Although foundational issues about meaning and concepts rarely take center stage in contemporary metaethics, metaethical debates are shaped by background assumptions about the nature of concepts. The issue of when there is direct logical agreement or disagreement among our normative thoughts, for instance, is at bottom a matter of concept identity. And the issue of whether our normative thoughts all pick out a single property depends in part on what is required for competence with the same concept and how concepts in general acquire their semantic values.

This paper sketches a new relational account of normative concepts that challenges widely held assumptions in metaethics. We start by articulating the central role played by concepts in keeping track of particular topics in thought and talk. Sameness of concept normally gives rise to the appearance of guaranteed sameness of topic, and this psychological appearance must be reliable if concepts are to play their characteristic role. We then introduce and criticize the standard broadly Fregean approach to concepts that has often been common ground in metaethical debates. The bulk of the paper articulates our alternative connectedness model of concepts and shows how this model is particularly well suited to normative concepts. The connectedness model builds social and historical facts into the foundations of concept identity. This aspect of the model, we suggest, reshares normative epistemology and provides new resources for a vindication of realism in ethics.

1. Concepts and keeping track of a topic in thought and talk

A personal crisis has plunged you into a state of serious moral disarray. You now wonder whether some of your deepest moral convictions are really justified and whether morality has the authority it is usually thought to have. You find you’re no longer motivated to act in accord with your moral judgments. Your own youthful moral enthusiasm now strikes you as naïve, and you wonder how your friends can still have so much confidence in moral views you now find so problematic. After watching an exposé on television that would have formerly left you morally outraged, for instance, you find yourself thinking:
Of course imprisoning asylum seekers is morally wrong. But does it really matter that it’s morally wrong? There’s no reason why I should care about what’s morally wrong.

In your reflection about the asylum seeker case, it seems obvious that there is a single topic in question — what’s morally wrong.1 In your conscious reasoning, your successive thoughts all present themselves as unquestionably about the same topic. And it seems equally obvious that your current disaffected thoughts concern the same topic as your earlier moral convictions: your denial that you should care about what’s morally wrong seems to directly logically contradict your past belief that everyone should care about what’s morally wrong. Similarly, your own thoughts strike you as obviously pertaining to the very same topic as the thoughts others express when they use the expression ‘morally wrong’ in conversation. Your belief that imprisoning asylum seekers is morally wrong, for instance, seems to directly logically contradict the assertions of politicians who claim that there’s nothing morally wrong about this policy.2

This phenomenon — the immediate appearance of two thoughts as obviously and unquestionably pertaining to the very same topic — is a reflection of the underlying semantic structure of thought and talk. It seems that we must have some epistemically basic ways of keeping track of particular topics in thought: without some basic way of recognizing sameness of topic among one’s own thoughts, learning and reasoning would be impossible. Consider your ability to learn about topics like President Obama, water, or what’s morally wrong. To accumulate a body of information about a topic, you must be able to immediately understand some new thoughts as pertaining to the very same topic as certain of your standing attitudes. Furthermore, your ability to keep a train of reasoning on topic, to recall pertinent stored attitudes, and to avoid direct logical contradictions all depend on this immediate appreciation of sameness of topic in thought.

Similarly, when other people use familiar terms like ‘Obama’, ‘water’, or ‘morally wrong’, their claims seem obviously to pertain to the very same topics you yourself associate with those terms: the question of sameness will strike you as closed. This distinctive appearance of sameness of topic reflects your basic ways of keeping track of particular topics in language, which enable full and fluid participation in interpersonal discourse. This basic linguistic capacity to keep track of sameness, for instance, allows you to immediately recognize a recurring topic in conversation and to detect direct logical disagreement with others.

These basic ways of keeping track of a particular topic in thought and talk, and the subjective appearances that they give rise to, have figured prominently in philosophical accounts of concepts. Consider the following constraint on concept identity, derived from Frege’s widely accepted criterion for individuating senses (Frege 1892):

\begin{quote}

**Cognitive difference principle:** If two token elements of thought don’t seem, from the perspective of the subject who entertains them, to obviously and unquestionably pertain to the same topic, then they express different concepts.
\end{quote}
This principle ties the notion of a concept to the subject’s own epistemic perspective on sameness of topic. Although your ‘Hesperus’ and ‘Phosphorus’ thoughts represent the very same topic (namely Venus), this sameness may not be obvious and unquestionable from your subjective perspective. It would then be compatible with minimal standards of rationality to doubt that Hesperus is Phosphorus, or even to accept that Hesperus has properties that Phosphorus lacks. According to Frege, such divergent attitudes toward the same topic indicate that these thoughts involve different ways of thinking about the topic—what we are calling concepts. As Allan Gibbard suggests, Moore’s open-question argument invokes a similar constraint on concept identity. The fact that the question ‘It’s good, but is it pleasure?’ strikes you as open (i.e., the identity claim ‘good is pleasure’ can be rationally doubted) suffices to establish that your understanding of that question involves two distinct concepts (Gibbard 2003).

There is room for theoretical disagreement about how exactly to formulate the cognitive difference principle. Our formulation is meant to capture how things appear to the subject when keeping track of a topic in conscious reasoning. When you ruminate about the asylum seeker case during your moral crisis, the question of sameness of subject matter seems closed: your thoughts present themselves to your conscious attention as obviously and unquestionably about the very same topic. In order to raise a coherent doubt about sameness of topic, you’ll need to adopt an explicitly meta-level perspective on your own thoughts: you can coherently doubt, for instance, whether the topic picked out by morally wrong in your chain of thought has really remained stable over time.3 But as long as you stick with your ordinary object-level perspective of thinking about what’s morally wrong, the question of sameness will seem just as trivially closed as the question of whether Hesperus is Hesperus. From this object-level perspective, certain thoughts appear guaranteed to pertain to the same topic. We’ll refer to this as the appearance of de jure sameness and contrast it with the appearance of de facto sameness that is characteristic of Hesperus/Phosphorus cases.

Our aim in this paper is to articulate an account of concept identity that can explain central aspects of normative thinking. Concept identity—whether two thoughts involve the very same concept—plays a key role in philosophical theorizing about mental states in general.4 Sameness of concept determines standards for assessing logical coherence in a subject’s thinking and provides the basis for rationalizing explanations of her reasoning and actions. Getting clear about what’s required for deploying the same concept helps to clarify the basic logical relations among elements of thought. In particular, concept identity determines relations of de jure sameness of topic that ground direct logical relations of contradiction, entailment, agreement, and disagreement among thought contents. And this issue of direct logical relations among thought contents has been central to metaethical theorizing.

Relations of direct logical contradiction, for instance, hinge on whether thoughts deploy the same concepts—not just whether the topics picked out are the same. In cases of direct logical contradiction between two thoughts, their contents must have the logical form of P and ~P; and to have this logical form, the two thoughts must be

---

3. We use small caps (e.g., dog) to denote the concept a subject expresses with the associated word (e.g., ‘dog’). We take concepts and context-neutral linguistic meanings to be systematically aligned, though the correspondence is not perfect: not all words express concepts, and not all of a subject’s concepts are associated with words.

4. A theory of concept identity is not the same as a theory of the metaphysical nature of concepts. At a very generic level, concepts can be understood as recurring constituents of whole thoughts, which make stable semantic contributions to the truth conditions of the thoughts in which they figure. The semantic role of a concept, for instance, may be to contribute an object, kind, property, or fictional entity to the complex state of affairs represented by a thought. But theorists disagree about how exactly to understand the metaphysical nature of concepts: e.g., whether they are mental representations, abstract objects, or cognitive abilities (Margolis and Laurence 2011). We’ll ignore this metaphysical issue in this paper and focus on the psychological conditions for when two elements of thought express the same concept. Our account of concept identity is consistent with different views about the intrinsic nature of concepts themselves.
composed of the same concepts and structured in the same way. Direct logical contradiction among thoughts is a much stronger relation than that of placing metaphysically incompatible constraints on the world. Logical contradictions among thoughts require relations of *de jure* sameness of topic, which are marked by sameness of concepts: although your current thought that Hesperus is a planet and your past thought that Phosphorus is not a planet cannot be true together, they do not logically contradict each other, since they have the logical form of *Fa* and ~*Fb*. Similarly, your normative judgment that imprisoning asylum seekers is morally wrong does not directly logically contradict local politicians’ judgment that doing so is not shameful — even if it turns out that the substantive commitments undertaken in judging something morally wrong are just the same as those involved in judging something as meriting shame. Here, too, there is a difference in logical form between the two thoughts.

Conceptually structured thought contents play the role of structured propositions, but we avoid using the term ‘proposition’ here in order to accommodate expressivists who deny that normative judgments express propositions. Such theorists, we contend, should still allow that normative judgments deploy normative concepts that compose whole thought contents. Even expressivists should admit that normative thoughts strike subjects as pertaining *de jure* to the same topic, and normative concepts ground this appearance.

Although metaethical realists and anti-realists may disagree about the nature of the commitments undertaken in normative judgments and about which psychological states ground logical relations among them, they can agree that there is no direct logical contradiction between these two thoughts: since there is room to rationally question whether being morally wrong just is being shameful, these two thoughts differ in logical form.

We take logical relations such as direct contradiction to be central to metaethics. Of course, there are many less stringent types of incompatibility among thought contents than direct logical contradiction. As in the Hesperus/Phosphorus case, the contents of your two beliefs may be *metaphysically incompatible*, even if this incompatibility cannot be known without further empirical information. Similarly, there may be an *indirect logical contradiction* between two of your belief contents that can be discerned only by means of a multi-step proof. Although metaethicists sometimes define ‘disagreement’ in ways that include weaker types of incompatibility in contents (Plunkett and Sundell 2013), the stricter notion of direct logical contradiction has been central to metaethics. Moore’s open-question argument, for instance, turns on the idea that (alleged) *a priori* metaphysical entailments like ‘it’s good; therefore it’s pleasant’ are rationally questionable (*i.e.*, they involve no direct

---

Following this philosophical tradition, our account starts with the cognitive difference principle. This principle is a partial constraint on concept identity: it provides a necessary condition for sameness of concept for the special case of co-conscious thoughts. A theory of concepts must extend from this core case to provide necessary and sufficient conditions for entertaining the same concept. In particular, the theory must specify conditions for sameness of concept among unconscious thoughts and thoughts entertained at different times. In addition, an account of concepts must address the question of concept identity in the interpersonal case: Can different individuals entertain the same concept, and if so, under what conditions? We’ll show how our connectedness model of concepts can provide general identity conditions for concepts shared between individuals.

2. The template model

The cognitive difference principle ties concept identity to the appearance of *de jure* sameness of topic among thoughts. But if these appearances were systematically misleading, concepts could not fulfill their main function — keeping track of a particular topic (*e.g.*, Obama, water, what’s morally wrong) in thought. So on a plausible account of concepts, the appearance of *de jure* sameness must be reliable: when two thought elements seem to pertain *de jure* to the same topic, this appearance is very likely to be veridical.

The standard, broadly Fregean, approach to concept identity is designed to explain this reliability. On this account, sameness of topic is ensured by the fact that token thought elements are each associated with a particular pattern of understanding that fixes the topic in question. The standard account can be formulated as follows:

**Template model**: two token elements of thought express the same concept iff each is associated with the same topic-fixing criterion.

---

logical contradiction), while the Frege-Geach problem for expressivism hinges on direct logical entailment relations between the contents of normative attitudes.
A topic-fixing criterion is a specific pattern of understanding (realized in the subject’s recognitional, inferential, or motivational dispositions) that suffices to distinguish the topic picked out by a thought element from all other possible topics. To take a standard example: Two elements of thought express the concept BACHELOR just in case the subject associates each element with the criterion of being an unmarried but eligible man. The guiding idea is that the topic picked out by a token thought element is determined by the criterion associated with that token. So when two tokens are associated with the same criterion, they are guaranteed to pick out the same topic.

This basic model is slightly modified by neo-Fregeans. Semantic externalists argue that subjective understanding alone, considered independently of one’s empirical context, does not always suffice to determine the precise topic picked out (Putnam 1970, 1972; Donnellan 1970; Kripke 1980; Burge 1979, 1982). Your purely subjective understanding of Obama or of water, for instance, may not suffice to distinguish those topics from the distinct topics picked out by a twin-earth counterpart who shares exactly the same topic-fixing criteria. In response, neo-Fregeans have argued (i) that sameness of topic-fixing criteria is still a necessary condition for expressing the same concept, and (ii) that topic-fixing criteria determine the topic as a function of specific facts about the subject’s empirical context (Lewis 1970, 1994; Peacocke 1992, 1998; Jackson 1988a, 1998b; Chalmers 2002a, 2002b). For instance, one might argue that two thought elements express the concept water only if each is associated with the criterion of being the basic potable liquid with which I have been acquainted, and two token thoughts associated with the criterion are guaranteed to pick out the same topic provided there has been no shift in historical context. So on this modified template model, sameness of topic-fixing criterion still provides a reliable conditional guarantee of sameness of topic: insofar as there is no hidden shift in relevant aspects of one’s empirical context, sameness of topic-fixing criterion guarantees sameness of topic.

The template model is widely accepted in contemporary metaethical discussions of normative concepts and meanings. The debate between expressivists and neo-descriptivists, for instance, is usually structured by a shared acceptance of that model. According to neo-descriptivists, topic-fixing criteria for normative concepts (including thin normative concepts like RIGHT OF REASON and thicker concepts like MORALLY WRONG, UNJUST, OR CRUEL) involve complex “folk theories” or “implicit conceptions” of what it takes for something to exemplify the properties picked out by those concepts (Jackson and Pettit 1995; Peacocke 2004). These descriptive criteria must be minimal enough that they are plausibly shared by everyone who is competent with normative concepts. All true claims about normative properties that go beyond these minimal commitments must then be justifiable on the basis of these core topic-fixing criteria. In contrast, expressivists typically take motivational states to constitute the core topic-fixing criteria for thin normative concepts, rather than any specific descriptive assumptions about what it takes for the concept to be applicable to particular cases. According to (Gibbard 2003), for

---

[7] This is unsurprising, given that the template model is the standard philosophical model of concept identity and provides the standard explanation of the difference between Hesperus/Hesperus and Hesperus/Phosphorus cases. The main exceptions to this rule in metaethics are theorists inspired by semantic externalism about natural kind terms who have challenged the template model both as an account of concept identity and as an account of the determination of semantic values (Boyd 1988; Brink 1989; Sayre-McCord 1997; van Roojen 2006). Our connectedness model is similar in spirit and motivation to this family of views. These theorists, however, have not sought to explain cognitive difference phenomena or to provide a systematic account of the identity conditions of normative concepts. These are the questions we focus on in this paper. One prominent anti-realist, Simon Blackburn, also rejects the template model as an account of linguistic competence with normative terms, on the grounds that it conflicts with the best account of linguistic conventions (Blackburn 1984, chap. 4, especially 130–134; 1991, 4–13; 1998, 59–68). Although our main focus is thought content, not linguistic communication, we find Blackburn’s criticisms of the standard template model of linguistic competence congenial. And we agree that competence with the same meaning hinges on a commitment to communal standards. Our account of interpersonal concept identity seeks to explain the individuation of linguistic communities and the nature of individuals’ commitments to communal standards — issues that Blackburn does not address.
instance, to think something is the all-things-considered right thing to do is roughly to have a plan to perform it should the circumstances arise. This core motivational profile is supposed to single out the particular normative topic picked out by the thin concept right from all other possible topics. The goal of Gibbard’s quasi-realist program is to vindicate the apparently representational features of our practice with normative concepts by deriving these features from this core motivational criterion for deploying the concept.9

Although there are many different ways of elaborating a template model of normative concepts along broadly expressivist or descriptivist lines, all template accounts posit a core topic-fixing criterion. Any change in the core criteria, according to the template model, will destroy the apparent guarantee of sameness of topic that’s distinctive of concepts.9

8. Other popular metaethical approaches also posit topic-fixing criteria. For instance, error theorists and fictionalists posit descriptive criteria that are unsatisfied in the actual world. Contextualists and relativists take their cue from semantic theories developed for indexicals like ‘I’ or ‘now’. On these accounts, the topic-fixing criterion associated with a normative predicate like ‘is morally wrong’ picks out a specific property, but only relative to a ‘hidden’ indexical parameter that imports conventional standards. (Contextualists and relativists disagree about whether the relevant standards are fixed by the context in which a thought was originally formed or instead by the context in which that thought is assessed.)

9. An exact match in topic-fixing criteria is essential to the way the template model explains concept identity and logical relations among thoughts. Concept identity is supposed to be an equivalence relation: reflexive, symmetric, and transitive. However, mere similarity of criteria will clearly fail the transitivity requirement. Mere similarity of topic-fixing criteria, moreover, cannot explain logical relations like direct logical contradiction (P & –P) or trivial identity (a = a). To ground such logical relations, the topic-fixing criteria must make it strictly impossible for there to be different semantic values assigned to the associated elements of thought (given the same context). If the topic-fixing criteria are not identical but merely similar, there will be empirical contexts in which the semantic values may diverge. For instance, if A associates ‘Hesperus’ with the criterion ‘the heavenly body at location x (with margin of error 5) in the evening’, and B associates the same name with the criterion ‘the heavenly body at location x (with margin of error 10) in the evening’, then there will be possible contexts in which A and B will be picking out distinct semantic values. It follows that even in their actual context in which they pick out the very same semantic value (i.e., Venus), A and B are not directly logically disagreeing with each other when A says ‘Hesperus is bright’ and B says ‘Hesperus is not bright’ — for there is a coherent way the world could be that would make both claims come out true.

10. According to Moore, the fact that you can rationally doubt that what’s right is what maximizes utility shows that one cannot analyze rightness in terms of maximization of utility. But our understanding of most topics is primarily constituted by implicit recognitional, inferential, or motivational dispositions rather than consciously accessible explicit definitions. So Moore’s argument cannot directly refute a putative analysis. The template model must be able to explain why sameness of one’s topic-fixing criteria is correlated with apparent de jure sameness of topic. But the template model is not committed to an explicit analysis of that criterion being itself obvious and unquestionable. According to the cognitive difference principle, what must be obvious and unquestionable is sameness of topic, not an explicit definition of that topic.

Normative Concepts: A Connectedness Model

The template model’s inflexibility about what’s required for apparent de jure sameness, however, clashes with our experience as normative thinkers (Blackburn 1991; Merli 2009). In the moral-crisis example, for instance, your implicit moral beliefs and motivations change radically, and yet it seems obvious to you that you’re thinking about the same topic. Indeed, it seems that any particular assumption could be dropped without threatening the appearance of de jure sameness. Your moral motivations can also undergo radical changes—you may become disaffected, cynical, or even morally perverse — without undermining the subjective appearance of de jure sameness. From the first-person perspective, it seems possible for normative concepts to persist through virtually any change in belief or motivation, even changes in one’s deep-seated implicit cognitive or motivational dispositions.

This commitment to flexibility, we suspect, is one important reason why many expressivists continue to be moved by Moore’s open-question argument. Moore’s own formulation of that argument turns on implausible views about our ability to recognize correct explicit analyses of our implicit understanding of a topic. Still, contemporary expressivists insist there is an important grain of truth in Moore’s argument: even if we acknowledge that explicit definitions needn’t be obvious, it still seems implausible that an inflexible core of topic-fixing descriptive criteria is required for apparent de jure sameness. After all,
we have an interest in being able to achieve logical coordination on the same topic with people who don’t share our moral views. And insofar as we take morality to be an objectively important topic, intellectual virtue seems to counsel against taking our naïve, pre-reflective understanding to provide strict constraints that settle what counts as moral. So imposing core descriptive preconditions for competence with the concept morally wrong would impose arbitrary limits on direct moral debate, disagreement, and inquiry. Normative concepts structured in this way would not serve our epistemic interests.

This line of argument against descriptivist topic-fixing criteria is sometimes taken to support motivational topic-fixing criteria instead. It’s important to see, however, that similar considerations militate against motivational topic-fixing criteria. No particular motivational state seems strictly required for the appearance of de jure sameness — much less a motivational state sufficiently distinctive to differentiate a given topic like moral wrongness from closely related topics like what’s prohibited by mere conventions (Blackburn 1991; Merli 2008). Moreover, insisting on a core motivational state would impose arbitrary limits on direct engagement on the same topic. The disaffected, cynical, or perverse would be arbitrarily excluded from our moral conversations: because of their motivations, they could not directly logically agree or disagree with the rest of us about moral claims. The general lesson is that flexibility is important to the role of concepts in underwriting open-ended debate and inquiry. A template model of concepts — whether descriptive or motivational — is structurally unsuited to vindicating such flexibility.\(^{11}\)

\(^{11}\) One might wonder how broadly the flexibility objection to the template model applies. Allan Gibbard (2003) and Ralph Wedgwood (2007), for instance, take purely motivational states (plans or dispositions to plan) to constitute the topic-fixing criteria for thin normative concepts like ought or all-things-considered right. We believe the flexibility objection applies to such motivation-based accounts of the thinnest normative concepts: e.g., even if you lose any disposition to form plans or preferences on the basis of what you judge right, you can still form thoughts that strike you as de jure pertaining to the same topic, and this judgment may be correct. We take Simon Blackburn (1991) to agree with us on this point, whereas David Merli (2008, 2009) confines his argument to thicker moral concepts. It is worth emphasizing, however, that the issue of concept identity cannot be settled by a mere appeal to intuitions — and this is doubly true of exotic cases such as subjects with radically aberrant motivational dispositions. Such cases must ultimately be decided by the best theory of concept identity. The connectedness model of concept identity we propose in this paper gives a principled reason for allowing flexibility across the board.
Providing identity conditions in terms of relations among particular token elements is a familiar metaphysical strategy. Consider the case of personal identity. Traditional theories explain the relation of personal identity in terms of matching: two person-stages are stages of the same person just in case they are both associated with, e.g., the same immaterial soul or with the same core set of memories. Connectedness theories, in contrast, explain personal identity in irreducibly relational terms: two person-stages are stages of the same person just in case, e.g., their mental states are causally connected in the right way. Our approach to concept identity follows a similar strategy. The key structural advantage of a relational model of concept identity is the same as that of a relational model of personal identity: relational models place no direct constraints on the independent properties of the two relata, so they can accommodate our commonsense commitments to variability in ways that matching models cannot.

This initial characterization of the connectedness model is purely structural. In order to provide a genuine alternative to the template model, this relational schema must be fleshed out in such a way as to vindicate the core role concepts play in keeping track of a topic in thought and talk.

The first task is to provide an account of the connections that are crucial to concept identity (§4). Our strategy is to start with the basic epistemic relation at the heart of the cognitive difference principle — the appearance of de jure sameness. It’s worth emphasizing that our approach privileges the level of thought rather than linguistic communication: to explain concept identity we focus in the first instance on a basic epistemic relation that connects elements within an individual’s co-conscious thoughts. We then show how this basic relation of apparent de jure sameness can be extended from the core case of co-conscious thoughts to the subject’s own past thoughts and to thoughts entertained by others. Just as connectedness accounts of personal identity may appeal to chains of memory links, our account appeals to chains of relations of apparent de jure sameness linking token elements of thought.

12. Moreover, the idea that recognition of sameness must always be based on explicit comparison and matching leads to a vicious explanatory regress. At some point, there must be a basic cognitive disposition to simply treat certain thoughts as pertaining to the same topic. Lewis Carroll made a similar point about our ability to reason in accord with modus ponens (Carroll 1895).
The following section (§5) tackles the reliability of the appearance of *de jure* sameness. On our account, apparent *de jure* sameness does not require matching topic-fixing criteria. But without such a match, what ensures that token elements of thought that appear *de jure* to pertain to the same topic really do pick out the very same thing? To address this challenge, we’ll need to take a closer look at the determination theory—the theory that explains how the semantic value of thought elements is determined by facts about the subject’s understanding, environment, history, etc.\textsuperscript{13}

We will then be in a position to provide necessary and sufficient conditions for concept identity in §6. It’s not obvious that a relational account like ours can ground a genuine identity relation—*i.e.* an equivalence relation that’s reflexive, symmetrical, and transitive. Some proponents of relational accounts of sameness of meaning, for instance, deny transitivity (Fine 2007; Pinillos 2011).\textsuperscript{14} Our account of concept identity seeks to answer this challenge.

Our guiding principle in developing the connectedness model is to capture the first-person epistemic perspective of those who actually use normative concepts. As we emphasized earlier, concepts are tied to the subject’s basic ways of keeping track of a topic in thought and talk. Our account starts with an immediate subjective appearance—taking two thoughts as pertaining *de jure* to the same topic—and explains why subjects naturally take this *de jure* sameness relation as extending to their past thoughts and to others’ thoughts. The same first-person perspective is crucial to our account of the reliability of these appearances. Reflection or empirical information can sometimes lead you to conclude that *de jure* appearances were mistaken: thoughts that had seemed to pertain to the same topic were in fact about distinct things. In explaining when *de jure* appearances are reliable and when they are not, our account aims to vindicate the first-person reflective perspective of the subject herself.

4. Tracing the connections

Let’s start with the appearance of *de jure* sameness and our basic ways of keeping track of a topic that give rise to such appearances in thought. As we saw in our discussion of the template model, it’s implausible to explain those *de jure* appearances in terms of matching topic-fixing criteria. The connectedness model proposes a different, irreducibly relational approach to explaining these phenomena. What grounds the appearance of *de jure* sameness, we suggest, are cognitive mechanisms that bind together token elements of thought directly, without relying on a match in the independent properties of the tokens so bound.

The metaphor of a mental file folder helps illustrate the approach we have in mind.\textsuperscript{15} Manila file folders in your filing cabinet serve to bind together a body of documents as pertaining to a given topic, such as health insurance, and segregate it from similar bodies of documents pertaining to different topics, such as taxes or car repairs. When you open your health insurance file, you’ll immediately take it that all the documents therein pertain to the same topic: your filing system presents the bound documents as pertaining to a single topic. Mental files serve a similar function. They bind together an evolving body of attitudes and cognitive dispositions as all pertaining to the same topic (*e.g.*, Obama, water, moral wrongness, injustice, or cruelty) and they

\textsuperscript{13} ‘Determination theory’ is Christopher Peacocke’s term (Peacocke 1992). Others have called a theory of how semantic values are determined a ‘metasemantic theory’ (Kaplan 1989) or a theory of ‘foundational semantics’ (Stalnaker 1997).

\textsuperscript{14} Similarly, Derek Parfit ultimately concludes that there is no such thing as a strict personal identity relation—instead he claims that a certain (non-transitive) psychological connectedness relation is what really matters to us when we talk about personal identity (Parfit 1984).

\textsuperscript{15} The mental file metaphor is a popular one, which has been developed in different ways (Strawson 1974; Evans 1973, 1982; Perry 1980, 2001; Forbes 1989; Recanati 1993, 2012; Millikan 2000; Lavorl 2001; Jeshion 2010; Dickie 2010; Schroeter 2008, 2013). What’s distinctive of the approach we assume here is that (i) mental files are not exclusively confined to thought about individuals—there are files for kinds and properties as well; (ii) mental files are not individuated by their contents, *i.e.*, the substantive understanding they bind together, but by the binding mechanism itself; and (iii) mental files persist over time in virtue of diachronic relations of apparent *de jure* sameness.
segregate those attitudes and dispositions from similar bodies that pertain to different topics.

The mental file metaphor highlights important structural features of our basic ways of keeping track of a topic in thought. Most importantly for our purposes, the binding relation established by files is irreducibly relational — it does not supervene on the properties of the bound elements considered independently. In a physical filing cabinet, the way documents are bound together into discrete files cannot always be deduced from the properties of the individual documents considered on their own. If a note from Joe has slipped into your Jill file, you’ll be disposed to immediately treat it as pertaining to the same person as other documents in your file. This disposition is not grounded in a comparison of the bound documents: the filing system itself is a mechanism that grounds this initial presumption of sameness of topic. Similarly, mental files establish a relation of apparent sameness among bound elements that is not determined by the properties of those elements considered on their own.

Of course the body of attitudes and dispositions bound together into a particular mental file at a given time and the structure of the associated representations play a central role in psychological explanations of categorization, induction, prototype effects, learning, etc. See (Murphy 2002) for an overview of the psychological literature on the nature of the mental representations involved in categorization. Our account of mental files is neutral with respect to this debate in the psychological literature over the nature of the actual patterns of understanding and the structure of mental representations associated with a file at a time. As we emphasized in §2, the primary motivation behind the connectedness model is to explain how normative concepts can remain stable despite changes in this substantive understanding.

Many psychologists and linguists have come to similar conclusions about the irreducibly relational nature of our disposition to treat elements of thought or discourse as pertaining to a single topic. Cognitive psychologists, for instance, have appealed to ‘object files’ or ‘visual indices’ to characterize our capacity to keep track of apparent objects in vision, binding information about size, shape, trajectory, etc., as pertaining to a single thing within the visual system: e.g., you’ll expect bound features to move together as a unit (Kahneman and Treisman 1984; Pylyshyn 1989). Developmental psychologists invoke object files in more general intermodal reasoning about objects in infants (for an overview, see Carey 2009). Psychologists argue that features are bound together in virtue of their relational properties, rather than in virtue of facts about those features considered independently of each other: for instance, the visual system seems to bind object-stages over time in virtue of their being linked by stages forming continuous pathways through occlusion, rather than in virtue of their resemblance in color, shape, speed, or trajectory. In linguistics, the phenomena of anaphora and discourse reference provide models for direct relations underwriting apparent de jure sameness of topic. Clearly, anaphora and variable binding establish relations that do not depend on matching independent features of the bound representations. While binding and co-indexing are often seen as purely syntactic phenomena, the relevant syntactic relations place constraints on semantic interpretation, and some authors have appealed to such syntactic relations to explain the general phenomenon of de jure sameness (Fiengo and May 1994, 2006). Discourse reference and file update semantics, in contrast, are meant to occur at the semantic level: they involve accumulating information over time about a putative topic. Here again, the relevant semantic relation among earlier and later thoughts cannot be explained by match in substantive understanding (Karttunen 1976; Kamp 1985; Heim 1983).

Moreover, the mental file metaphor helps us see why the relation of apparent de jure sameness should not be confined to co-conscious thoughts. Manila file folders establish stable binding relations among persisting documents. Your ‘Jill’ file, for instance, grounds stable dispositions to treat the documents it binds as pertaining to the same topic: your letters from Jill persist over time, and your filing system grounds a stable disposition to treat them all as pertaining to the same topic whenever you access them. Similarly, mental files establish stable relations among persisting representational states — your standing attitudes. Consider your standing attitudes pertaining to cruelly: e.g., your belief that imprisoning asylum seekers is cruel, your intention to avoid cruel actions, your episodic memory of being the victim of a schoolyard bully’s cruel behavior, and your theoretical understanding of the psychology of cruelty. Your mental filing system stably binds these persisting attitudes together in such a way that when you entertain these attitudes together in conscious thinking, they’ll immediately strike you as pertaining de jure to the same topic — cruelly. The point is that this relation of apparent de jure sameness is not an ad hoc relation that comes into existence on a particular occasion — it is a stable relation connecting your standing attitudes. The mental file metaphor helps highlight this stability and the fact that multiple attitudes are bound together in this way.
By appealing to the mental file metaphor, we have taken a first step toward identifying a connection that can ground concept identity. The cognitive difference principle focuses on the relation of apparent *de jure* sameness linking elements of co-conscious thoughts: being linked in this way is a necessary condition for expressing the same concept. The notion of a mental file helps us see why the relevant relation extends beyond conscious episodes of reasoning: a mental file links together a network of standing attitudes in such a way that the subject is disposed to treat elements of those attitudes as *de jure* pertaining to the same topic. When elements of thoughts are bound together in a file, they satisfy the core condition on sameness of concept.

The file metaphor also helps us to understand how past and present attitudes can be bound by a relation of apparent *de jure* sameness. Among the attitudes bound by a file at a particular time, there will be episodic memories of the past. Your moral wrongness file, for instance, will bind your childhood conviction that abortion is morally wrong together with your memory of rejecting this doctrine as a mere cultural prejudice as a teenager. These standing memories share a common element that strikes you as pertaining *de jure* to the very same topic, moral wrongness, which figures in your current moral thinking. The key point is that, from your current perspective, your memories seem to afford direct access to the content of your past attitudes: in ordinary thinking, you don’t distinguish the content of your current memory and the content of the childhood attitude from which it derives. So when you consciously entertain these memories, the appearance of *de jure* sameness won’t be confined to your attitudes right now—it will also extend into the past. Although your current understanding of moral wrongness is very different from that of your childhood self, you’ll understand your past beliefs and intentions as pertaining *de jure* to the very same topic as your more recent ones. By linking memories together with your current standing attitudes, mental files give rise to the appearance of *de jure* sameness of topic over time.

Your current memories, however, afford very patchy, selective, and somewhat unreliable access to your past thinking: you’ll have forgotten most of your childhood attitudes pertaining to what’s morally wrong, for instance. Your current mental file cannot generate any appearance of *de jure* sameness between your current attitudes and forgotten mental states. Nevertheless, it is very natural to assume that your current attitudes pertain *de jure* to the same topic as forgotten past states in virtue of chains of relations of apparent *de jure* sameness. At each stage in this chain, it strikes you as obvious and rationally unquestionable that the topic of your attitudes is identical. So taking this appearance at face value commits you to this *de jure* sameness extending from your present attitudes to appropriately linked past attitudes that you no longer remember. *Prima facie*, the appearance of *de jure* sameness at each stage of the chain seems to commit you to *de jure* sameness throughout the chain. The structure of this chaining relation is familiar from the literature on personal identity: at any given time, a mental file will bind a rich body of episodic memories and other attitudes derived from the immediate past, and at each stage your current attitudes seem to pertain *de jure* to the same topic as bound attitudes deriving from the recent past. Such chains of memory links among files demarcate historically extended representational traditions. Being connected by such chains, we suggest, is the core requirement for elements of thought to express the same concept.¹⁸

The story so far does not explain how concepts could be shared between different individuals. But it is very natural to extend our account to the interpersonal case. As we noted earlier: in parsing another person’s speech, you automatically hear their use of a familiar English expression as pertaining *de jure* to the very same topic you yourself associate with that expression. Even if you think your interlocutor is saying something false when she says “This action is morally wrong”, you’ll immediately understand her as thinking and talking about the very same topic you yourself associate with the term

¹⁸. Taking these chains of apparent *de jure* sameness at face value seems to commit one to the transitivity of *de jure* sameness relations. As many authors have emphasized, however, unrestricted transitivity of *de jure* sameness gives rise to difficulties for concept individuation. We address this issue in §6.
'morally wrong'. Just as you take your current memories to afford direct access to the topic of the past beliefs from which they derive, so you take your current literal understanding of an interlocutor's words to afford direct access to the topic of the beliefs from which they derive. Moreover, we take ourselves to be de jure samesaying not just with our immediate acquaintances but also with those with whom our acquaintances take themselves to be de jure samesaying. As in the intrapersonal case, chains of apparent de jure sameness can be used to construct historically extended representational traditions: chains of apparent de jure samesaying demarcate shared representational traditions within a given linguistic community. For two individuals to share the same concept, we suggest, their mental files must be systematically connected by chains of apparent de jure sameness relations that constitute a shared representational tradition.19

According to the connectedness model of concepts, token elements of thought express the same concept only if they are connected in the right way. The current section has focused on cashing out the relevant connections, starting from the core case of apparent de jure sameness of topic among co-conscious thoughts. We've sketched how the core condition on concept identity can be extended to cover dispositional relations among standing attitudes, diachronic relations with past thoughts, and interpersonal relations among the thoughts of different individuals. These relations demarcate shared representational traditions that reflect subjects' dispositions to treat thought elements as de jure pertaining to the same topic. Being connected in this way is merely a necessary condition for token thought elements to express the same concept. In §6, we will propose a second condition on concept identity designed to ensure that concepts really do represent a single topic. But before we do so, we must address the question of the reliability of de jure appearances.

19. For an account of how different speakers can coordinate their internal mental filing systems via natural language that is congenial to our approach, see (Cumming 2013a, 2013b).

5. Vindicating the appearances

On the template model, two token elements of thought express the same concept just in case each is associated with precisely the same topic-fixing criterion. This match in criteria underwrites the appearance of de jure sameness. And since the criterion associated with a token fixes its semantic value, sameness of criterion explains why such appearances are veridical.20 No similar explanation of the reliability of de jure appearances is available to proponents of the connectedness model. Although representational traditions may foster similarity in understanding over time and between subjects, the connectedness model does not require a precise match in topic-fixing criteria. This is a key virtue of the account, since apparent de jure sameness is not in fact beholden to any precise match. But the challenge for the connectedness model is to explain the veridicality of apparent de jure sameness given the scope of variation it tolerates in substantive understanding. The connectedness model must explain how the mechanisms that link us together into a single representational tradition allow us to reliably keep track of a topic in thought and talk.

To address this challenge, the connectedness model proposes a distinctive approach to the determination of semantic values. The connectedness model relies on a tradition-based determination theory: the fundamental units to which semantic values are assigned are not token elements of thought considered in isolation from each other, but rather an entire representational tradition (i.e., the entire set of token thought elements bound together by relations of apparent de jure sameness). On this approach, the determination theory seeks to assign a univocal semantic value to the tradition as a whole, taking into account the understanding, environment, and history of the entire diachronic and interpersonal tradition. Token elements of thought then inherit their semantic values from the traditions to which they belonging.

20. According to the modified neo-Fregean template model, sameness of criterion is merely a necessary condition on concept identity. Sameness of criterion then provides a conditional guarantee of sameness of topic, conditional on the empirical context remaining stable (see above §2).
are bound. On this account, then, we start by determining which tokens are bound together into the same representational tradition, and then we determine the semantic value of that tradition as a whole. Focusing on referential expressions, we can sum up the connectedness approach by calling it a “co-reference first, reference second” approach.

The “co-reference first” approach is implicit in many externalist theories of names and natural kind terms, which take historically extended representational traditions to be crucial to fixing the semantic values of token expressions. For instance, Kripke’s “causal chain” account of names starts with the fact that individual speakers are connected via networks of co-referential intentions. Your use of a name like ‘Gödel’ on a given occasion, Kripke argues, does not acquire its reference on the basis of facts about you considered in isolation from others in your linguistic community: your token use of ‘Gödel’ inherits its semantic value from that of the shared representational tradition as a whole (1980, 96–97). In effect, Kripke treats the shared representational tradition as the basic unit for semantic interpretation. Putnam’s so-called division of linguistic labor invokes a similar “co-reference first” approach: the ignorant individual’s use of a term like ‘elm’ inherits its semantic value from the linguistic community as a whole, in which some tree experts are delegated the “job” of distinguishing elms from non-elms (Putnam 1972, 143–146). Tyler Burge generalized this “co-reference first” approach from language to thought: it’s not just the reference of your word ‘arthritis’ that depends on your linguistic community; the content of the thoughts you express with that term also depends on these social facts (Burge 1979, 1982). In order to determine which property your ARTHRITIS thoughts represent, we must first identify the shared representational traditions to which those thoughts are connected.21

21. Although our account is similar in spirit to Kripke’s and Putnam’s accounts of names and natural kind terms, these authors focus on particular natural language expressions, and they do not offer any general determination theory for thought contents. In contrast, Burge does focus on thought contents and argues for a general thesis about the determination of thought content: that contents depend on external facts about the subject’s historical, social, and physical environment. However, Burge stops short of providing a general determination theory or theory of concept identity.

A complete determination theory, of course, requires more than just a specification of the basic units of interpretation. The determination theory must also explain how exactly these units acquire their semantic values: what are the principles that determine the semantic value of a given representational tradition?

This is a very different question from the question of the basic units of interpretation, but the difference may be obscured if one focuses exclusively on Kripke’s causal theory of proper names, where the very same links that are relevant to individuating the units of interpretation are also responsible for determining semantic values. Kripke sketches a quasi-anaphoric model of how the semantic value of a proper name is determined: each token use of a name simply inherits its reference from the earlier uses to which it is appropriately connected. The resulting chain of inherited reference stretches back to an initial baptismal event, where the reference was fixed by ostension or description. On this quasi-anaphoric model, the semantic value of any use of a name will depend on what was picked out by the original anchoring event, regardless of the subject’s current understanding or use of the name.

Although a quasi-anaphoric model correctly predicts the reference of many names, the model does not fit all names—and it certainly not all expressions. In particular, the idea that the semantic value of normative concepts is fixed by some original baptismal event is highly implausible. Different, more generally applicable principles for assigning semantic values are needed.

The correct principles, we suggest, are simply an idealization of a subject’s own reflective methods for refining her understanding of the precise subject matter of her words and thoughts. From the first-person perspective, object-level questions about what exactly it takes for something to count as Gödel, water, or morally wrong are equivalent to meta-level questions about the precise nature of the topic picked out by the corresponding words and thoughts. So we can use our
epistemic practices in deciding how to answer object-level questions as a guide to our meta-level commitments about the principles for determining precisely what it is we are thinking and talking about. In effect, we can build our semantic determination theory from the first-person reflective epistemology of the topic in question.

Elsewhere, we have suggested that the semantic values of our words and thoughts are determined by holistic rationalizing interpretation of the whole representational tradition (Schroeter forthcoming; Schroeter and Schroeter 2009, forthcoming-b). On this account, the input into deliberation about the correct semantic assignment is not restricted to a causal chain of co-referential intentions stretching back to an original baptismal event; instead deliberation takes into account the whole set of attitudes, dispositions, social practices, and environmental feedback loops associated with the historically and socially extended representational tradition. The correct assignment seeks to vindicate the most important aspects of the tradition as a whole. Our shared representational traditions thus demarcate the relevant input into reflection about the precise nature of the topic picked out. Principles of wide reflective equilibrium then determine the semantic value to be assigned.22

One might worry that holistic rationalizing interpretation will divorce the semantic value from the subject’s current substantive understanding of the topic in question. Shouldn’t a determination theory mirror a subject’s current understanding of a topic if it’s to capture the semantic values of her actual thoughts rather than the

22. More specifically, in rationalizing interpretation we seek to identify the point of the representational tradition: What are the most important practical and theoretical interests that have been subserved by the categorizing practices? The correct semantic assignment is the one that best vindicates the interests that have been most important to justifying and sustaining the representational tradition as a whole. Thus, the semantic value to be assigned to the token element of thought expressed by ‘water’ or ‘morally wrong’ on a given occasion is determined by wide reflective equilibrium about the point of the representational tradition to which that token is linked. For more details and an extended discussion of the case of water, see (Schroeter and Schroeter forthcoming-b).

semantic values of the thoughts she should have after philosophical or scientific reflection?

Clearly the semantic values of a subject’s thoughts must reflect an important aspect of the subject’s own epistemic perspective—else they would be arbitrary posits that would be useless in characterizing the subject’s states of mind. But the role of a semantic value is not simply to capture every quirk in an individual’s current substantive understanding of a topic. Semantic values determine the truth- and satisfaction-conditions for a subject’s attitudes, and they determine the logical relations among different attitudes. If one were to tie semantic values too tightly to current patterns of understanding, false belief would become impossible and direct logical relations across time and between subjects would vanish. So if semantic values are to play their role in assessing truth and validity, they must abstract from the fine-grained details of a subject’s current understanding.

We take this point to be well illustrated in literature on semantic externalism (e.g., Putnam 1970, 1972; Burge 1979, 1986; Schroeter 2008). But it’s worth emphasizing that the idea that the subject’s current epistemic perspective is fallible is not an arbitrary theoretical posit—it’s grounded in individual subjects’ own epistemic commitments. Ordinary individuals take themselves to be fallible about the precise nature of the topics they pick out in thought and talk: by your own best epistemic lights, you can be ignorant or mistaken about what it takes to be water, or Gödel, or morally wrong. And it’s natural to assume you can get closer to the truth about the nature of these topics through rational inquiry: revising your substantive understanding of water, or Gödel, or moral wrongness does not eo ipso change the topic under consideration. Thus, a subject’s own epistemic commitments suggest that semantic values mark the standards that she should be held accountable to in assessing the truth or satisfaction of her attitudes—not just the (incomplete and confused) standards that she takes herself to be accountable to.

This point is widely acknowledged in the metaethical literature: the topic picked out by ‘morally wrong’ does not simply mirror
the biases, idiosyncrasies, and confusions of your current moral understanding — some idealization is required to determine a topic that remains stable through changes in substantive commitments and that determines standards for the semantic correctness of your moral judgments.23

This sketch of our determination theory suffices to address our earlier challenge of explaining the reliability of apparent de jure sameness. In particular, we’re now in a position to explain how the fact that we’re linked to the same representational tradition allows us to reliably track sameness of topic, despite significant variation in individual subjects’ substantive understanding.

Take the moral-crisis scenario. Your beliefs and motivations have shifted in important ways, and yet you treat your past and present morally wrong thoughts as pertaining de jure to the same topic: it seems obvious and unquestionable to you that you’ve changed your mind about the very same topic you’ve been thinking about all along. Our tradition-based determination theory relies on such de jure appearances to demarcate the historically extended representational tradition to which a semantic value is assigned. Rationalizing interpretation then seeks to assign a univocal semantic value that vindicates the most important aspects of this representational tradition as a whole — taking into account both your past and present understanding and use. On this account, the appearance of de jure sameness will be veridical whenever a univocal rationalizing interpretation of the whole tradition is possible. Given the way you’re connected to the historical tradition, there is no need for you to preserve some core topic-fixing criteria at each moment throughout your life in order to ensure sameness of topic: the continuity of your representational tradition itself helps justify a common interpretation. Similar considerations apply at the interpersonal level.

We can summarize our solution to the reliability problem by saying that on the connectedness model we bootstrap our way to de jure sameness: the appearance of de jure sameness among token elements of thought helps make it the case that those elements really do have the same semantic value (Schroeter 2012).

This bootstrapping explanation of the reliability of apparent de jure sameness nicely captures the epistemic perspective of the subject herself. When you try to determine the precise nature of a topic — e.g., what it takes to be morally wrong, or Gödel, or an elm — you don’t (and shouldn’t) take the input into deliberation to be confined to your current substantive understanding. Given that your current attitudes are mostly dispositional states derived from past experience and from the testimony of others in your community, it makes sense to take this shared representational tradition into account when deciding what exactly you’re currently thinking about. After all, you aim to characterize the topic everybody has been thinking and talking about all along. Moreover, when you come to a verdict about the nature of that topic, you take that conclusion to apply not just to your own current thoughts considered in isolation, but to all of the appropriately related

23. Some recent work in metaethics and philosophy of language focuses on questions about the semantic values one should express with a given word, as opposed to the semantic values one does express with that word (Burgess and Plunkett 2013a, 2013b; Plunkett and Sundell 2013; Sundell 2012). For instance, in a dispute over the claim ‘Waterboarding is torture’, two speakers may be in complete agreement about the degree of pain and degradation inflicted but disagree about the relevant level that should constitute torture. If one identifies semantic values of a subject’s use of ‘torture’ via her current substantive understanding of the term, then we can characterize this as a disagreement over which semantic value should be associated with the term ‘torture’. But if the semantic value of an individual’s use of a word is determined by rationalizing interpretation of a shared representational tradition, then there will be less scope for this sort of metalinguistic disagreement. Presumably the two interlocutors in the ‘torture’ debate are linked to a shared representational tradition, in which we have used this term to mark unacceptable violations of a prisoner’s dignity, whether they be friend or foe. Given this determination theory, it will be much more difficult to establish that ‘Waterboarding is not torture’ conforms with the correct semantic value of proponents’ use of the term ‘torture’. It would not suffice, for instance, to simply appeal to U.S. Justice Department policy in defining ‘torture’ as “inflicting pain rising to the level of death, organ failure or permanent impairment” (cf. Plunkett and Sundell 2013, 19). Of course, there would still be scope for metalinguistic disagreement on our account, particularly in cases of obvious context-dependent natural language expressions like ‘tall’. But rationalizing interpretation will not favor divergent semantic assignments to individuals’ use of an expression simply on the basis of a divergence in their actual patterns of understanding.
thoughts of your past self and of others in your community. From
your own first-person perspective, the relevant unit of interpretation
is not your current understanding considered in isolation, but the
whole representational tradition to which you are linked by chains of
apparent de jure sameness relations.

Our determination theory thus respects interpretive charity. Our
account does not impose an arbitrary external semantic value on a
participant in a tradition. When we take the subject’s thoughts to
inherit their semantic value from a shared representational tradition,
we are simply vindicating the subject’s own epistemic commitment to
that tradition.

6. Disambiguation and concept identity

We have shown how the appearance of de jure sameness can be
reliable on our connectedness account. However, we don’t claim
that such appearances are infallible. In fact, it’s highly implausible to
suppose de jure appearances are always accurate. When you say ‘Pears
are delicious’, I may hear you as saying ‘Bears are delicious’, thereby
taking your thoughts to pertain de jure to the same topic as my own
thoughts about bears. Clearly this is a case of mistaken appearances:
your thought is not about the same topic as my bear thoughts.

Consider also the British and American uses of ‘corn’. British
speakers normally take ‘corn’ to apply to any type of grain, whereas
Americans take ‘corn’ to apply exclusively to maize. Even so, it’s easy
to imagine contexts in which British and American speakers would fail
to recognize this divergence and hear each other as de jure thinking
about the same topic. Intuitively, the appearance of de jure sameness
is mistaken in this case: despite apparent de jure sameness, the two

In the bear/pear case, the appearance of de jure sameness may not be suf-
ciently stable to forge a link in a shared representational tradition. Your (ap-
parent) claim will strike me as strange, so I’m likely to seek further clarifica-
tion about your experience with eating bears. In the normal course of things,
our ensuing conversation will quickly lead to a correction of my original
misunderstanding. But not all mistakes about de jure sameness are so easily
discernible — and the more difficult cases call for an enrichment of our deter-
mination theory.

linguistic subgroups are thinking and talking about two distinct kinds.
Moreover, such mistakes also occur through gradual referential drift.
No doubt the restrictive American usage evolved gradually from the
more liberal British usage, as Americans came to use the generic
predominantly for the indigenous species. And it’s plausible that a
chain of apparent de jure sameness relations connect contemporary
American uses of the term with 17th-century British uses. Intuitively,
there was a gradual undetected shift in the semantic value associated
with Americans’ use of the term ‘corn’.

Such examples suggest that being connected by a shared
representational tradition (by chains of apparent de jure sameness)
does not suffice for sameness of semantic value. If a tradition-based
determination theory were to treat the appearance of de jure sameness
as infallible, then it would make undetected shifts of reference over
time impossible. But this seems to flout our reflective interpretive
verdicts about such cases.

We can accommodate these reflective verdicts within our ‘co-
reference first’ determination theory, by introducing a second
step in the assignment of semantic values to representational
traditions — disambiguation. In the first instance, rationalizing
interpretation seeks a univocal semantic value for the entire
representational tradition associated with ‘corn’. However, when we
consider the British/American tradition as a whole, rationalizing
interpretation yields two equally good referential candidates: grain
and maize. If interpretation were to stop at this point, the conclusion
might be that contemporary American and British speakers all share
indeterminate reference: the semantic value of all of our corn thoughts
is undecided as between grain and maize. But commonsense

24. For similar examples in the literature, see Gareth Evans’ ‘Madagascar’ case
(1973) and Tyler Burge’s ‘slow-switching’ case (1988).

25. This indeterminacy should not be confused with ambiguity. In the case of
ambiguous expressions like ‘bank’, some token uses of the term refer to river
banks, and other tokens refer to financial institutions. But on the indetermi-
nacy interpretation of ‘corn’, each token use indeterminately refers to corn
and also indeterminately refers to maize. Nor should indeterminate reference
interpretation doesn’t (and shouldn’t) stop here. We can vindicate our presumption that corn is a single determinate kind of stuff by disambiguating our representational tradition: we look for a way of partitioning our initial representational tradition with ‘corn’ such that each partition can be assigned a univocal semantic value. In the case at hand, it’s clear how to make this partition: British users of the term ‘corn’ refer univocally to grain, and after a certain point in history, American users refer univocally to maize.

This enrichment of the determination theory allows us to specify identity conditions of concepts.

**Concept identity:** two token elements of thought express the same concept if and only if:

i. (connectedness) they are connected to each other via a shared representational tradition (constituted by apparent *de jure* sameness relations); and

ii. (congruence) the understanding and historical context associated with them does not diverge so radically as to undermine a univocal interpretation.

The connectedness constraint says that two thoughts cannot share the same concept unless they are causally related in the relevant way. The congruence constraint allows for the possibility that the appearance of *de jure* sameness can be misleading. As in the ‘corn’ case, there may be referential drift or undetected referential ambiguity within a given representational tradition. When there is too much variation in different speakers’ understanding of what they initially be confused with reference to a disjunctive kind grain-or-maize: on the disjunctive interpretation, Americans would be right to call barley ‘corn’. On the indeterminacy interpretation, as Hartry Field puts it, your use of the term ‘partially denotes’ grain and ‘partially denotes’ maize: see (Field 1972) for a defense of indeterminate interpretations, together with a supervaluation semantics for deciding on the overall acceptability of particular uses of the term. For a response to Field that advocates a disambiguation approach similar to ours, see (Kitcher 1978).

took to be a single topic, as in the corn and bear cases, the initial presumption of *de jure* sameness can be defeated.27

Similarly, in the case of normative concepts like *morally wrong*, our theory says that the presumption of *de jure* sameness is a reliable but fallible guide to concept identity. It could turn out that the person you’re arguing with belongs to a subgroup whose use of the term ‘morally wrong’ is so divergent from that of the rest of the community that it should be assigned a distinct semantic value. Or your own understanding and application of ‘morally wrong’ may change so radically over time that your current beliefs should be assigned a semantic value distinct from that of your past beliefs. But such cases of undetected conceptual change will be the exception to the rule.

We’ve motivated disambiguation in the determination theory and the congruence constraint on concept identity by appealing to first-person reflective verdicts about the nature of the topic picked out by our words and thoughts. But congruence is also crucial to answering an important line of objection to essentially relational accounts like ours, which make concept identity (or meaning identity) depend on relations of apparent *de jure* sameness (Fine 2007; Pinillos 2011).

The aim of our relational theory of concept identity is to show how a direct logical guarantee of sameness of topic, which concepts are supposed to provide, can be built from the subjective appearance of *de jure* sameness. By itself the relation of apparent *de jure* sameness seems unsuited to securing direct logical relations. Even if the appearance of *de jure* sameness is a reliable guide to the semantic facts, that’s not enough to secure a logical guarantee. And long chains of apparent *de jure* sameness aren’t always reliable. Moreover, an equivalence relation

27. Exactly what types of variation in the understanding and history associated with token thoughts would defeat an initial presumption of *de jure* sameness? We will not attempt to provide a full account of such defeating conditions here. It suffices for our purposes to leave the congruence condition at an intuitive level, appealing to commonsense interpretive intuitions. The precise considerations that ground defeating conditions will be open-ended and will vary depending on the topic in question. For further discussion of the principles of disambiguation and how disambiguation affects concept identity, see (Schroeter 2008, 2012).
like concept identity must be transitive. But if we take relations of apparent de jure sameness to determine concept identity, there will be a failure of transitivity: contemporary speakers in America and Britain who use ‘corn’ are linked by chains of apparent de jure sameness to the same historical tradition, and so would both count as co-referencing with past users; but they do not take themselves to co-refer with each other.

Adding the congruence constraint to the connectedness constraint allows us to respond to these objections in a principled way. The congruence constraint corrects the failure of transitivity by partitioning representational traditions so as to respect our reflective interpretive commitments about sameness of semantic value. On our account, for two token elements of thought to express the same concept is for them to belong to the same disambiguated representational tradition. All tokens within the same disambiguated representational tradition will share the same semantic value. So sameness of concept entails sameness of semantic value on the connectedness model of concept identity. Moreover, this relation of belonging to the same disambiguated representational tradition respects transitivity. In effect, disambiguation imposes a strict identity relation on the non-transitive relation of apparent de jure sameness. The resulting relation, we submit, respects the direct logical relations of guaranteed sameness of content to which we should hold our thoughts accountable.28

7. New resources for realism

In the last three sections, we have articulated a relational alternative to the standard template model of concepts. Our connectedness model, we’ve suggested, provides a better explanation of the appearance of de jure sameness and of the open-ended variation in substantive understanding characteristic of normative concepts. This model, we believe, reshapes debates in metaethics. Most obviously, bootstrapping makes de jure sameness much easier to achieve on this model than on the template model: the mere fact that we take ourselves to be thinking and talking de jure about the same topic helps make it the case that we are in fact doing so. The connectedness model also has important consequences for the debate over the nature of the topic picked out by normative concepts. In this section, we explain how moving from the template to the connectedness model of concepts makes the prospects for traditional, context-invariant normative realism significantly brighter.

A general model of concepts, considered in itself, is neutral about the semantic values associated with normative concepts. For all we have said so far, our shared representational tradition with ‘morally wrong’ could pick out a single property (as traditional moral realists maintain).

28. Many theorists will be tempted to press a different sort of objection to a connectedness model. The connectedness model makes real causal-historical connections a necessary condition for concept identity. In this respect, our relational account of concept identity is similar to relational accounts of personal identity or biological species identity: just as a perfect intrinsic duplicate of you living on a distant, causally isolated, intrinsic duplicate planet doesn’t count as the same individual person as you and doesn’t belong to the same biological species as you, that duplicate won’t share the same concepts as you. Many metaethicists will find this result counterintuitive. Moreover, the influential Moral Twin Earth argument treats such intuitions about sameness of meaning across distinct linguistic communities as an important constraint on an acceptable theory of concept identity (Horgan and Timmons forthcoming-a). In a nutshell, to vindicate normative thought and talk, a theory of concepts must ground de jure sameness of topic over time and between subjects despite open-ended variation in subjects’ substantive understanding of that topic. In this paper, we’ve argued that the mechanisms that are best suited to playing this core theoretical role require real-world causal connections among token thoughts, rather than a precise match in topic-fixing criteria. Our theory can, of course, acknowledge that perfect intrinsic duplicates would share the dispositional states that constitute their substantive understanding of the various topics they think about. But without the right sort of causal connection, these duplicates would lack the very feature that explains how two ordinary speakers of the same language (or a single thinker at different times) manage to coordinate de jure on the same topic despite variation in beliefs and motivation. What’s lacking in the duplicate case is precisely what’s crucial to securing ordinary semantic coordination across time and between subjects: a shared representational tradition.
But it’s also compatible with our model that our moral concepts could have an indexical semantic value that picks out different properties on different occasions (contextualism), or an assessor-relative semantic value that assigns properties relative to an audience’s standards of assessment (relativism). Alternatively, moral concepts could serve to express hypothetical plans or systems of preferences (expressivism); they could pick out a purely intentional or necessarily uninstantiated property (error theory or fictionalism); or they could have no semantic value whatsoever (because the representational tradition is irredeemably confused).29

Despite the fact that they don’t directly pronounce on semantic values, the determination theories associated with the template and connectedness models will have important consequences for how to adjudicate the debate between these different metaethical positions.

The template model imposes a straitjacket on interpretation: the correct semantic assignment is determined by the topic-fixing criterion each individual subject independently associates with a token element of thought at a given time. This individualistic approach to the determination of semantic values seems to put traditional context-invariant realism about the normative domain out of reach. Normative realism would require token elements of thought to be independently associated with the very same topic-fixing criterion, and this criterion would have to be rich enough to single out the very same property for everyone. But given the flexibility characteristic of normative concepts — virtually any variation in beliefs and motivations seems compatible with apparent de jure sameness of topic — it’s hard to see how the template model can provide enough shared topic-fixing assumptions to determine the same property as semantic value for all.

The connectedness model’s tradition-based approach to the determination theory removes this individualistic straightjacket on interpretation. The immediate appearance of de jure sameness that links subjects’ thoughts to an historically extended representational tradition, we’ve suggested, will and should dispose them to hold their understanding accountable to the best interpretation of that tradition as a whole. This commitment to a shared representational tradition significantly increases the prospects for realism in the normative domain: it enriches the relevant input into interpretation and it mandates the search for a mutually acceptable interpretation.30

First consider the inputs. On the connectedness model, the inputs into rationalizing interpretation will be facts about the shared representational tradition as a whole. This means that the relevant inputs will be the same for all participants in a given tradition. Whereas the template model implies that the correct semantic assignment is determined by a subset of what’s currently “in the head”

29. It is worth noting as well that our model is also neutral about how precisely to demarcate normative concepts from non-normative concepts. To a first approximation, we could characterize normative concepts as those that are associated with a characteristic action-guiding role. But one might locate the relevant action-guiding role at different levels in one’s account. On one approach, normative concepts are demarcated at the level of their semantic values: e.g., normative semantic values might be expressive of a certain pro-attitude (Gibbard 2003), or normative semantic values might be identified with properties that are metaphysically intrinsically action-guiding (McDowell 1998, 83; Enoch 2011). Alternatively, normative concepts might be demarcated at the conceptual level by appealing to representational traditions and associated understanding: e.g., perhaps there must just be a critical mass of participants in a representational tradition who accept a certain action-guiding role in order for it to count as expressing a normative concept (for an account in this spirit, see Blackburn 1998). We favor a version of this last approach (Schroeter and Schroeter forthcoming-c), but our account of concepts is consistent with different views.

30. NB: The connectedness model can grant that some representational traditions may exhibit some template-like features. Perhaps we would not count someone as a legitimate interlocutor about bachelors unless they relied on a specific topic-fixing criterion for the term ‘bachelor’. If so, rationalizing interpretation might take the property picked out to be determined by that conventionally mandated criterion. But the connectedness model is not committed to this conclusion, since there may be other aspects of our shared representational practice that affect rationalizing interpretation. Arguably this was the case with our shared representational practice with ‘atom’ at a certain point: although scientists accepted a standard topic-fixing definition (‘indivisible particle’), their systematic practice of using the term to identify and learn about particles supported an interpretation that falsified this topic-fixing criterion.
of the individual, the connectedness model implies that historical and social facts about which the subject may be ignorant or mistaken are equally relevant to justifying a semantic assignment. This expansion of the input into interpretation means the realist no longer needs to identify some invariant core understanding, shared by all, that suffices to single out a determinate property. Since the cognitive and motivational dispositions relevant to grounding a correct interpretation are distributed among participants in the tradition, it may be that no individual actually grasps all aspects of understanding relevant to determining the semantic value of her representational tradition. By increasing the substantive input into interpretation and standardizing these inputs among subjects, this anti-individualist approach brightens the prospects for finding a univocal realist interpretation — and indeed for finding any univocal interpretation at all.

Next, consider the interpretive methods. The correct interpretive methods, we suggested, must reflect the subject’s own reflective dispositions to refine her understanding of the precise nature of the topic. So one might worry that the problem of divergent understanding will simply reappear at the level of subjects’ epistemic methods. We should not expect participants in a representational tradition to start out with precisely matching dispositions to evaluate inputs into reflection: their methodological dispositions can vary just as their substantive understanding varies. So even if subjects were presented with the very same information about their shared representational tradition with the term ‘morally wrong’, they may be disposed to weigh this evidence in different ways. After ideal reflective equilibrium based on their current methodological dispositions, different subjects could take the same input to support divergent verdicts about precisely what it takes to be morally wrong. If so, the anti-realist may argue, subjects’ initial presumption of de jure co-reference would be defeated.

The connectedness model rejects this individualistic approach to interpretation. The inputs into interpretation depend on what other participants in the tradition have been thinking and doing with their concepts. Likewise, the methods of interpretation depend on what other similarly committed participants would find plausible. Individuals’ commitment to a univocal semantic interpretation of their shared representational tradition should lead them to seek mutually acceptable methods of interpretation. So holding oneself accountable to shared representational traditions will exert rational pressure to revise one’s interpretive dispositions in such a way as to bring them into harmony with the group.

The idea of mutually acceptable interpretive methods should be less controversial for metaethicists than it may initially seem. When proponents of expressivism, relativism, and traditional context-invariant realism engage in debate about the semantic value of normative terms, for instance, they each try to show that their own preferred interpretation vindicates the most important aspects of our communal representational tradition. Each side seeks to show that their interpretation is justifiable, not just from their own idiosyncratic perspective, but from the point of view of all participants. Moreover, such arguments don’t typically fall on deaf ears: although you may not agree with your dialectical opponent, you can usually acknowledge the relevance and at least prima facie cogency of the reasons they cite in favor of their position. The very fact that metaethicists engage each other in this way, seeking intersubjectively valid reasons in favor of their preferred interpretations, suggests they are committed to mutually acceptable interpretive methods.\footnote{31 For a defense of the idea that individual members of a linguistic community do and should take themselves to be beholden to communally acceptable epistemic standards for specifying the topic of their own words and thoughts, see (Burge 1986, 1989) and (Petit 1993, ch. 4; 2002, 20–24).}

Consider a case of disagreement about how to best interpret our shared representational tradition with racial terms like ‘black’ and ‘white’. Some theorists hold that our dominant historical interest at stake in keeping track of such categories has been to track alleged biological differences between populations with African or European ancestry. If this is right, one may conclude that the racial categories are empty: there are no blacks or whites, since there are no robust
biological differences of the relevant kind. Other theorists hold that
the dominant interests suberved by our shared representational
tradition is not in tracing biological differences but in regimenting
and enforcing a normative system of racial subordination of one
group to the other. If this is right, one may conclude that racial
categories are social constructs that play a certain (objectionable)
normative role in guiding action and emotion. The mere fact
that theorists disagree about the best interpretation of a shared
representational tradition does not entail that they are talking past
each other. Indeed, both sides of the debate take themselves to be
making a claim about the shared representational tradition, and
each side seeks to identify mutually acceptable reasons in favor of
its favored interpretation. It’s perfectly coherent to suppose that one
side of the debate may be right about the dominant interests at stake
in the shared representational tradition.32

We believe this is the correct verdict about how we do and should
adjudicate questions about the semantic values of our words and
thoughts: we seek mutually acceptable interpretations of our shared
representational traditions. The connectedness model vindicates
this verdict by building an anti-individualist epistemology into
the determination theory. The initial inputs into reflection about
the nature of a particular topic are determined by interpersonal
representational traditions, and the relevant methods for evaluating
those inputs must also be interpersonally acceptable. In effect, our
determination theory is wide reflective equilibrium writ large and
stripped of its individualistic bias: it aims at finding an interpersonally

32. Of course, it could turn out that the considerations on each side of the arg-
ument are equally balanced: the different background interests have been
equally important to justifying and sustaining our representational tradition,
and those interests ground competing interpretations of the semantic values
expressed by our use of racial terms like ‘black’ and ‘white’. In that case, ra-
tionalizing interpretation calls for disambiguation: both sides of the debate
should admit that we’ve been implicitly treating two distinct concepts (BLACK
and BLACK,) as if they were one, and we should distinguish these different
conceptual practices. For more details, see (Schroeter 2007, 2008).

justifiable interpretation that resolves tensions within the whole
communal representational tradition.

This anti-individualistic determination theory has important
consequences for the debate over normative realism. Many theorists
take persistent normative disagreement to be the central problem
facing traditional, context-invariant normative realism: how can
different subjects be thinking about precisely the same property if
they would, on reflection, come to divergent verdicts about which
actions count as morally wrong? As we noted earlier, this problem
seems especially daunting if one accepts the template model of
concept identity: any shared topic-fixing template, it seems, will be
too minimal to ensure that different subjects will converge on the
same instantiation conditions for normative properties. Moreover,
the template model requires convergence to be independent:
relative to the same context, each individual must be disposed to
arrive independently at precisely matching verdicts. The variability
of subjects’ substantive understanding makes such independent
convergence unlikely.

By rejecting the individualistic bias of the template model, the
connectedness model provides new resources for vindicating
realism. On the connectedness model, assigning the same property
as the semantic value of the concept morally wrong does not
require different individuals to come to matching verdicts about
the property picked out on the basis of individualistic reflection alone.
Individualistic reflection leaves out two key factors that legitimately
affect subjects’ ideal verdicts about the precise nature of the topic: (i)
the substantive understanding and history of other participants in the
shared representational tradition, and (ii) the interpretive methods
that would be acceptable to other participants. Convergence on a
single property is more plausible if reflective equilibrium is conceived
of as an essentially social project, where different individuals are all
committed to finding a mutually justifiable interpretation of a shared
representational tradition.
Anti-realists might object at this point. Getting rid of the individualistic bias of the template model will make it easier to justify a univocal semantic assignment despite variation in individuals’ substantive understanding and methodological dispositions. That should be good news for everyone, since ensuring convergence in reflective verdicts despite variability in initial understanding is a serious challenge facing any univocal semantic interpretation. But why suppose that the correct semantic assignment will be a single property? Why couldn’t rationalizing interpretation of our shared representational tradition with ‘morally wrong’ assign, say, a purely expressivist semantic value or an indexical function from contexts to properties? Why, in other words, should we suppose the connectedness model is good news for normative realists in particular?

The short answer is that rationalizing interpretation is biased in favor of a traditional, context-invariant realist interpretation of normative concepts because these realist assumptions are central to our normative thought and talk. Insofar as an interpretation can vindicate these realist assumptions, it should be preferred to semantic interpretations that cannot.

Realist assumptions are not just a superficial aspect of our use of normative concepts. Of course, all metaethicists agree that normative talk exhibits the surface syntax characteristic of paradigmatic representational discourse. But there is a deeper point to be made about the nature of normative thought. Normative concepts seem predominantly directed toward stable categorization: we categorize actions as morally right or wrong; we accumulate a body of general beliefs and dispositions pertaining to these topics; we reason and theorize about the principles governing categorization; we debate the nature of that category with others; and we correct, refine, and precisify prior categorizations in light of this reflective theorizing. In short, our categorizing practice with a normative concept like morally wrong treats the topic picked out as a stable, judgment-independent property of actions—a property we can learn about through armchair theorizing, empirical inquiry, testimony, or argument. In these respects, our practice of judging actions to be right or wrong resembles the paradigmatically realist practices of classifying something as gold, or arthritis, or Gödel.

This basic categorizing function contrasts sharply with the structure of thought associated with an expressive use of ‘Ouch!’, a contrastive use of ‘but’, or an imperatival use of ‘Help!’ And the practice of seeking to accumulate a stable and coherent body of information about a topic also contrasts with the context-dependent categorizing practices associated with terms like ‘fashionable’, or ‘legal’, or ‘tall’, where it’s mutually obvious to all speakers that categorizations are relative to contingent social conventions or conversational interests. In contrast with these domains, moral disagreement does not tend to disappear when it’s pointed out that the parties to the dispute have different interests or belong to communities who accept different social standards. On the face of it, then, our practice with central normative concepts like morally wrong seems to presuppose that a single property is picked out by all competent subjects independently of their idiosyncratic interests or circumstances.

This presumption of stable reference over time and between subjects plays a foundational role in normative epistemology: it explains and justifies virtuous epistemic practices. If you assume that your deployments of a normative concept always pick out the very same property, then there will be direct pressure to integrate your conflicting evaluative dispositions into a coherent whole. And if you assume that we are all co-referring on normative questions, there will be direct epistemic pressure to take others’ testimony at face value, to give credence to their normative reasoning, and to work toward resolving interpersonal disagreements through distinctions, justifications, and rationales that are intersubjectively acceptable.

Note that context-invariant realism requires that the same judgment-independent object, kind, or property is picked out by all deployments of the same concept. It does not make any claim about the special metaphysical status of this feature: e.g., it does not require that in order to count as real, an object, kind, or property must be metaphysically fundamental, or causally explanatory, or play some wide cosmological role.
Without such a presumption of stable co-reference, it’s not obvious that these virtuous epistemic practices would be warranted, at least in their full generality.

Given the centrality of the presumption of a stable reference to our normative thinking, our determination theory suggests that traditional context-invariant realism should be the preferred default semantic interpretation. Rationalizing interpretation, after all, seeks to vindicate the most important aspects of our shared representational traditions. A vindicating semantic interpretation of our use of normative concepts such as **morally wrong** is not just a matter of making certain commonly accepted sentences come out true and logical relations among commonly accepted forms of argument come out valid. Crucially, a rationalizing interpretation should seek to vindicate our stable categorizing dispositions and our epistemic practices of self-correction associated with these concepts—and to do so in a **non-ad hoc** way. Interpreting these normative concepts as picking out stable properties, we submit, is the simplest and most charitable way of vindicating these foundational aspects of our representational tradition with central normative concepts.34

This is not to say normative realism is guaranteed to be correct. It may turn out that there are simply too many divergent interests at stake in our practice with normative terms to ground any univocal interpretation. In that case, rationalizing interpretation should look for some other way of making sense of our shared representational tradition. But from the point of view of our determination theory,

34. One might think that the quasi-realist program, if successful, could provide an equally good vindication of our representational traditions with normative concepts (Blackburn 1988). Starting from a purely expressive semantic value, the quasi-realist seeks to explain why it is semantically correct for subjects to talk as if there were a single stable property picked out, construing certain claims as ‘true’ and certain argument forms as ‘valid’. But unlike the realist, the quasi-realist needn’t posit any stable property as the semantic value picked out by every token use of a normative concept like **morally wrong**. Such an interpretation, we submit, would not fully vindicate subjects’ stable categorizing and theorizing practices. For a critique along these lines of one quasi-realist proposal (Gibbard 2003), see (Schroeter and Schroeter 2005).

anti-realist interpretations will always be second best: justifying an anti-realist interpretation depends on showing that the realist option is not viable. The implicit hold of the template model on metaethical theorizing, we believe, has led many metaethicists to settle too quickly for second best.35

References


35. Many thanks to David Enoch, Tristram McPherson, Mark van Roojen, Pekka Väyrynen, and two anonymous referees for this journal for helpful comments on earlier drafts.


------. forthcoming-c. Why Go Hybrid? A Cognitivist Alternative to Hybrid Theories of Normative Judgment. In *Having It Both Ways:


