this community contains deontologists, and there need not be any evidence of consequentialists in the vicinity). But the communities don’t disagree over whether there are consequentialists in any interesting sense.8

This does not mean that there is nothing for a realist meta-semantics to explain. Rather there is a very natural property of normative language that is very close to the alleged disagreement data that any plausible realist theory should account for. This is the contrast with semantically plastic terms, which change their reference easily with relatively little change in how a community uses the terms in question. Color-terms are paradigms of plasticity. For example we apply ‘red’ to surfaces that reflect light wavelengths between 620 and 750 nm. (This is the conventional demarcation of the red spectrum.9) And we thereby refer to a property that has objects with the relevant surfaces in its extension. But another community could easily apply ‘red’ to surfaces that reflect light wavelengths between 600 and 730 nm. (Thus they ‘red’ to some colors in the red-orange range as well.)

8Thanks to John Hawthorne for discussion. For more discussion of potential problems with disagreement as a diagnostic here, see DOWELL CITE
This isn’t a case where both communities are talking about redness and one makes systematic mistakes about its distribution. Rather both communities speak different languages, where ‘red’ refers to different properties in their respective mouths. ‘Red’ is semantically plastic.

‘Ought’ by contrast is not plastic in this way. We are not disposed to treat another community that uses ‘ought’ differently as speaking truly, in their own language (where in their language ‘ought’ refers to some other property other than obligation). Rather the natural diagnosis, when another community deviates in their usage of ‘ought’ (perhaps by not taking whether coercion of rational agents is required to bring about an optimific outcome into account) we treat them as making false judgments about obligation. When they apply ‘ought’ to an optimific act that requires taking the life savings of a small handful of moderately well-off people against their wishes, we don’t think they have a true normative belief about a property closely related to, but distinct from, obligation. Rather we think, when they make the relevant judgment, that the falsely believe that they ought to take the money for the greater good. This is a symptom of a degree of semantic stability in ‘ought’.

Semantic stability is a deigned phenomenon. That is: terms can be semantically stable to varying degrees, and there is no guarantee that most terms will be as unstable (or “plastic”) as ‘red’ is. What is distinctive of ‘ought’ and some other normative vocabulary is that it is highly semantically stable. This means that there are many ways for a community to use their term ‘ought’ in different ways than we do, without changing what they are talking about. They succeed in referring to obligation (the same property we refer to) even though they use their terms differently. Their usage of ‘ought’ would have to substantially differ from ours to constitute a usage pattern that refers to a different property; roughly we can think of the degree to which usage would have to differ to generate a change in reference as corresponding to the degree to which ‘ought’ is semantically stable.

One primary aim for a meta-semantics for ‘ought’, I will assume, is to explain the high degree to which ‘ought’ is semantically stable. If realism is correct, then there must be some explanation for this meta-semantic phenomenon. I will sketch in the next subsection what I take to be the most promising account of the data. But it is worth emphasizing here that the Argument from Risk does not depend on the details of this account; any implementation of realism that explains the semantic stability of ‘ought’ successfully will face the challenge.

2.2 The Metaphysics and Meta-semantics of Realism

If the Argument from Risk is going to present a successful epistemic challenge for realism, it must derive epistemic difficulties from the distinctive aspects to realism—and not from features it shares with its anti-realist competitors. So it is

\[10\]For related discussion see WILLIAMS CITE, EKLUND CITE, Dorr and Hawthorne (2014), and Schoenfield (forthcoming).
important to isolate an essential, rather than accidental, feature of realism about the normative.

Whatever this central feature of realism is, it will have to be a part of the explanation of the high degree of semantic stability in ‘ought’. The best account of this appeals to the high degree of metaphysical eliteness of obligation. There are two components to this account. The first explains why realism about obligation should be thought of as a thesis about the metaphysical status of the property. The second is a meta-semantic account that explains the semantic stability of ‘ought’ in terms of the metaphysical eliteness of obligation.

Realist views of the normative include versions of Moore’s (1903) non-naturalism, as well as the naturalist views found in Railton (1986) and Schroeder (2007). According to the non-naturalist view, obligation is a sui generis part of reality, that cannot be reduced to or identified with any natural property. Naturalist views hold that the normative is continuous with (identical to, or reducible to) natural properties. The properties in question are fairly simple, non-gerrymandered natural properties that are built out of components that play a role in natural-scientific explanations: they include being desired by one’s fully informed self (Railton), or promoting one of the desires one actually has (Schroeder).

I take most promising account of what unities these different realist views of the normative with each other, and with realism about other non-normative domains, to hold that realist views entail that the relevant properties are highly elite. (Dunaway (MS), Dunaway and McPherson (2016)) ‘Elite’ here is to be understood as a metaphysical feature of properties in the family of ‘naturalness’ in the sense of David Lewis (1983, 1984). Roughly this means that the properties in question are not very gerrymandered (they are more like being metallic than being either prime or located in Mongolia). And they will be properties that play explanatory roles some legitimate theoretical discipline—whether it be physics, genetics, social science, or metaphysics.

Eliteness is a degreed notion. Some properties are fully elite, such as (perhaps) mass and charge, or parthood. Other properties are not fully elite, but close to it: perhaps properties such as being a gene or chemical bond are like this. And others will be not very elite at all: astrological properties such as being a virgo, or highly disjunctive properties such as being either prime or located in Mongolia are like this. Further I will assume (unlike Lewis) that degrees of eliteness cannot be reduced to the notion of perfect eliteness: Lewis held that the eliteness of a property is tied to the length of a definition of the property in terms standing for perfectly elite properties alone, but this definitional approach is implausible and I will work with an undefined notion of eliteness instead.

Suppose then that realism about obligation entails that obligation is highly

---

11I use ‘elite’ instead of ‘natural’ here to avoid confusion with the sense of ‘natural’ at issue in the naturalism/non-natural distinction, which is not Lewis’s notion.

12See Lewis (1983) for an initial presentation of this idea, and Hawthorne (2006, 2007) and Williams (2007) for discussion.
fundamental. What is the connection between this thesis and semantic stability? There is a very natural connection, given a plausible component on any theory of reference-determination: *reference magnetism*. This is the thesis that highly elite properties are easy to refer to. One doesn’t need to have fully accurate beliefs about gold, and apply ‘gold’ to all and only samples of gold, in order to successfully refer to gold, rather than some gerrymandered property that better fits one’s limited usage. Sincerely including some samples of “fool’s gold” in one’s usage of ‘gold’ doesn’t shift the referent of one’s term away from gold.

More precisely: since eliteness comes in degrees, and ease of reference comes in degrees, then a property will be easy to refer to to a degree $d$ just in case it is elite to degree $d$. Given that obligation is, for a normative realist, highly (if not fully) elite, it is a reference magnet and will be very easy to refer to. Color-terms like ‘red’ are examples of not-very-elite properties that will not be very stable. But other descriptive terms will be stable to some degree, since they refer to highly elite properties. ‘Ought’ is like these.

Reference magnetism does not imply that *every* possible use of of ‘ought’, even among those connected with motivational and attitudinal roles, will be a use that refers to obligation. Some uses are too poor of a fit with obligation, highly elite as it is. And these uses fit much better with other somewhat elite properties. A meta-semantics that includes reference magnetism entails that ‘ought’ when used in these ways by a community refers some property distinct from obligation.

Here is an example. Suppose we are a community who uses the normative ‘ought’ exclusively for actions that are best, all-things-considered, to perform. Another community might use ‘ought’ exclusively for actions that satisfy a kind of *bounded optimality* constraint: they are the best, among the actions that can feasibly be performed. In this case it is natural to interpret this community’s use of normative language not as a misguided attempt to talk about obligation, but rather as a community which uses ‘ought’ to talk about a different normative property: bounded optimality. (It is also natural to treat the community as not disagreeing with us, assuming we can make sense of the relevant notion of disagreement.) Of course each community uses a normative term ‘ought’ that is fairly stable. Slight departures from perfect usage in either community will not produce a new referent. But the fact that wholesale shift in usage from one community to the other can produce a change in referent shows that ‘ought’ is only moderately semantically stable. Reference magnetism protects against frequent semantic shifts in the reference of ‘ought’, but it does not guarantee that they are absent entirely.

To sum up: a realist view of the normative needs to explain the semantic stability

---

13 Of course ordinary English has a context-sensitive ‘ought’ that can express either of these notions in different contexts (Cf. Kratzer (1977)). I will ignore this complication here for the same of exposition, though the points made here could be accommodated within a contextualist framework.

14 Thanks to John Hawthorne for discussion of this point.
of ‘ought’. There are promising avenues for such an explanation: the realist can draw on analogies with highly stable descriptive terms, and hold that obligation is, like the referents of other stable terms, highly elite. The meta-semantic thesis of reference magnetism will imply that obligation is easy to refer to, and hence stable.

We can introduce some terminology to describe this idea succinctly. In the actual world, our use of ‘ought’ refers to obligation. Since ‘ought’ is highly stable, many nearby worlds which are almost identical to the actual world, but contain only slight differences in how we use ‘ought’, will also be worlds where ‘ought’, as used in the relevant world, refers to obligation. (According to the realist explanation I have outlined here, these are worlds where obligation is the property that scores best on the metrics of degree of eliteness and fit with how the community uses ‘ought’.) Call these stable worlds. Not every world is a stable world. In some potentially nearby worlds, usage of ‘ought’ is sufficiently different from our own to produce a world where ‘ought’, as used by the relevant community, has shifted its reference to a property distinct from obligation. Call these shifty worlds.

Since ‘ought’ is moderately semantically stable, there will be many stable worlds. But not every world will be stable; some worlds will be shifty worlds. In shifty worlds one’s community uses ‘ought’ in a way that produces a referent distinct from obligation—there is a sufficiently elite property that better fits than obligation with use of ‘ought’. There are different ways for usage in a shifty world to constitute a usage on which ‘ought’ does not refer to obligation. But in some shifty worlds one won’t know that one’s community’s use has undergone a change, or won’t know that the changed usage constitutes a shift. The next section sets the groundwork for worlds like these to constitute epistemological problems for normative beliefs in the actual world. That is: even if we aren’t actually in a shifty world, the threat of being in one could prove devastating for the epistemological status of normative beliefs in stable worlds. This will allow us to focus on the Argument from Risk in greater detail.

3 Risk

The Argument from Risk aims to turn the moderate semantic stability of ‘ought’ into an epistemological problem for realism. Success would involve showing that, according to realism, no one knows, or has justified beliefs about normative propositions. One desideratum for a convincing epistemological argument against realism is that it should avoid relying on a controversial principles about what knowledge or justification are. So I will focus here on a simple and commonly accepted necessary condition on knowledge and epistemic justification. This is an anti-risk principle: roughly, a belief cannot be knowledge if it is at risk of being false. This principle needs refinement, but the upshot is clear: if realism about the normative is plausible, it then among its consequences must be the epistemic fact
that normative beliefs are not at risk of being false. Otherwise realism entails that there is no normative knowledge.\(^{15}\)

The moderate semantic stability of normative terms threatens to put normative beliefs at risk of being false. So the core tenets normative realism appear to entail that normative beliefs cannot satisfy some central conditions on knowledge or justification. This is the phenomenon that the Argument from Risk

The relevant notion of \textit{risk} is to be cashed out in terms of what goes on in nearby worlds, or worlds that could easily have obtained.\(^{16}\) When one risks dropping one’s phone in a pool by standing on the edge of the pool and tossing the phone in the air, this amounts to there being a nearby world where one tosses the phone in the air and it falls into the pool. Likewise when a belief is at risk of being false, this amounts to there being a nearby world in which the belief is false.

Call the nearby false belief that puts one’s actual belief at risk and prevents it from being knowledge a \textit{bad counterpart}.\(^{17}\) Beliefs that are not knowledge, by definition, have bad counterparts. This picture illuminates the ways in which mere guesses, beliefs based on the testimony of serial liars, and Gettierized beliefs are not knowledge. A few notes on which beliefs are bad counterparts are in order.

First, a belief has a bad counterpart only if there are \textit{similar} beliefs that are false in nearby worlds. I can know that I had breakfast this morning even if there is a nearby world where I misremember the name of a new acquaintance. But bad counterparts do not need to be \textit{identical} to the beliefs they are bad counterparts for. If one is guessing at the answer to questions about the sums of moderately large numbers, then one’s correct guesses won’t have nearby worlds where the same belief is false. If one correctly guesses that \(634 + 399 = 1033\), then one has a true belief, and moreover this very belief is not false in any nearby world (in all nearby worlds, \(634 + 399 = 1033\)). So one can’t have a false belief that \(634 + 399 = 1033\) in any nearby world.

But correctly guessing doesn’t bring knowledge. If one is guessing at the relevant sums, then even if one actually gets the answer right, there is a nearby world where one instead comes to believe (for instance) that that \(634 + 399 = 893\). This belief is sufficiently similar to one’s actual belief, and can serve as a bad counterpart for it.

Here is a second qualification: not all similar false beliefs are bad counterparts. Some nearby similar false beliefs are arrived at in a suitably different way, and thereby do not threaten the epistemic status of one’s actual beliefs. Suppose I

\(^{15}\)What about epistemic justification? I assume that one way to lose justification for a belief is to learn that it could easily have been false in a sense that is incompatible with knowledge. So if knowledge is absent because of the presence of a kind of objectionable risk, one will lose justification for a normative belief when one learns that the relevant kind of risk is present.

\(^{16}\)Cf. “safety” principles in Sosa (1999), Williamson (2000), and Pritchard (2004). Here it is important to emphasize a point that Williamson makes, which is that the relevant sense in which a world “could easily” have obtained may ultimately be understood in partially epistemic terms. The anti-risk principle should not be understood as ipso fact offering a reductive analysis of (a component of) knowledge.

\(^{17}\)Cf. Dunaway and Hawthorne (forthcoming)
see you walking past my office door, and come on that basis to believe that you are in town. I can know that you are in town. But I could easily have believed that you are not in town: suppose in addition that I know that you live in a different city, and I have no other evidence that you are in town. I could easily have not looked out my office door at the moment you walked by (or I could easily have left the door closed), and believed on the basis of my knowledge our your typical place of residence that you are not in town. This is a nearby false belief. But it is not a bad companion in the present sense: it doesn’t destroy the status of your actual true belief as knowledge.18 The reason is that nearby beliefs aren’t candidates for bad companionship if they are formed by very different token processes. Here my actual true belief is formed on the basis of my seeing you walk by my door, whereas the would-be bad companion is formed on the basis of general knowledge about your usual residence. The difference in the causal process that produces the nearby belief prevents it from threatening the status of my very near belief.

So beliefs that have bad companions are at risk of being false in an epistemically relevant sense. They are not knowledge. This is a fairly uncontroversial anti-risk principle in epistemology, supplemented with a few tweaks to accommodate the finer details of the kind of risk that is epistemically relevant.

We can now spell out the idea that the moderate semantic stability of ‘ought’ subjects normative beliefs to the relevant kind of risk. The argument begins with the semantic stability of ‘ought’, and concludes that normative beliefs are not knowledge because they have bad companions. There are two crucial premises which are needed to connect the dots; these are premises 2 and 3 in the Argument from Risk:

**Argument from Risk**

1 ‘Ought’ is moderately semantically stable.

2 If ‘ought’ is moderately semantically stable, then one could easily be in a world where ‘ought’ doesn’t refer to obligation.

3 If one could easily be in a world where ‘ought’ doesn’t refer to obligation, then one could easily have a false normative belief.

4 If one could easily have had a false normative belief, then one’s actual normative beliefs have bad companions, and are not knowledge.

Here is an illustration of how the Argument from Risk applies to particular normative beliefs that we have used to illustrate the the degree to which ‘ought’ is semantically stable. Suppose, as before, that our community uses ‘ought’ to refer to obligation, and moreover that among the obligatory actions are the act of giving 10% of one’s annual income to charity. And suppose that the linguistic situation

is one where our use of ‘ought’ shifts in a nearby world to produce a referent for ‘ought’ that does not have giving away 10% of one’s income in its extension: we could easily have applied it only to actions that are psychologically feasible for typical agents to perform. The community in effect requires “obligatory” actions to be those normal agents could perform without significant emotional distress, and in the absence of external coercion or bribery. And let’s suppose a typical agent could not feasibly give 10% of her income in her situation. (If that sounds implausible, we could find another area where morality appears to be very demanding. I will stick with this simplified example for illustration.)

In this example ‘ought’ could easily refer to the property that has the intersection of obligatory and psychologically feasible acts in its extension. Call this property obligation*. Worlds where ‘ought’ refers to obligation* are “shifty” worlds. In shifty worlds, it needn’t be that one’s own usage differs in any way from one’s usage in worlds where ‘ought’ refers to obligation. One might continue to apply ‘ought’ to the act of giving 10% but nonetheless unknowingly be using a term that refers to obligation*, which does not have giving 10% in its extension. One will then come to believe, in the shifty world, that giving 10% of one’s income to charity instantiates obligation*. This is a false belief—giving 10% of one’s income is, by definition, obligatory but not obligatory*. So being at risk of a semantic shift puts one at risk of having false normative beliefs. That, at least, is the idea behind the Argument from risk.

4 Evaluating the Argument from Risk: Advantages

In this section I will outline some aspects of the Argument from Risk that mark a significant improvement over existing epistemological arguments against realism.

4.1 Similarity in content and process

The Argument from Risk is especially threatening to realism because it does not flounder on a full appreciation of the details of what is required for bad companionship. The argument targets beliefs that are at risk of being false owing to a nearby shifty world where ‘ought’ refers to a property distinct from obligation. There is no straightforward way to claim that these worlds do not contain bad companions for our normative beliefs on the grounds that the false beliefs formed in these worlds do not have one of the necessary properties for bad companionship.

Take the requirement that the would-be bad companions be similar beliefs. In a shifty world, the beliefs formed about the property ‘ought’ refers to in those worlds will still be normative beliefs. They will be connected to motivational states, attitudes of blame, praise and guilt, etc. in the same way actual normative beliefs are. (Granted, they are not identical beliefs because by stipulation ‘ought’ refers to a property distinct from obligation in shifty worlds. But we have seen that identity of content is not necessary for bad companionship.)
Or take the requirement that would-be bad companions must be formed by a sufficiently similar process. Other arguments that highlight the possibility of false moral belief do often ignore this. For instance someone might witness a murder and come to believe on this basis that it is wrong. It is true that, had the witness possessed a very different moral sensibility, the same individual, considering the same murder, could have come to believe that the murder in question is not wrong. But this conclusion couldn’t have come about by a causal process that is very similar to the original belief.

The Argument from Risk doesn’t make the same mistake. In a stable world, one’s normative belief is formed by a normal process of witnessing the murder, and forming a normative judgment on the basis of the observed facts and one’s normative sensibility. If there is a nearby shifty world where one forms a false belief, it is nonetheless a belief formed by a nearly identical process. The event one witnesses, and the moral sensibility influencing the judgment are unchanged in the shifty world. The Argument from Risk cannot be accused, in general, of mistaking false beliefs that are formed by very different processes for bad companions for our actual normative beliefs.

4.2 Avoiding overgeration: plasticity and other stable terms

Another virtue to the Argument from Risk is that it does not rely on premises which overgenerate: if analogous premises could establish skepticism for any domain, then the skeptical threat to the realist about normativity would be minimal. But for non-normative terminology there is good reason to think that bad companions for the relevant beliefs will not be inevitable byproducts of semantic shift.

Some terms—like ‘red’ and other color terms—admit of semantic shifts more easily than normative terms. These are the semantically “plastic” terms. One might think that this means that risks of false beliefs are more serious for plastic terms, since the semantic shifts for color-terms are easier to come by. But in fact plastic terms have other features of use that safeguard against errors produced by reference shifts.

Suppose our community consistently applies ‘red’ to surfaces that reflect light wavelengths between 620 and 750 nm. (This is the conventional demarcation of the red spectrum.\textsuperscript{19}) We could easily have used ‘red’ differently: for example, by consistently applying it to surfaces that reflect wavelengths between 600 and 730 nm. In a world where the reference of ‘red’ shifts to refer to a property that applies surfaces that reflect wavelengths between 600 and 730 nm, one would form a false belief if one uses the term ‘red’ to form a belief about an object that reflects wavelengths of 745 nm.

However if one is using ‘red’ properly one won’t be forming false beliefs in this way. If one discovers that one’s linguistic community regularly applies ‘red’

\textsuperscript{19}https://en.wikipedia.org/wiki/Visible_spectrum
to surfaces that reflect wavelengths between 600 and 730 nm (and hence applies it to some red-orange shades), the correct response is to modify one’s own usage to match the community’s. One won’t insist that since redness “really” is the property of reflecting wavelengths between between 620 and 750 nm, all of one’s linguistic peers are mistaken and ought to reform their usage. This is sensitivity to one’s community’s usage, and semantically plastic terms display it.

By contrast, if one discovers that one’s community uses ‘ought’ without regard for which actions require manipulation or coercion, but instead uniformly applies the term to optimific actions, one will not on this basis alone change one’s usage, and apply ‘ought’ only to optimific actions. Instead one will think that one’s community is forming some false beliefs about what ought to be done (for instance believing that some morally problematic manipulation is obligatory) and should revise their usage of ‘ought’ to better track the morally relevant property that it in fact refers to.²⁰

Perhaps other semantically stable terms will be vulnerable to the Argument from Risk. But here there are other differences with normative terms that mitigate against the epistemic problems that meta-semantic risk generates. Many semantically stable descriptive terms are used by scientific experts to refer to highly elite properties like mass and charge.²¹ Here the experts’ usage is tied to the highly elite properties by attention to experiment and evidence. Such experts will not use terms like ‘mass’ and ‘charge’ in nearby worlds in slightly different ways that better fit non-elite properties. Instead, while there are potential semantic shifts (where for example ‘mass’ is applied with significant accuracy to charge, and refers to it), there are no intermediate nearby worlds where experts are using ‘mass’ where mass and charge are equally bad fits with usage. With scientific terms, experts will not find themselves in a situation where they cannot know that a semantic shift has occurred.²²

5 Evaluating the Argument from Risk: Limitations

Realism implies that normative terms are moderately semantically stable. The Argument from Risk claims that this has drastic epistemological consequences: normative beliefs have bad companions, and therefore lack one of the necessary conditions on knowledge. It will be fruitful to look in detail at the premises 2 and 3 to draw some substantial conclusions about the epistemic prospects for realism.

5.1 Premise 3: reference shift to falsity

I will take these premises in reverse order. Premise 3 says:

²⁰See EKLUND CITE for more on the absence of sensitivity of this kind in ‘ought’.
²¹See Putnam (1973) for a description of the semantic division of labor along these lines.
²²What about non-experts? Insofar as non-expert usage of ‘mass’ and ‘charge’ follows expert usage, and knowledge of mass and charge depends on expert knowledge, non-experts will not be faced with epistemic problems from semantic shifts if the experts aren’t.
If one could easily be in a world where ‘ought’ doesn’t refer to obligation, then one could easily have a false normative belief.

The central idea behind this premise is that we form beliefs about a subject-matter by relying on the terms in public language we have for talking about it. Thus one will typically token beliefs about the properties which one has lexical items to refer to. If one could easily be in a shifty world, then some of one’s beliefs that depend on the reference of ‘ought’ in one’s public language will be false. This does not imply, however, that one must have false normative beliefs simply by virtue of being in a shifty world. More nuance is needed.

5.1.1 Alternative guises

In a shifty world one can’t rely on what ‘ought’ refers to in one’s language to form beliefs about obligation. But one might feasibly do this by another route. For instance, if obligation is instantiated by the actions $a_1, a_2, \ldots$, but the referent of ‘ought’ in the nearest shifty world is not instantiated by all of these actions, one might form beliefs about obligation (rather than the shifty referent) under the guise ‘the most natural property that is instantiated by action $a_1$ in circumstance $c_1$, $a_2$ in $c_2$, \ldots’. There might be other methods to cognitively grasp obligation in a shifty world as well. Call these alternative guises for obligation. Forming beliefs about obligation by using alternative guises will not, however, provide much protection against meta-semantic risk.

First, in stable worlds where one has true beliefs about obligation, one fails to have knowledge when (to a first approximation) one could easily have been in a shifty world and had a false belief. But one won’t know when one is in the shifty world, and so one won’t be able to avoid having a false belief by switching to an alternative guise in the shifty world. If one wants to use alternative guises as a buffer against meta-semantic risk, one will have to use them even in the stable worlds, where the normal guise for obligation is a perfectly good tool for forming beliefs about what one ought to do. It is unlikely that most agents in fact do this.

Second, most agents are not in a position to use an adequate alternative guise for obligation. One guise that would in principle do the trick is the alternative guise which enumerates every possible action/circumstance pair that is obligatory. But many agents do not know which possible actions are obligatory, because they haven’t formed beliefs about all of them, or because they have some false beliefs about obligation. And arguably ordinary humans don’t have the cognitive capacity to form beliefs about the (likely infinitary) list of actions that this method requires. So even if one tried to use this method for forming an alternative guise in a stable world, it would likely prevent one from forming true beliefs about obligation in the stable world, and not eliminate the risk of false beliefs in shifty worlds.

Finally, other alternative guises may themselves be subject to meta-semantic shift. For example one might attempt to use an alternative guise that characterizes
the functional role of obligation: it is a property such that there is some reason to blame agents who fail to do actions that have it, and such that there is some reason to feel guilt if one fails to do an action which has it, etc. But such guises contain terminology that will likely also refer to different properties in shifty worlds. Speakers in one’s community in shifty worlds will treat ‘obligation’ as bearing quasi-analytic ties to ‘reason’, ‘blame’, ‘guilt’ and the like. So if a world is one where ‘obligation’ refers to something other than obligation, ‘reason’, ‘blame’, ‘guilt’, and the like will also refer to something other than reasons, blame, and guilt. The alternative guise will not be one for obligation.

Relying on the meaning of terms in one’s public language is in general beneficial, allowing one to form thoughts and beliefs about much more than what one could entertain by speaking a private language with terms whose referents are determined entirely by one’s own mental activity. Perhaps it is in principle possible for some agents to form beliefs about obligation without relying on the term ‘ought’ in a public language. But such agents will need to be cognitively very special. Avoiding meta-semantic risk by not relying on public language to form beliefs about obligation will, for most of us, introduce additional epistemic defects in our normative beliefs.

5.1.2 Reference shift and false belief

When one’s normative language undergoes semantic shifts, one can have some false normative beliefs, assuming one relies on the reference of the public-language term ‘ought’ to form them. For example if one applies one’s word ‘ought’ to the act of giving 10% of one’s income to charity, then in a stable world where ‘ought’ refers to obligation, the connection between language and thought will give rise to a true normative belief, namely the belief that one ought to give 10% of one’s income to charity.

But take a shifty world where one’s linguistic community’s usage has changed in a way that produces a reference shift. For example, take the world is one where the community uses ‘ought’, with sufficient uniformity and consistency, only for actions that are psychologically feasible for typical agents to perform. The community in effect requires “obligatory” actions to be those normal agents could perform without significant emotional distress, and in the absence of external coercion or bribery. If typical agent could not feasibly give 10% of her income in such a situation, then the community in question does not apply ‘ought’ to the action. (If that sounds implausible, fill in the details of the situation appropriately, or find another area where morality appears to be very demanding.) ‘Ought’ in the mouths of this community refer to the property that has the intersection of obligatory and psychologically feasible acts in its extension. Call this property obligation*.

In the shifty world where ‘ought’ refers to obligation*, it needn’t be that one’s own usage has shifted. One might continue to apply ‘ought’ to all and only actions that instantiate obligation, but nonetheless have one’s referent shift to obligation*,
in virtue of changes in one’s linguistic community. This is an artifact of the community-wide aspect of reference-determination. By hypothesis, some actions that are obligatory do not instantiate obligation*. But one will apply ‘ought’ to them, and by the connection between language and thought, will frequently form the corresponding beliefs. So one will have false normative beliefs. This is, in schematic form, the argument for Premise 3.

But this doesn’t guarantee that all normative beliefs in the shifty world will be false. While the belief formed by tokening the sentence ‘one ought to give 10% of one’s income to charity’ will be false, there are other, less precise normative beliefs that don’t have false counterparts in shifty worlds. For instance even in the shifty world, one forms a true belief by tokening the sentence ‘one ought to give some money to charity’, (even though in the shifty world, this is a belief about obligation*, and not obligation).

Likewise some logically complex normative beliefs will remain true: the belief formed by tokening the sentence ‘if one ought to give 10% of one’s income, then one ought to give 5% of one’s income’ is true in the shifty world. Similarly for ‘It is not required that one give 50% of one’s income to charity’. Again, these are true beliefs but are about obligation* and not obligation.

This doesn’t necessarily mean that beliefs formed by tokening these sentences in stable worlds will be free of bad companions. Recall that a bad companions need only to be similar enough to put true beliefs at risk of error. They need not be exactly the same as the beliefs they serve as bad companions for. The possibility remains that the belief formed by tokening the sentence ‘one ought to give some money to charity’ in a stable world will have a bad companion in a shifty world that is formed by tokening a different sentence. But the point here is that, while semantic shift can produce some false normative beliefs, it does not make every normative belief go false. Premise 3 needs to be read so as not to entail the possibility of such widespread false belief.

5.2 Premise 2: Stability to reference shift

So one can acquire a false belief in virtue of the reference shift of a moderately stable term like ‘ought’. Whether this happens depends on the guise one uses to form the relevant normative beliefs, and on which normative beliefs one forms. But in addition these false beliefs will need to be beliefs one could easily have formed, in the relevant sense. If normative beliefs in stable worlds all have similar false beliefs in shifty worlds, but the shifty beliefs are too distant, they won’y be bad companions.

Premise 2 claims that the nearbyness of such false beliefs follows from the moderate stability of ‘ought’:

2 If ‘ought’ is moderately semantically stable, then one could easily be in a world where ‘ought’ doesn’t refer to obligation.
We should note at the outset that there is a sense in which 2 is clearly true. Worlds where ‘ought’ has shifted in reference, and does not refer to obligation, can be *phenomenally indistinguishable* from worlds where ‘ought’ still refers to obligation. This is a consequence of the communal contribution to reference-determination: since reference depends on usage of one’s community, and changes in communal usage needn’t present themselves in the phenomenology of a competent speaker, the occurrence of a semantic shift needn’t make a difference to what a competent speaker is phenomenally aware of.23

So in many situations one could easily have a false belief owing to a semantic shift in ‘ought’, in the sense that one can’t phenomenally discriminate between states where a shift has occurred and states where it hasn’t. If someone asks you to cite an aspect of the phenomenally presented world as evidence that a shift hasn’t occurred, in most situations one will be forced to respond, “I can’t point to any conclusive evidence; a shift very well might have occurred.”

This kind of nearbyness—call it “phenomenal closeness” since worlds that are close in this sense are worlds that cannot be phenomenally distinguished from the actual world—is not the only sense we might give to the kind of closeness that matters for bad companionship. Someone who is in a normal perceptual environment and is at no risk of being envatted is not at risk of having false perceptual beliefs. But they cannot phenomenally distinguish their actual environment from one in which they are envatted. Likewise someone who has been told by a knowledgeable interlocutor that there is a traffic delay on the highway is not putting herself at risk by believing that there is a traffic delay, even though nothing in her phenomenology distinguishes the interlocutor from someone who is a habitual liar. These cases can be easily multiplied: the sense in which bad companionship involves the nearbyness of a false belief does not require that all worlds where one has a false belief can be ruled out on the basis of phenomenally presented evidence alone. Such an approach would be much too strict, ruling out knowledge of all but the most trivial truths.

This doesn’t guarantee that these worlds are close in the sense required to make Premise 3 guarantee the presence of bad companions for one’s actual normative beliefs. Phenomenal closeness does not guarantee bad companionship. (Holding fixed a reading of ‘could easily’ in terms of phenomenal closeness in the Argument from Risk, Premise 4 is false.) Premise 2 might still be true when we fix on the relevant reading of closeness, but focusing on the phenomenal closeness will not be helpful for establishing this fact.

Suppose then that we fix on a reading of “close” in 2 on which close worlds are, by definition, those that are close in the sense required for bad companionship. Call this *epistemic* closeness. This will ward of an understanding of the Argument from Risk on which the premises are all true, but rely on equivocal readings of

23In addition, competent speakers might not know when a change in usage generates a semantic shift, since they needn’t know the specific character of the supervenience relation between meaning and usage facts. See Williamson (1994) on this point in relation to vague terms.
“could easily be” (and cognate expressions) in order to secure their truth.

Understood in terms of epistemic closeness, 2 is not necessarily true. That is: at every world, the material conditional ‘if ‘ought’ is moderately semantically stable, then one could easily be (in the sense tied to epistemic closeness) in a world where ‘ought’ doesn’t refer to obligation’ is not true at every possible world. These worlds are those in the “center” of the sphere of stable worlds. In these worlds, if the sphere is large enough, one is as it were not at risk of leaving the sphere, and 2 is false.

More precisely: the moderate semantic stability of ‘ought’ guarantees that there is a set of worlds, M, where at every world in M, ‘ought’ (as used by one’s community) refers to obligation. Moreover the worlds in M will be fairly continuous in how one’s community uses ‘ought’ in those worlds: the worlds form a sphere in the sense that for every world in M, there are other worlds also in M where the community usage of ‘ought’ differs only in minimal respects. (In other words: there is no world in M where, in order to get to another world in M, one has to jump to a world where the community usage of ‘ought’ is substantially different in order to do so.) Of course there might be pairs of worlds in M where the community usage facts in the worlds are substantially different—what the sphericity of M amounts to is the claim that such worlds will be connected by a chain of worlds, also in M, where usage facts in each world in the chain are minimally different from the usage in its neighbors.

Some worlds in the M-sphere are at the center: one could not easily have been in a world that is not in M; which is to say every world one could easily have been in will also be a world in M. If we reserve the label $E_w$ for the set of worlds that are epistemically close to w, this amounts to the claim that some world $w^*$ in M is such that every world in $E_{w^*}$ is also in M. These will be worlds at which Premise 2 is false. So it is not necessarily true. Once we fix on a reading of Premise 2 on which closeness is read as epistemic closeness, the Argument from Risk contains a premise that is possibly false.

But the premise is not necessarily false. If Premise 2 is evaluated at a world which is at the edge of the sphere M, there are guaranteed to be some worlds which are epistemically close and which ‘ought’ has not shifted reference. But there will be other worlds at which it has shifted—if we find ourselves in a world which is in a precarious position at the edge of M, we can find another world within the bounds of epistemic closeness for the precarious world, but which do not fall within the boundaries of M. That is, if $w'$ is a world at the fringes of M, there will be some worlds in $E_{w'}$ which are not in M. The epistemic evils threatened by the Argument from Risk will be very real in these worlds.

The lesson here is that it is important to be cautious about how far these evils spread. The mere fact that there are some worlds in M which are epistemically close to worlds outside M does not entail that all worlds in M are epistemically close to worlds outside M. (As an analogy, I might go to a party where there is a risk, but no guarantee, that I will gamble $10. And if I gamble the $10, then there
is the risk that I will lose it. But just by going to the party I do not thereby risk $10."

Of course it might be that every world in M is close to a world outside M. But without establishing that the epistemic space is structured in this way, the Argument from Risk cannot guarantee success.24

The Argument from Risk is possibly but not necessarily sound, if one of its premises is contingently true. A separate question then arises concerning our ability to know whether it is true or not. That is, suppose 2 is true in some stable worlds and not in others. In the worlds where it is true, the Argument from Risk is sound and entails that normative propositions are not known. In other stable worlds 2 is false, and the conclusion of the Argument from Risk is false as well. There is no grounds for skepticism in these worlds. But whether we can know that the premises are not true in these worlds, and on this basis come to know that we have normative knowledge is a separate question. That is, it might be that even in stable worlds where we can have normative knowledge, we cannot, owing to the possibility of semantic shift, know that we have the knowledge.

There is a strong argument that we cannot. Call the worlds where ‘ought’ both refers to obligation, and in addition are not epistemically close to worlds where the referent of ‘ought’ has shifted, safe worlds. In safe worlds, one can have normative knowledge. But suppose in addition in one of these worlds that one believes that one knows some relevant normative propositions. For instance: consider a safe world where it is true that I ought to give 10%, I know that I ought to give 10%. Moreover I can believe that I know that I ought to give 10%. Even though the belief that I ought to give 10% has no bad companions (recall that we are in a safe world), the belief that I know I ought to give 10% does.

There are some cases which appear to show that sometimes risk of a risk collapses into risk simpliciter. For instance Dorr and Hawthorne make the following offhand comment:

If aliens are in fact going to invade the Earth in 2050, it would be rather tendentious to claim that we can still know that they will not invade in the next year. (Dorr and Hawthorne 2014: 321)

Here we can think of worlds where aliens invade in 2049 as epistemically close to worlds where they invade in 2050; worlds where they invade in 2048 are epistemically close to worlds where they invade in 2049, etc. Assuming they in fact invade in 2050, then we can’t know in 2049 that they won’t invade, since there is an epistemically close world where that belief is false. But it also seems that we can’t know that they won’t invade next year, and so the natural conclusion to draw is that being in a world that is close to a world that is close to where the invasion happens in 2050 is sufficient for being close to the invasion. (This principle would reduce any number of iterations of closeness to a single iteration by repeated applications.)

What this case suggests is not that some iteration of closeness is sufficient for epistemic closeness. Rather it shows that we need to careful about how we circumscribe the worlds that are epistemically close simpliciter. Just because in fact the aliens will invade in 2050, it doesn’t follow that a world where they invade next year is distant. (Separation by 33 years does not imply separation by 33 closeness relations.) The case as described doesn’t suggest any particular mechanism by which the actual invasion date is decades after one forms the relevant belief. So for all the description of the case says, the invasion could easily happen next year. Since the world where this happens is a world where one has a false belief, one doesn’t know that the invasion won’t happen next year. Getting to this conclusion doesn’t require collapsing iterations of closeness.
Since we are in a safe world, we are not epistemically close to worlds where we have false normative beliefs owing to a semantic shift. There are other non-safe worlds where the shift has not happened, but which are epistemically close to worlds where the shift has occurred. Call these unsafe worlds. We are epistemically close to some unsafe worlds. But in unsafe worlds it is false that we have normative knowledge. That is, even if it is true in such a world that I ought to give 10%, is is false that I know that I ought to give 10%. So if I believe that I know that I ought to give 10% in an unsafe world, I have a false belief. Moreover such belief will satisfy all of the conditions for bad companionship with the belief that I know that I ought to give 10% in a safe world: it will be sufficiently similar, formed by a nearly identical process, and in an epistemically close world. So I can’t know that I know I ought to give 10%, even in a safe world where first-order normative knowledge is available.

Finally it is worth returning to the question of which normative propositions we fail to know when we are in a stable but unsafe world. As we noted above, a shifty world won’t be one where all of one’s normative beliefs are false. One will still have true beliefs by tokening the sentence ‘one ought to give some money to charity’, even though by doing so one won’t form any beliefs about obligation. But other normative beliefs will be false: in our earlier example, one will believe something false by tokening the sentence ‘one ought to give 10% of one’s income to charity’. Perhaps in unsafe worlds, false beliefs like these will be bad companions for every normative belief. If so, then normative skepticism will be wholesale in unsafe worlds.

The case for wholesale skepticism in unsafe worlds is flimsy. Some normative beliefs in these worlds will have clear bad companions. By definition, and unsafe world is close to shifty worlds. And in a shifty world, there will be false normative beliefs: for instance one might falsely believe that giving 10% of one’s income is obligatory*. But on reflection these false beliefs cannot serve as bad companions for every normative belief. Simply because the belief that one ought to give some money, and the belief that one ought to give 10%, are both normative beliefs, it doesn’t follow that a bad companion is a bad companion for the other.

Start with logical truths about obligation: for the truth that either one ought to give 10% of one’s income to charity, or it is not the case that one ought to give 10% of one’s income to charity. Believing this in a safe but unstable world is compatible with knowing it—presumably one does not fail to know logical truths just because in believing them one could easily have believed a different logical truth.

Even some beliefs in non-logical truths will be sufficiently removed from bad counterparts that they will not fail to be knowledge, even in unsafe worlds. Take an analogy, inspired by Williamson (2000): if one is standing at a distance from a tree which is in fact 55 ft tall, and one has normal perceptual powers and is in a normal environment, one won’t be in a position to know that the tree is precisely 55 ft tall. If one believes that it is 55 ft tall, one will have an unknowledgeable
true belief that could easily have been false: since one’s perception isn’t accurate enough. One could easily have believed that the tree is 56 ft tall instead. But there are other beliefs about the height of the tree based on perception that are knowledge in this situation. For instance if one believes that the tree is between 50 and 60 ft tall, one will not only have a true belief, but will also be knowledge. Even if one had believed that the tree was between 51 and 61 ft tall, which could easily have happened, one would have believed truly. So beliefs with a significantly wide margin for error (in Williamson’s terms) can safeguard against false belief in nearby worlds that destroys knowledge.

One can insert margins for error in normative beliefs to guard against meta-semantic shift as well. Forming a normative belief by tokening the sentence ‘one ought to give some money to charity’ is an extreme example. Very likely there are more (though not fully) specific beliefs—such as the belief formed by tokening the sentence ‘one ought to give at least 2% of one’s money to charity’ that will not have bad companions. Meta-semantic risk can destroy some normative knowledge, but not all of it.

6 Conclusion

Realism about the normative implies that normative terms will be semantically stable to some degree. And since they are semantically stable, when one uses the public-language meaning of normative terms to form normative beliefs, one will be at risk of forming false normative beliefs when one is at risk of semantic shifts. This is the thought encoded by the Argument from Risk.

The Argument from Risk is potentially threatening to normative realism, since (unlike other arguments in the literature) it begins with a central feature of realism about the normative, and targets a largely uncontroversial necessary condition on knowledge and justification. That is, it claims that it follows from the moderate semantic stability of ‘ought’ that normative beliefs are subject to epistemic risk. Thus the Argument from Risk does not allow the normative realist to evade it by simply modifying the version of realism that she accepts, or denying the relevant epistemic virtues. Everyone can recognize that a view which entails the absence of knowledge or justification due to risk is in trouble.

But on closer investigation the Argument from Risk does not show enough to deliver a devastating epistemic blow to realism. If normative terms are sufficiently stable, one might be in a world where there is no risk of a reference shift. Since one may be unable to know how close the danger of a reference shift is, one won’t be able to know if one has normative knowledge or not. But it is in principle possible to know. Moreover even being in danger of a semantic shift does not entirely destroy normative knowledge. Logical truths about normative properties, and normative beliefs with a sufficiently large margin for error will not be at risk of knowledge-destroying error, even if shifts are close.

This should obviate the fact that the Argument from Risk points to some
striking epistemological consequences for realism. Even if one isn’t guaranteed to be at risk of a meta-semantic shift, one might be at risk of such a shift, and if so, many common normative beliefs won’t be knowledge. Moreover one can arguably never know whether one has this knowledge that is contingent on semantic shifts or not. If so, higher-order normative knowledge is difficult to come by. And finally when the shifts real dangers, one will have to form beliefs that have sufficiently wide margins for error in order to retain knowledge. Ordinarily we would suppose that our normative knowledge is less contingent, more wide-ranging, and more specific. If realism entails that it is not, this is surprising indeed.

Of course this is just a sketch of the epistemology of normative realism with moderate semantic stability in the picture. There are many places where further details would be nice to have. For instance more precision on the extent of the semantic stability of ‘ought’ would be nice to have. For some possible degrees of stability for the term, there won’t be any “safe worlds” where all of the worlds one could easily have been in are not shifty worlds. Every world where ‘ought’ refers to obligation will be close to some shifty worlds. In this case it is not a contingent matter whether one is subject to meta-semantic risk, and normative knowledge will be absent to some extent throughout.

Alternatively the degree of semantic stability in ‘ought’ could be quite high. This might not only make it (contingently) possible to have full first-order normative knowledge, but if the stability is robust enough, one will contingently have some higher-order knowledge about one’s own normative knowledge as well: all the nearby worlds where one believes one knows will be worlds where one believes truly, since all to the worlds nearby to them will not be shifty.

And finally one would like to know more about exactly which normative beliefs are similar enough to the beliefs the go false in shifty worlds. This will tell us in more detail which beliefs are not knowledge when we are at risk of a shift. We know that beliefs which encode a sufficiently high margin for error will not be at risk. But more detail on the closeness-relation, and what the candidate alternative referents for ‘ought’ are, will tell us more about exactly which beliefs encode sufficient safeguards against meta-semantic risk.

References


Dunaway, B. (forthcominga), ‘Ethical Vagueness and Practical Reasoning’, *The Philosophical Quarterly*. 24

Dunaway, B. (MS), ‘Grounds and Knowledge: Epistemological Motivations for Anti-Realism’.


Hare, R. M. (1952), The Language of Morals, Oxford University Press. All citations to the 1964 edition.


Moore, G. (1903), Principia Ethica, Cambridge University Press.


