UNDERSTANDING THE INTERNALISM-EXTERNALISM DEBATE:

WHAT IS THE BOUNDARY OF THE THINKER?

Brie Gertler

Forthcoming in *Philosophical Perspectives 2012: Philosophy of Mind*.

Since the work of Burge, Davidson, Kripke, and Putnam in the 1970's, philosophers of language and mind have engaged in extensive debate over the following question: Do mental content properties—such as *thinking that water quenches thirst*—supervene on properties intrinsic to the thinker? To answer affirmatively is to endorse internalism (or "individualism"); a negative answer is an expression of externalism.

There is no consensus about the correct answer to this question; a 2009 survey indicates that a bare majority of philosophers now characterize themselves as externalists.[1] The recent literature on this topic largely focuses on the implications of externalism and internalism. There is no consensus here either. Philosophers are sharply divided as to whether externalism is compatible with privileged access to one's own thoughts; whether externalism implies that we can achieve knowledge of the external world from the armchair; whether internalism is compatible with physicalism about the mental; and whether internalism implies that thoughts are incommunicable.

Disagreements are philosophers' stock in trade. But the disputes just mentioned have proven exceptionally intractable. The culprit, I think, is an ambiguity in the terms "externalism" and "internalism", which they inherit from an ambiguity in the notion of "intrinsic to the thinker" operative in these disputes. As employed in the debate over mental content, "externalism" and "internalism" are associated with a shifting set of claims encompassing a heterogeneous array of topics; these include the organism's contribution to thought contents, links between the individual and her community, the epistemic availability of thoughts, and relations between phenomenal character and intentional content.

I will argue that this ambiguity is ineliminable. Any way of explicating "intrinsic to the thinker" will clash with the usual taxonomy of leading externalist and internalist views, or construe these positions as involving claims that are standardly regarded as orthogonal to them—and, in some cases, explicitly rejected by their most prominent exponents.[2] The moral is stark. The sense that there is a substantive, defining commitment of externalism or internalism—even one that is vague or underspecified—is illusory. There is no univocal thesis of externalism or internalism.

The ambiguity of "externalism" and "internalism" helps to explain why contributors to this literature often seem to be arguing at cross-purposes, disagreeing about the truth and implications of externalism and internalism, and about the nature of the evidence that could resolve these disputes. Now this ambiguity would not be too worrisome if its effects were confined to disputes about mental content. But because the claims associated with externalism and internalism cover a diverse range of topics, philosophers routinely invoke externalism or internalism (or purported implications thereof) in evaluating a range of other questions—in the

philosophy of language, epistemology, and the philosophy of mind. These include: Does the meaning of an utterance correspond to elements understood by the speaker? Do thinkers generally enjoy privileged access to their own mental states? Can we know contingent facts about the external world through introspection and *a priori* reasoning? Does phenomenal character supervene on intentional content, or vice versa? Can content be naturalized? The ambiguity endemic to discussions of externalism and internalism thus threatens progress on a broad spectrum of philosophical questions.

I begin by arguing, in Section 1, that an adequate explication of "externalism" or "internalism" must employ a criterion of "intrinsic to the thinker". The next three sections evaluate candidate criteria. Section 2 discusses the most familiar type of criteria, which explicate this notion in physical or spatial terms. Section 3 examines a recently proposed epistemic criterion. Section 4 considers the idea that what is intrinsic to thinkers are thoughts themselves—the bearers of content—which may not exhaust the factors determining content. Each of these candidates fails. Each commits externalists or internalists to positions that are strictly optional, according to the ordinary understanding of these views; conflicts with established classifications of particular views as externalist or internalist; or lacks the informativeness needed to illuminate this debate. Section 5 argues that other possible criteria of "intrinsic to the thinker" will likely share these inadequacies.

The debate about mental content, as it is currently framed, cannot be salvaged. I conclude by briefly suggesting more profitable uses for the philosophical energies conserved by abandoning this debate.

1. "Intrinsic to the thinker"

Internalism and externalism are standardly expressed as follows.

> **(I)** Thought contents always supervene on properties intrinsic to the thinker.
>
> **(E)** Thought contents do not always supervene on properties intrinsic to the thinker.[3]

My plan is to demonstrate that there is no univocal thesis of externalism or internalism, by showing that (I) and (E) are irremediably ambiguous: no way of explicating "intrinsic to the thinker" will cash out these statements in a way that makes sense of the existing debate.

Someone could object to my project by noting that the term "intrinsic to the thinker" is not present in every formulation of externalism and internalism. But while this term is not crucial, the distinction it marks—between properties intrinsic to the thinker and properties extrinsic to her—will be invoked in any plausible formulation of these positions. To see this, consider Kirk Ludwig's particularly clear formulation of externalism, which does not use the term "intrinsic to the thinker".

> The externalist thesis is, in short, that content properties are in part relational properties. A property *P* is a relational property just in case, necessarily, for any object *O*, if *O* has *P*, then there is an *X* such that *X* is (i) not an abstract object and (ii) *X* is not identical to *O* or to any part of *O*. (Ludwig 1993, 251)

On this interpretation, the content property *thinking that water quenches thirst* satisfies the externalist thesis *iff* my having a thought with that content entails the existence of some concrete entity (or other) distinct from myself. But arguably, X is a distinct concrete entity just in case *being such that X exists* is not intrinsic to me. So the notion of properties intrinsic to the thinker is implicit in this formulation of externalism. And this is how it should be, since—as the labels "externalism" and "internalism" indicate—these positions' defining theses make crucial use of the notion of features instantiated within (or outside) the thinking subject.

A more promising objection to my project denies that understanding "intrinsic to the thinker" requires identifying a *criterion*. This objection might take the following form.

> The search for a criterion here is misguided. Surely factors standardly regarded as internal, such as brain states, occur within the thinker. And those that serve as examples of external factors, such as the presence of H_2O in the environment, and the use of "arthritis" by community experts, are external to the thinker. We should treat *being in brain state B* as a paradigm case of an intrinsic property; and we should treat *inhabiting an environment in which the watery stuff is H_2O* and *belonging to a community where experts use "arthritis" to refer to a joint disease* as paradigm cases of non-intrinsic properties. While the status of other properties may be less clear, these examples illuminate what "intrinsic to the thinker" means. We understand this term well enough, even if we are unable to specify a precise *criterion*.

Here is my response. Although the properties mentioned appear to be clear examples of intrinsic and non-intrinsic properties, it is conceivable that the best way to understand "intrinsic" and "non-intrinsic" will reclassify one or more of them. All else being equal, a way of drawing this distinction should count *being in brain state B* as an intrinsic property; and it should count *inhabiting an environment in which the watery stuff is H_2O*, and *belonging to a community where experts use "arthritis" to refer to a joint disease*, as non-intrinsic. But we cannot assume that an understanding of "intrinsic to the thinker" that remains loyal to widespread perceptions of the basic commitments of externalism and internalism will neatly match our intuitions about which properties fit this description. In other words, all else may *not* be equal. So even these seemingly clear instances of intrinsic and non-intrinsic properties are open to reclassification.

This last point is controversial. To see why these instances of (apparently) intrinsic and non-intrinsic properties should not be treated as sacrosanct, recall that a seminal externalist argument (Putnam 1975) uses *inhabiting an environment in which the watery stuff is H_2O* as an example of an *external* property. Many commentators have noted that this property could be regarded as intrinsic to the thinker, since humans are partly composed of H_2O. This complication is usually brushed off with the observation that H_2O is an unfortunate example. But it carries a valuable lesson: particular examples of properties claimed to be intrinsic (or non-intrinsic) may sit uneasily with the *intentions* guiding the use of these terms. And sometimes, as in the H_2O case, the intention is more important than the particular example. As we will see below, one philosopher has proposed that loyalty to the relevant referential intentions will count *being in brain state B* as a non-intrinsic property (Farkas 2003). Regardless of that proposal's ultimate merits, it seems reasonable not to foreclose, from the outset, the possibility that seemingly paradigmatic cases of intrinsic or non-intrinsic properties could conceivably be reclassified.

Here is another way to put this point. The standard examples of intrinsic and non-intrinsic properties are not genuine paradigms, in the strict sense of "paradigm" that is at work in paradigm case arguments. (In that strict sense of "paradigm", a paradigm case of an *F* cannot fail to be an *F*.) Rather, as the H_2O example illustrates, these examples are chosen because it is assumed that they qualify as intrinsic (or non-intrinsic) according to some principled, albeit unarticulated, conception of the boundary of the thinker: a boundary dividing factors within the thinker from those outside her. Making this implicit conception *explicit* requires identifying the criterion of "intrinsic to the thinker" that operates behind the scenes in the externalism-internalism debate.[4]

In attempting to unpack the notion of "intrinsic to the thinker" operative in this debate, we must balance a variety of factors. We must accord some weight to intuitions about how to categorize specific properties. But such intuitions may not carry the day, for they may conflict with standard classifications of particular views as internalist or externalist, or with widely shared assumptions about the commitments of internalism and externalism.

2. The Spatial Approach

Externalists often express their view by denying that thought content supervenes on properties instantiated within the subject's skin, brain, or head. The idea here is that the skin, brain, or head constitutes the outer spatial limits of the individual, conceived as an organism, or of that part of the individual directly involved in thought. It is easy to see why contributors to this debate have not felt it necessary to choose between these various biological boundaries. The central externalist claim is that some thought contents metaphysically depend on features of the physical environment or social practices, and these are presumed to fall squarely outside the human organism. (For convenience, I will use "the skin" to represent biological boundaries more generally.)

Expressions of externalism commonly assume that the supervenience base spatially located within the skin is constituted by physical properties. For instance, the normal test case for externalist claims are imaginary twins, characterized as "molecule-for-molecule duplicates". So the most familiar formulation of externalism relies on a spatiophysical construal of "intrinsic to the thinker": it interprets externalism as the thesis that thought contents can differ between individuals who are precisely alike as regards the physical properties instantiated within the space delineated by their skins. This suggests the following criterion. (Throughout the paper, "*S*" refers to a thinker and "*F*" refers to a property *S* instantiates.)

> **(Spatiophysical Criterion)** *F* is intrinsic to *S* <u>iff</u> *F* is a physical property instantiated within the spatial boundary constituted by *S*'s skin.

While the Spatiophysical Criterion fits classic ways of stating the externalist thesis, it is plainly inadequate. This criterion interprets externalism as the claim that thought contents do not supervene on physical factors within the spatial boundary of the organism; it thereby links externalism to a seemingly unrelated question about mental ontology. Perhaps the clearest indication of this flaw is that this criterion classifies Descartes—standardly regarded as the archetypal internalist—as an *externalist*. For Descartes denies that mental contents metaphysically supervene on any physical properties. (Burge (2003a) notes that this flaw was present in his earlier (1986b) characterization of internalism.)[5]

Sensitivity to this issue about mental ontology has led some philosophers to take special care in formulating externalism (and internalism). Here is a good example of a carefully formulated externalist claim.

> [I]t is possible for thinkers that are alike in all intrinsic physical respects to differ in the contents of their thoughts by virtue of differences in their environments. (McLaughlin and Tye 1998, 349)

By specifying that it is environmental differences that are responsible for the difference in thought contents, this formulation adds a condition for externalism not present in the previous formulation. It is not clear whether McLaughlin and Tye intend this as a necessary condition for externalism, a sufficient condition, or both. But it will serve our purpose of expressing the externalist thesis only if it is both necessary and sufficient; so we must consider whether it satisfies that role.

Assume, for the moment, that the environmental differences in question are *physical* differences. (We revisit this assumption below.) This formulation then suggests the following criterion.

> **(Modified Spatiophysical Criterion)** *F* is intrinsic to *S* <u>iff</u> either (i) *F* is a physical property instantiated within the spatial boundary constituted by *S*'s skin, or (ii) *S*'s instantiating *F* does not metaphysically depend on any physical features of the environment.

Using this criterion, externalism is the claim that a difference in the physical environment can suffice for a difference in thought contents between two persons who are intra-skin physical duplicates.

The Modified Spatiophysical Criterion improves on the original Spatiophysical Criterion in that it classifies Descartes as an internalist. A Cartesian soul's thinking a particular thought (instantiating a particular content property) is independent of the physical features of the environment, and hence is intrinsic to the thinker, according to this criterion. And this criterion fits nicely with some of the principal examples used to support externalism. In these examples, the thought contents of physical duplicates differ purely in virtue of physical differences between their environments: e.g., differences in the microstructure of the local watery stuff (H_2O vs. XYZ).

But the Modified Spatiophysical Criterion contains the same flaw as the original Spatiophysical Criterion, though in a less obvious form. By construing externalism as the claim that thought contents depend on specifically *physical* features of the environment, the Modified Spatiophysical Criterion links externalism to a seemingly unrelated ontological issue. To see this, consider a view constituted by two claims.

(1) Possessing the concept <u>arthritis</u> is an irreducibly mental (i.e., nonphysical) property.

(2) A thinker's ability to entertain <u>arthritis</u> thoughts metaphysically depends on the possession of the concept <u>arthritis</u> by experts in her community (and on no other environmental factor).

On this view, the fact that community experts possess the concept arthritis is a nonphysical feature of the environment. *Thinking that arthritis is painful* thus satisfies condition (ii) of the Modified Spatiophysical Criterion, and is therefore *intrinsic* to the thinker, according to that criterion. So the Modified Spatiophysical Criterion will count the conjunction of (1) and (2) as an internalist view. However, this view seems patently externalist. In fact, Burge may hold something like this view.[6] (I will refer to the conjunction of (1) and (2) as Externalist Dualism, though of course it is only one brand of externalist dualism.)

So the Modified Spatiophysical Criterion is inadequate. It correctly classifies some externalist positions, viz., those that claim that two physical duplicates' thoughts can differ purely in virtue of differences in their physical environments. But it misclassifies another plainly externalist view, because it counts, as intrinsic to the thinker, an apparently non-intrinsic property (being in a community in which experts possess the concept arthritis).[7]

Both of the criteria we have considered cash out "intrinsic to the thinker" in partly physical terms. This leads to problems with each: the initial Spatiophysical Criterion misclassified Cartesianism, and the Modified Spatiophysical Criterion misclassified Externalist Dualism. The lesson is clear. Definitionally linking intrinsic (or non-intrinsic) properties with the physical entangles externalism and internalism with ontological issues that are orthogonal to them.

On reflection, this result is unsurprising. For internalism and externalism are, in spirit, ontologically neutral. This neutrality is reflected in the fact that each of the following positions has been defended by influential philosophers: internalist dualism (Descartes, David Chalmers); internalist materialism (Jerry Fodor[8], Frank Jackson[9], Gabriel Segal); externalist dualism (Tyler Burge and perhaps Donald Davidson[10]); externalist materialism (Fred Dretske, Hilary Putnam, Michael Tye, and numerous others).

An obvious strategy for avoiding these ontological complications is to abandon the assumption, present in condition (ii) of the Modified Spatiophysical Criterion, that environmental features are physical features. This tactic is suggested by the formulation of internalism (or "individualism") on which Burge seems to have settled.

> According to individualism about the mind, the mental natures of all a person's or animal's mental states (and events) are such that there is no necessary or deep individuative relation between the individual's being in states of those kinds and the nature of the individual's physical *or social* environments. (Burge 1986b, 3-4, my emphasis; compare Burge 2006, 152.)

Externalism is then the claim that this "necessary or deep individuative relation" sometimes does obtain.

This formulation correctly classifies Descartes, since Descartes would deny that thoughts are individuated by relation to the physical or social environment. And it also seems to yield the desired classification of Externalist Dualism, since community experts' possession of the concept arthritis is a feature of the social environment.

Crucially, this latter consequence depends on the assumption that the social environment qualifies as *external* to the thinker even if it is not a matter of *physical* features of the world beyond her skin. This assumption invites the question: in what sense is the social environment *external* to the thinker? One answer, which retains the desired ontological neutrality,[11] is that the

social environment is external to the thinker in a *spatial* sense. The corresponding demarcation of the thinker's intrinsic properties is as follows.

> **(Spatial Criterion)** *F* is intrinsic to *S* iff either (i) *F* is instantiated within the spatial boundary defined by *S*'s skin, or (ii) *S*'s instantiating *F* does not metaphysically depend on any features of the environment outside the spatial boundary defined by *S*'s skin.

The Spatial Criterion avoids the ontological entanglements on which the previous criteria foundered. And it generates the appropriate classifications of Cartesianism (as internalist) and Externalist Dualism (as externalist).

However, the Spatial Criterion is disloyal to the spirit of the externalism-internalism debate. This point is aptly demonstrated with an ingenious case devised by Katalin Farkas (2003). Farkas imagines twins who are precisely similar except for one particular. One twin, on Earth, suffers from meningitis. The other, on Twin Earth, suffers from a disease that is superficially similar to meningitis, and is called "meningitis" on Twin Earth, but involves a bacterium different from the meningitis bacterium (meningococcus). Farkas designs this case to closely parallel Putnam's argument for externalism regarding *water*. A further similarity is that Farkas' case takes place in 1750, before the bacterium associated with meningitis was identified.

Putnam's example challenges internalism by prompting the intuition that two physical duplicates who differ only in the makeup of the watery stuff in their environment (H_2O vs. XYZ) entertain different contents when they think (what they would express by saying) "water quenches thirst". Given that Farkas' meningitis case parallels Putnam's example, one would expect internalism to be challenged by the intuition that Farkas' twins entertain different contents when they think (what they would express by saying) "meningitis is dangerous". But the Spatial Criterion does not deliver that result. According to the Spatial Criterion, the presence of the bacterium is *intrinsic* to each twin, since the bacterium is present within the spatial boundary defined by the skin. (Meningitis is a brain disease, so its presence falls within more restrictive spatial boundaries as well.) The claim that the difference between those bacteria can suffice for a difference in thought contents thus presents no challenge to internalism—it is perfectly compatible with internalism. The upshot is that the Spatial Criterion does not capture the spirit of the externalism-internalism dispute. Using that criterion, an argument relevantly similar to a classic argument against internalism does not threaten internalism.

The meningitis case fails to challenge internalism because meningitis occurs within the spatial boundary of the thinker. Its presence thereby satisfies the first clause of the Spatial Criterion. We might try to resolve this problem by eliminating that clause, and understanding "intrinsic to the thinker" solely by reference to the second, environmental clause.

> **(Modified Spatial Criterion)** *F* is intrinsic to *S* iff *S*'s having *F* does not metaphysically depend on any features of the environment outside the spatial boundary defined by *S*'s skin.

This criterion may be more loyal to Burge's intentions, since his formulation of externalism quoted above focuses exclusively on the contribution of the environment and says nothing about what occurs within the subject's skin.

The Modified Spatial Criterion is initially promising. But it is threatened by a variant of the meningitis example. (This variant is my own twist on Farkas' thought experiment.) Compatibly with the meningitis example as previously described, the twins' environments may be perfectly similar: this would be the case if each of the respective bacteria first appeared in the twins, and neither was yet present in their environments (outside their skins). Suppose this is the case. While this additional supposition weakens the parallel with Putnam's original case somewhat, it does not affect the basis for the intuition that drives the challenge to internalism. In the original case, the intuition was this: subjects can think *water* thoughts without being in a position to distinguish water (H_2O) from stuff that is only superficially similar (XYZ). In the meningitis case, the intuition is this: subjects can think *meningitis* thoughts without being in a position to distinguish meningitis from a disease that is only superficially similar (twin meningitis). In both cases, the difference in thought contents derives exclusively from the difference in natural kinds. The fact that the relevant natural kind is instantiated within the spatial boundary of the skin, rather than outside that boundary, has no bearing on the thrust of the thought experiment.

The insignificance of spatial location nicely explains why early discussions ignored the fact that the "twins" in the water example are not genuinely "molecule-for-molecule duplicates". These discussions treated the presence of H_2O as an external factor, despite the fact that water is present within the *spatial* boundary of the individual organism.

The Modified Spatial Criterion construes externalism and internalism as views about where content-individuating factors can be spatially located. Using that criterion, the claim that the twins in the meningitis case would differ presents no challenge to internalism. As Farkas convincingly argues, the meningitis case parallels the H_2O case in all crucial respects: if internalism is challenged by the intuition that the twins' thought contents differ in the latter case, it should be equally challenged by the corresponding intuition in the former case. If there is a single, clear notion of "intrinsic to the thinker" at work in this classic externalist argument, it is not a spatial notion.

To respect the ontological neutrality of externalism and internalism, an adequate formulation of these positions cannot employ a criterion that defines "intrinsic to the thinker" in physical terms. Retreating to a less committal, purely spatial criterion has some advantages. But this strategy ultimately fails, as a spatial criterion plainly conflicts with the spirit of the externalism-internalism debate.

3. The Epistemic Approach

Farkas' meningitis scenario reveals that what divides externalism from internalism is not a claim about the spatial location of content-individuating factors. She suggests that what leads us to take the meningitis case to be similar to Putnam's water example, as regards the potential challenge to internalism, is an epistemic feature: the subjects in both cases are blind to the differences between their thoughts and their twins'. On her view, the point at issue between externalism and internalism concerns the epistemic status of thought contents—specifically, whether differences in thought content are subjectively distinguishable.

In a nutshell, Farkas' argument is as follows. The question of what is intrinsic to the thinker is primarily intended to concern the mind; in these discussions, the brain is at best a stand-in for the mind. And "[w]hat it is to have a mind is inseparable from what it is for example to have experiences, and this latter is a thoroughly epistemic notion." (Farkas 2003, 205)

Moreover, most philosophers believe that externalism faces, and internalism avoids, at least a *prima facie* problem of compatibility with the phenomenon of privileged access. Farkas concludes that externalism and internalism are, at bottom, views about thinkers' epistemic relations to their thoughts. Specifically, internalism is the thesis that

> facts individuate mental contents only insofar as they *make a difference* to the way things appear to us. This means that any difference in the content of thoughts should be distinguishable from the subject's point of view and hence remains within the reach of privileged access. (ibid., 203)

The following criterion captures Farkas' proposal.

> **(Epistemic Criterion)** *F* is intrinsic to *S* iff *S*'s instantiation of *F* makes a difference to how things appear to *S*, in a way that enables *S* to have privileged access to the fact that she instantiates *F*.

This proposal has significant benefits. As Farkas observes, it makes sense of the widespread impression that externalism faces a special burden in explaining privileged access. It correctly classifies Descartes, as an internalist. It also correctly classifies Externalist Dualism, as externalist (assuming that whether an expert in my community has the concept *arthritis* makes no difference to "how things appear" to me). Finally, this proposal captures the spirit of Putnam's argument and, relatedly, yields the appropriate construal of the meningitis case. The externalist reading of these cases is that one can think a determinately *water* (or *meningitis*) thought without being in a position to distinguish this thought from a *twin water* (or *twin meningitis*) thought.[12]

While Farkas acknowledges that the epistemic approach is unorthodox, she contends that it reflects the "motives [that] lie behind the externalist thesis" more accurately than spatial criteria (ibid., 193). It's not entirely clear to me whether Farkas' proposal is intended purely as an explication of the current debate. But our purpose is explicatory: we must examine whether the Epistemic Criterion reflects the current debate.[13]

The Epistemic Criterion has some problematic consequences. First, it ensures that externalism is incompatible with privileged first-person access, as a definitional matter. The Epistemic Criterion glosses externalism as the claim that content properties don't supervene on (and hence, aren't identical to) properties to which the thinker enjoys privileged access. Farkas embraces this consequence, saying that "one way to sum up my proposal is to say that externalism is a thesis about the nature of our access to our thoughts" (ibid., 204). While most externalists concede that their view initially appears incompatible with privileged access, most also maintain that these are ultimately compatible. Regardless of whether compatibilism is true, the controversy surrounding this issue casts doubt on the idea that incompatibility with privileged access is a simple analytic consequence of externalism.

A second worry about the Epistemic Criterion is that, by defining properties "intrinsic to the thinker" as those which (in Farkas' words) "*make a difference* to the way things appear", it renders externalism about the phenomenal incoherent. For surely phenomenal differences "*make a difference* to the way things appear". This result is especially troublesome because most advocates of phenomenal externalism take phenomenal character to be a species of intentional content (Dretske 1996, Lycan 2001, Tye 2000). So the sense of "externalism" operative in phenomenal externalism is precisely the sense operative in content externalism.

Finally, the Epistemic Criterion has difficulty making sense of the pivotal externalist claim that some intensional thought contents are wide. In Burge's terms, we must sometimes individuate thoughts widely in order to capture the thinker's "epistemic perspective": "how things seem to him, or in an informal sense, how they are represented to him" (Burge 1979, 25). This claim arguably constitutes the core externalist challenge to internalism. Internalists can grant that extensional content (e.g., what a *water* thought refers to, in a given context) is wide. So the key externalist claim is that some *intensional* contents—contents that reflect "how things seem to [the thinker], or … how they are represented to him"—fail to supervene on his intrinsic properties.

A dilemma emerges when we try to make sense of this externalist claim, using the Epistemic Criterion. This dilemma centers on the question whether a difference in a thinker's intensional contents (that is, in her epistemic perspective) must be subjectively distinguishable by her. Suppose the answer is "yes". On this supposition, the Epistemic Criterion classifies any factor on which the epistemic perspective depends as intrinsic to the thinker: hence, any way of individuating thoughts that captures the epistemic perspective will be a version of *internalism*. So on this first horn of the dilemma, a key externalist claim—that some intensional contents are wide—is incoherent.

The other horn of the dilemma is generated by denying that differences in intensional contents must be subjectively distinguishable. This horn allows for a coherent reading of the externalist claim just mentioned. But it implies that it is not (merely) a difference in intensional content that "enables *S* to have privileged access to the fact that she instantiates *F*". Some factor other than intensional content must explain privileged access. The only plausible alternative seems to be a thought's phenomenal character: *what it's like* to think that thought. Intrinsic properties—properties that make a difference to how things appear, in a way that allows for privileged access to the corresponding thoughts—would then be phenomenal properties. (Williamson (2000, 49) suggests identifying the internal with the phenomenal, as a way of sidestepping issues about physicalism.) Now if a difference in phenomenal character is what renders two thoughts subjectively distinguishable, then, given the Epistemic Criterion, the question dividing internalists and externalists is whether thought contents supervene on phenomenal character. But that question belongs to a different debate, one that is orthogonal to the debate over externalism. (Burge explicitly denies that the target of his arguments against internalism is the claim that content supervenes on phenomenal character.)

So the second horn of the dilemma is this: if differences in intensional content need not be subjectively distinguishable, then the only feature that could ground subjective distinguishability seems to be phenomenal character. On this horn, the Epistemic Criterion construes externalism as the view that thought contents fail to supervene on phenomenal character.[14] (Farkas accepts this implication in her 2008 book.)

The Epistemic Criterion is superior, in significant respects, to the previous criteria. It avoids entanglements with extraneous ontological issues, and makes sense of some classic externalist arguments (such as Putnam's "water" argument). Moreover, an epistemic approach to delineating the thinker seems more salient to philosophical concerns than physical or spatial approaches. But the Epistemic Criterion seriously distorts the current debate. It makes the denial of privileged access a simple analytic consequence of externalism. It renders phenomenal externalism incoherent. And it either renders a key externalist claim incoherent or mistakenly

construes this debate as centering on the question whether intentional content supervenes on phenomenal character. The Epistemic Criterion does not satisfy our search for a univocal criterion implicit in the externalism-internalism debate.

4. The Neutral Approach

None of the criteria for "intrinsic to the thinker" we have considered provides an accurate construal of the mental content debate. These criteria cash out externalism and/or internalism as involving commitments that seem wholly unrelated to them—and which, in some cases, their leading proponents explicitly disavow. This pattern suggests that, to do justice to the current debate, an interpretation of "intrinsic to the thinker" must be relatively neutral, at least about ontological and epistemic matters.

In a valuable discussion, Richard Fumerton describes obstacles to establishing a precise definition of externalism and internalism. He responds to these obstacles by retreating to a highly neutral—even austere—understanding of what is intrinsic (or "internal") to the thinker.

> I suspect that in the end we will simply need to understand internal states as including both nonrelational properties of the self and the self's standing in certain sorts of nonnatural relations (such as acquaintance) with certain entities. Though inelegant, that's the only way I can see how to define internalism so that paradigm internalists stay in the right camp. (Fumerton 2003, 262)

The "certain entities" Fumerton mentions are universals. In effect, his proposal is similar to Ludwig's proposal (quoted in Section 1 above), with a verbal difference about whether standing in relation to a (presumably abstract) universal is a "relational property".

Reserving "relational property" for relations to concreta, the following roughly captures the Ludwig/Fumerton approach.

> **(Thinker Criterion)** *F* is intrinsic to *S* <u>iff</u> *S*'s instantiating *F* does not entail the existence of any concrete entity wholly distinct from *S*.

The Thinker Criterion achieves the ontological neutrality required to correctly classify both Descartes' view and Externalist Dualism. Descartes qualifies as an internalist, since he would presumably deny that one's having a particular thought depends on (or entails) the existence of any other concrete thing. Externalist Dualism qualifies as externalist so long as community experts are concrete entities distinct from the thinker.

Another strength of the Thinker Criterion is that it captures at least part of the spirit of the externalism-internalism debate. For it characterizes externalism as the claim that, for some thought contents, having a thought with these contents requires that the thinker is appropriately related to certain contingently existing things distinct from her. And the classic externalist arguments center on the thinker's relation to contingently existing things distinct from her (H_2O, experts who use "arthritis" in a certain way, etc.).

One consequence of the Thinker Criterion may at first be surprising. This criterion classifies the "extended mind" view (Clark and Chalmers 1998)—also known as "vehicle externalism"—as neutral between externalism and internalism. According to this view, factors "external" to a thinker, such as a notebook, sometimes perform genuinely cognitive functions

for the thinker, and hence partly constitute his beliefs and other attitudes.[15] Such factors thereby qualify as *part of* his mind and, hence, part of the thinker himself. The mind and the thinker are *extended* to include factors like notebooks.

> [The subject] himself is best regarded as an extended system, a coupling of biological organism and external resources. (Clark and Chalmers 1998, 18)

The claim that a notebook could partly constitute the thinker illustrates a point anticipated in Section 1: that even seemingly paradigmatic external factors may be glossed as intrinsic to the thinker.

Now if I am an extended system that includes my notebook, then my notebook is not wholly distinct from me. So the fact that my believing that *p* depends on my notebook does not entail externalism, according to the Thinker Criterion. Whether externalism is true depends on a question on which vehicle externalism is neutral, namely, whether my content properties entail the existence of any contingent entity that (unlike my notebook) is not within my extended mind or self. By contrast, the spatial criteria outlined in Section 2 classify vehicle externalism as externalist, since vehicle externalism denies that content properties supervene on properties instantiated within the skin.[16] (How the Epistemic Criterion classifies vehicle externalism is a complicated question.[17])

That the Thinker Criterion construes vehicle externalism as compatible with (content) internalism is not a strike against it. After all, vehicle externalism differs markedly from the paradigmatic content externalist positions of Burge, Davidson, and Putnam. These positions do not imply the vehicle externalist thesis that external factors can *partly constitute* mental states. Stephen Yablo (1997) highlights this contrast when he notes that Putnam's famous slogan "meanings ain't in the head" mischaracterizes Putnam's own conclusion. That slogan implies that external factors *partly constitute* meanings (and, by extension, beliefs). But classic externalist views say only that external factors sometimes *individuate* contents, making it the case that a belief is the belief that *p* rather than the belief that *q*. Moreover, Chalmers (2002) embraces both vehicle externalism and content internalism. Far from a strike against it, then, the result that vehicle externalism is neutral on the question of content externalism is plausibly a strength of the Thinker Criterion. (The label "vehicle externalism" reflects the influence of spatial construals of "intrinsic to the thinker".)

The Thinker Criterion does, however, face a serious problem. It fails to provide *informative* truth conditions for externalism or internalism. Consider the kind of truth conditions provided by the Spatiophysical Criterion. According to that criterion, externalism is true (and internalism is false) *iff* two thinkers who are precisely similar, as regards physical properties instantiated within the skin, may differ as to whether they think that *p*. This criterion has the potential to shed light on the debate about mental content, for it generates truth conditions for externalism and internalism that are informative, albeit ultimately flawed. By contrast, the Thinker Criterion says that externalism is true (and internalism is false) *iff* two thinkers can differ, as to whether they think that *p*, purely by virtue of differences in concreta existing outside them. But this is uninformative. To say that an entity exists *outside*—is wholly distinct from—the thinker is just to say that *being such that that entity exists* is not among the thinker's intrinsic properties.

In effect, the Thinker Criterion reintroduces our original question: how should we understand "intrinsic to the thinker" in (I) and (E)?

> **(I)** Thought contents always supervene on properties intrinsic to the thinker.
>
> **(E)** Thought contents do not always supervene on properties intrinsic to the thinker.

The Thinker Criterion does not illuminate these statements. The truth conditions for externalism and internalism generated by the Thinker Criterion are precisely those already inherent in the statements we are trying to explicate. Externalism is true (and internalism is false) *iff* two thinkers who are precisely similar, as regards intrinsic properties, may differ as to whether they think that *p*.

The Thinker Criterion's neutrality enables it to avoid saddling externalism or internalism with extraneous commitments. But this criterion is too neutral to be informative.

Clearly, what is needed is a criterion of "intrinsic to the thinker" that is informative (and thereby improves on the Thinker Criterion) yet also neutral in relevant respects (and thereby avoids entanglements with orthogonal issues). The contrast between vehicle and content externalism suggests a new tack. Construe externalism as the claim that some content-determining factors are external to content vehicles—e.g., to the thoughts possessing that content. In other words, thought contents don't always supervene on properties intrinsic to thoughts themselves. This yields the following construal of the externalism-internalism debate.

> **(Vehicle Construal)** The defining thesis of internalism is that thought content always supervenes on properties intrinsic to the thought. The defining thesis of externalism is the denial of this claim.

This construal nicely matches the kind of relationship between thoughts and contents envisioned by (at least some) traditional externalists. Davidson (1987) illustratest his relationship with a sunburn analogy. A sunburn is located on the skin, but what makes it a sunburn is an external factor: that it was caused by sun exposure. Since a cause other than sun exposure could lead to precisely similar damage, two intrinsically similar bits of skin (on intrinsically similar organisms) could differ in that only one is sunburned. So the property *being sunburned* does not supervene on properties intrinsic to the skin (or organism). Analogously, according to content externalists some factors that contribute to fixing a thought's content may be external to the thought itself: such factors include the use of "arthritis" by experts in the community and (in the meningitis case) the presence of a certain bacterium in the brain. So the thought I'd express by saying "meningitis is dangerous" may have the same intrinsic properties as the thought my twin would express with those words, even if my thought is a *meningitis* thought whereas hers is a *twin meningitis* thought.

Unlike the proposals we've considered thus far, this construal of the debate is not based in a criterion for "intrinsic to the thinker". Nor does it provide such a criterion, since "intrinsic to the thinker" is not equivalent to "intrinsic to the thought" or even to "among the intrinsic properties of the thinker's thoughts". Being among the intrinsic properties of *S*'s thoughts is plausibly *sufficient* for being intrinsic to *S*. But it is much less clear that this condition is *necessary* for being intrinsic to *S*. To restrict intrinsic properties of thinkers to intrinsic properties of their thoughts is to endorse the bundle theory of the self, or something very close to it. Because the bundle theory is highly controversial, it's unlikely that that theory (or anything close to it) is a foundational assumption of the debate about mental content.

This means that the question the Vehicle Construal takes to define this debate—whether thought contents supervene on properties intrinsic to thoughts—is not a plausible interpretation of the question ordinarily taken to define this debate, namely, whether thought contents supervene on properties intrinsic to the thinker. So an immediate worry about the Vehicle Construal is that it seems to conflict with the ordinary understanding of the point at issue between externalism and internalism. Whereas previous proposals were explications of this ordinary understanding, the Vehicle Construal is a competitor to it.

Let's put this worry aside for the moment, and consider how the Vehicle Construal fares in other respects. This construal appears to correctly classify Externalist Dualism. It may also correctly classify Descartes' view, though this is somewhat less clear.[18] It avoids the problem posed by the meningitis case, since even if meningitis occurs within the thinker, in some sense, *occurring in a brain in which meningitis is present* is plausibly a non-intrinsic property of a meningitis thought. And this construal shares, with the Thinker Criterion, the virtue of classifying vehicle externalism as neutral between internalism and externalism. The defining claim of vehicle externalism is that some content vehicles are partly constituted by factors outside the organism's biological boundary: vehicle externalism is silent on the question whether properties intrinsic to content vehicles exhaustively determine content properties.

But the Vehicle Construal faces a problem, stemming from its reliance on the distinction between the factors determining thought contents and thoughts themselves. This distinction is an instance of the more general distinction between *total* realizations and *core* realizations. A property's total realization is the set of conditions that jointly suffice for its being instantiated.[19] Its core realization is that part of the total realization corresponding to the thing that has the property. For example, the total realization of *being sunburned* is something like *having damage caused by sun exposure*. The core realization is just the skin, as it is the skin that has the property *being sunburned*. A thought's total realization is the set of conditions that jointly suffice for the instantiation of its content properties. E.g., if externalism is true the total realization of a particular thought that *water quenches thirst* may include the presence of H_2O in the environment. This thought's core realization is just the thought itself, which has this content.

According to the Vehicle Construal, the externalist thesis is that content properties sometimes fail to supervene on the properties intrinsic to thoughts. To cash out this thesis, we need some way of distinguishing properties intrinsic to a thought's core realization, on the one hand, from those that only belong to its total realization. In other words, we need some criterion for "intrinsic to a thought". The effect of replacing "intrinsic to the thinker" with "intrinsic to the thought", in our formulation of the point at issue in this debate, is to replace the need for a criterion for the former with a need for a criterion for the latter. Instead of asking how *thinkers* are delineated, in this context, we now need to ask how *thoughts*—core realizations of content properties—are delineated.

In some cases, like the case of sunburn, the distinction between core and total realizations is easily drawn. Properties intrinsic to the core realization of *sunburn* are distinguished from other parts of its total realization along biological and temporal lines. The properties intrinsic to the core realization (the damaged skin) are limited to those within a biologically salient region—in this case, the skin itself. And they concern the present time, whereas *having been caused by sun exposure* concerns the past. By contrast, properties intrinsic to a thought's core realization cannot be distinguished from other parts of its total realization in biological or temporal terms.

Delineating a thought's core realization in biological terms would entangle the debate about mental content with questions of physicalism. A temporal delineation would construe plainly externalist claims, to the effect that thought contents are partly fixed by the natural kinds present in the environment *at the time of the thought*, as perfectly compatible with internalism.

It should be clear why spatial or epistemic approaches to understanding "intrinsic to the thought" will also be inadequate. These approaches will fail for precisely the reasons they failed regarding "intrinsic to the thinker": they will conflict with the ordinary taxonomy of views, or commit externalists or internalists to positions on which they are neutral (or, in some cases, which they explicitly reject). For example, identifying properties intrinsic to a thought with properties to which a thinker is epistemically sensitive, in a way that explains privileged access, would make the denial of privileged access a simple analytic consequence of externalism.

We should look for a new approach, one that diverges from the approaches to understanding "intrinsic to the thinker" we've previously considered. One obvious strategy for delineating something's core realization is to construe properties intrinsic to a core realization as those that underwrite the causal features of the thing. The total realization of a penny includes *being produced at a U.S. Mint*. But this part of the total realization seems irrelevant to the penny's causal features. A perfect duplicate of a penny that differed only in *not* being produced at a U.S. Mint would possess the same causal features: when run over by a train, both would flatten in precisely the same way; proffering a handful of such duplicates, as payment in a store, is as likely to exasperate a cashier as proffering a handful of pennies. So we might say that properties intrinsic to a thought are those directly responsible for the thought's causal features; causally irrelevant properties may belong to its total realization, but are not part of the thought itself.

But this strategy will not work. One problem is that the issue of causal relevance is not as straightforward as my example suggests. Some arguably causal explanations invoke properties not usually regarded as belonging to a core realization. That I gave the clerk a genuine penny seems to causally explain why I now have less money (legal tender) than I did a moment ago, whereas my handing over a counterfeit penny would not.

A more serious difficulty with this strategy is that it ensures that wide content is causally irrelevant. On the Vehicle Construal, narrow content is content that supervenes on the intrinsic properties of a thought's core realization. Wide content is content that metaphysically depends on factors beyond those intrinsic properties. (Because of this dependence, these latter factors belong to the thought's total realization.) So if properties intrinsic to core realizations are exclusively responsible for a thought's causal features, then we need not advert to wide content to explain a thought's effects on cognition or behavior. But the idea that wide content is irrelevant to such explanations is a standard *objection* to externalism, and is rejected by most externalists. So no plausible construal of externalism will interpret that view as straightforwardly entailing the causal inefficacy of wide content.

We might cast about for other ways to delineate thoughts, distinguishing properties intrinsic to a thought's core realization from properties merely belonging to its total realization. But this exercise is not likely to illuminate externalism and internalism. Any substantive principle used to distinguish core realizations is in danger of being insufficiently neutral. This was the flaw in the proposal just considered: that proposal used a substantive claim about what sorts of

factors are relevant to causal explanations, and thereby committed externalists to a position that most of them reject.[20]

There is a more general reason to doubt that any way of delineating a thought will (when combined with the Vehicle Construal) yield an adequate explication of externalism and internalism. This is the worry expressed earlier: the Vehicle Construal does not provide for a suitable criterion of "intrinsic to the thinker", and therefore conflicts with the ordinary understanding of the externalism-internalism debate. Given the pervasiveness of the ordinary understanding, abandoning it in favor of the Vehicle Construal seems unwarranted.

There does seem to be something right about the Vehicle Construal. Internalists *may* generally accept, and externalists *may* generally deny, that a thought's content always supervenes on properties intrinsic to the thought. But I submit that, to the extent that the Vehicle Construal identifies a question that divides these two camps, this is because the distinction it relies on—between the factors determining content (a thought's total realization) and the thought itself (its core realization)—derives from a prior, more fundamental distinction between those properties of a thinker that are intrinsic to her and those that are not. If internalists disagree with externalists about whether a thought's content always supervenes on properties intrinsic to the thought, this is because "intrinsic to the thought" is understood by reference to what is *intrinsic to thinkers*.

Since this debate is not premised on the assumption that thoughts are the only features intrinsic to thinkers, the vehicle/content (or core realization/total realization) distinction will not explicate the notion of "intrinsic to the thinker". The Vehicle Construal accurately reflects an aspect of the current debate (if it does) only insofar as it relies on the assumption that properties intrinsic to a thought are intrinsic to thinkers—where the notion of "intrinsic to the thinker" remains unarticulated. So it cannot explicate that notion, or the externalism-internalism debate.

Let us review. The Thinker Criterion avoids the problematic commitments of previous criteria by defining externalism relative to a neutral conception of the thinker. While this criterion may be accurate, its neutrality prevents it from *explicating* the internalism-externalism debate. The Vehicle Construal aims to improve on the Thinker Criterion by providing informative truth conditions for externalism and internalism, while preserving the Thinker Criterion's ontological and epistemic neutrality. To achieve this latter goal, it exploits the neutral distinction between thoughts, as vehicles of content, and the factors that suffice for determining thought content. But absent a criterion of "intrinsic to the thought", the Vehicle Construal is no more informative than the Thinker Criterion. And any such substantive criterion—e.g., identifying properties intrinsic to a thought as those that ground its causal features—will threaten the Vehicle Construal's accuracy. This construal expresses a point on which externalists and internalists disagree (if it does) only by implicitly restricting properties "intrinsic to the thought" to properties intrinsic to the thinker. So it sheds no light on the externalism-internalism debate, or on the sense of "intrinsic to the thinker" operative therein.

5. Prospects for defining internalism and externalism

We have examined three approaches to defining internalism and externalism. The first approach accepts familiar construals of "intrinsic to the thinker" at face value. It glosses

externalism as the claim that thoughts don't metaphysically supervene on (perhaps physical) properties instantiated within a certain spatial region, or that they metaphysically depend on (perhaps physical) properties instantiated outside that region. The second approach interprets externalism as the claim that distinct thoughts can be subjectively indistinguishable, perhaps because they are sometimes phenomenally similar. The third approach construes externalism as the claim that some content-determining factors are relational features of thinkers or of thoughts. None of these approaches succeeds in explicating the current debate. The first two approaches are overly committal about the nature or limits of the thinker. This lack of neutrality leads both of these approaches to conflict with the usual classification of familiar views, or to interpret internalism or externalism as committed to claims—about mental ontology, the subject's access to her own thoughts, or the relation between the intentional and the phenomenal—generally regarded as orthogonal to those positions. The third approach generally avoids these pitfalls. But its more promising versions implicitly rely on the distinction between intrinsic and non-intrinsic properties of the thinker. So this approach does not illuminate that distinction.

Should we persist in the search for a suitable criterion of "intrinsic to the thinker", one that is loyal to how philosophers ordinarily construe externalism and internalism? Participants in this debate do seem to have *some* common understanding of what is at issue; and there is relatively wide consensus about which sorts of positions are externalist and which are internalist. So perhaps there is some shared, implicit notion of intrinsic properties remaining to be discovered. In other words, perhaps our situation is similar to that which J.S. Mill described as the situation in ethics. Mill claimed that there was widespread agreement about which particular actions are right, and which are not right, but little consensus about the criterion of rightness. On his diagnosis, this curious situation was due to "the tacit influence of a standard [of rightness] not recognised"—namely, the Principle of Utility (Mill 1863, 3).

But there are strong reasons to doubt that there is a single, unrecognized criterion of "intrinsic to the thinker" operative in the current debate about content. First, the usual explicit gloss of this concept, in spatial (or spatiophysical) terms, has become deeply ingrained. Traces of this approach are ubiquitous in discussions of internalism and externalism: they are present in Putnam's famous slogan that meanings "ain't in the head", in the standard description of twins as "molecule-for-molecule duplicates", and in characterizations of intrinsic properties as those instantiated "within the skin". These familiar phrases have shaped our intuitions about what kinds of properties are intrinsic to thinkers, and about which views count as externalist and which as internalist. As noted above, the spatial approach is likely responsible for the fact that the extended mind view, which appears neutral about content externalism, nonetheless carries the label "vehicle externalism". As we saw above, the spatial approach is clearly inadequate: many of the intuitions rooted in this approach clash with the spirit of externalism and internalism, as ordinarily understood. Still, its influence on our intuitions dims the prospects for alternative approaches, as such alternatives will inevitably conflict with those intuitions.

The initial promise of each of the spatial criteria, and of the various ways of unpacking the Epistemic Criterion, supplies a second reason to doubt that there is a uniform tacit standard of "intrinsic to the thinker" at work here. Each of these proposals fits some dimension of the internalism-externalism debate, as ordinarily understood. For each of the following issues is implicated in *some* aspect of this debate: the relation between an organism's thoughts and the

natural kinds in its physical environment; the division of linguistic and conceptual labor within a social community; the linguistic communicability of thoughts; privileged access to one's own mental states; and the relation between the phenomenal and the intentional. Because each of these issues is closely associated with some aspect of the externalism-internalism debate, and *no single criterion will capture all of them*, we have reason to doubt that there is a criterion of the internal that will do justice to the usual terms of this debate.

I propose, then, that we abandon the search for a criterion of "intrinsic to the thinker" that will capture the terms of the externalism-internalism debate, and discontinue the debate as it is now framed. To make progress on the diverse range of issues linked with this debate, we might focus our attention on more well-defined questions, of the sort that emerged from this discussion. We might ask whether thought contents supervene on physical properties that fall within the spatial boundary constituted by the skin; whether a difference in concepts possessed by experts (distinct from the thinker) can suffice for a difference in thought contents; or whether distinct thought contents are subjectively distinguishable by the thinker; etc.

Alternatively, we might try to rehabilitate the question at the heart of the current debate, namely, "Do thought contents always supervene on properties intrinsic to the thinker?" This rehabilitative process involves two stages. The first, negative stage consists in surrendering our implicit associations with this question, including our present opinions about the implications of particular answers to it. The second, positive stage begins with an exercise in metaphysics: establishing a precise, principled conception of the boundary of the thinker, which can be used to unpack "intrinsic to the thinker". Only once such a conception is in hand can we address the question of supervenience.

One moral of this discussion is that any way of delineating the thinker will significantly reframe the debate over whether thought contents supervene on properties intrinsic to the thinker.[21] But if my arguments here succeed, they show that a fresh approach is overdue.[22]

References

Burge, T. (1979) "Individualism and the Mental". *Midwest Studies in Philosophy 4*, P. French, ed. (University of Minnesota Press), pp. 73-122. Reprinted in P. Ludlow and N. Martin (1998).

Burge, T. (1986a) "Cartesian Error and the Objectivity of Perception". In *Subject, Thought, and Context*, edited by P. Pettit and J. McDowell. Oxford: Oxford University Press, 117-36.

Burge, T. (1986b) "Intellectual Norms and Foundations of Mind", *Journal of Philosophy* 83: 697-720.

Burge, T. (2003a) "Descartes, Bare Concepts, and Anti-Individualism: Reply to Normore". In Hahn, M. and B. Ramberg (2003) *Reflections and Replies: Essays on the Philosophy of Tyler Burge* (MIT Press), pp. 291-334.

Burge, T. (2003b) "Epiphenomenalism: Reply to Dretske". Hahn, M. and B. Ramberg (2003) *Reflections and Replies: Essays on the Philosophy of Tyler Burge* (MIT Press), pp. 397-403.

Burge, T. (2006) "Postscript to 'Individualism and the Mental'". In Burge, *The Foundations of Mind* (Oxford University Press, 2007), 151-81.

Chalmers, D. (2002) "The Components of Content". In Chalmers, ed., *Philosophy of Mind: Classical and Contemporary Readings* (Oxford University Press, 2002), pp. 608-633.

Clark, A. and Chalmers, D. (1998) "The Extended Mind". *Analysis* 58: 7-19.

Davidson, D. (1970/1980) "Mental Events", in his *Essays on Action and Events*, Oxford University Press, pp. 207-224.

Davidson, D. (1987) "Knowing One's Own Mind". Proceedings and Addresses of the American Philosophical Association 60: 441–458.

Dretske, F. (1996) "Phenomenal Externalism". In E. Villanueva , ed. *Philosophical Issues* 7: Perception (Ridgeview Publishing).

Farkas, K. (2003) "What is Externalism?" *Philosophical Studies* 112: 187-208.

Farkas, K. (2008) *The Subject's Point of View* (Oxford University Press).

Fodor, J. (1980) "Methodological solipsism considered as a research strategy in cognitive psychology". *Behavioral and Brain Sciences* 3: 63-110.

Fumerton, R. (2003) "Introspection and Internalism", in S. Nuccetelli, ed., *New Essays on Semantic Externalism and Self-Knowledge* (MIT: Bradford Books).

Gertler, B. (2007) "Content Externalism and the Epistemic Conception of the Self". *Philosophical Issues* 17: 37-56.

Gertler, B. (2011) *Self-Knowledge* (Routledge).

Jackson, F. (2003) "Narrow Content and Representationalism - or Twin Earth Revisited." Patrick Romanell Lecture, *Proceedings of the American Philosophical Association*, 77: 55-71.

Ludlow, P. and N. Martin, eds., (1998) *Externalism and Self-Knowledge*, Stanford, CA: CSLI Publications, 21-83.

Ludwig, K. (1993) "Externalism, Naturalism, and Method". *Philosophical Issues 4: Naturalism and Normativity*. (Ridgeview Publishing), pp. 250-64.

Lycan, W.G. (2001) "The Case for Phenomenal Externalism". In J.E. Tomberlin, ed., *Philosophical Perspectives, Vol. 15: Metaphysics* (Ridgeview Publishing).

McLaughlin, B. and M. Tye (1998) "Is Content-Externalism Compatible with Privileged Access?" *Philosophical Review* 107: 349-380

Mill, J.S. (1863/2002) *Utilitarianism*. G. Sher, ed. (Hackett Publishing).

Putnam, H. (1975) "The Meaning of 'Meaning'". In K. Gunderson, ed., *Language, Mind, and Knowledge* (Minneapolis: University of Minnesota) pp. 131-193.

Segal, G. (2000) A Slim Book about Narrow Content (MIT Press).

Tye, M. (2000) Consciousness, Color, and Content (MIT Press).

Williamson, T. (2000) *Knowledge and Its Limits* (Oxford University Press).

Wilson, R.A. (2004) Boundaries of the Mind: the individual in the fragile sciences (Cambridge University Press).

Yablo, S. (1997) "Wide Causation". *Philosophical Perspectives* 11 (11):251-281.

---

[1] Of the 931 "target faculty" responses to the Phil Papers 2009 survey, 51.1% chose the response "accept or lean toward externalism". Interestingly, only 19.9% chose "accept or lean toward internalism"; 28.8% chose "other". *Source*: http://philpapers.org/surveys/results.pl

[2] I am not the first to notice these difficulties. Katalin Farkas (2003) and Richard Fumerton (2003) provide especially insightful discussions of them; I am indebted to both of these authors for helping me to appreciate the force of this problem. But while I regard these difficulties as fatal, Farkas and Fumerton each advance a proposal aimed to resolve them. I discuss their proposals in Sections 3 and 4, respectively.

[3] By "thought contents" I mean content *properties*, such as the property *thinking that p*. I will usually talk of such properties as properties of thinkers, but in Section 4 I will discuss content properties—such as *having the content p*—as properties of thoughts themselves. (I assume that thoughts just are instantiations of contents; "having the content *p*" serves as shorthand for *being an instantiation of p*.) Some standard formulations of internalism and externalism use "internal" rather than "intrinsic"; nothing will turn on my choice of terminology.

[4] Note that a viable criterion may allow for vagueness. The idea that there is a gray area, in which some properties are neither clearly intrinsic nor clearly extrinsic, is consistent with the availability of a general, principled criterion that distinguishes intrinsic from extrinsic features. By analogy: there is a general, principled criterion, along the lines of "having relatively few hairs on the head", that distinguishes those who are bald from those who are not bald. While this criterion is not specific enough to deliver a verdict in every case, it does explain why "bald" accurately describes Howie Mandel but not Oprah Winfrey. My search for an explication of "intrinsic to the thinker" would be satisfied by a similarly principled criterion, one that explains why "intrinsic to the thinker" accurately describes some properties but not others. A criterion could be adequate for this purpose even if it is less than maximally specific and hence fails to deliver a verdict in some cases. Indeed, a successful criterion may explain *why* certain cases are borderline.

[5] Burge says that his earlier formulation (in Burge 1986b) "misleadingly suggests that failure of local supervenience of intentional states on the individual's physical states is to be identified with anti-individualism." (Burge 2003a, 302)

[6] Burge seems committed to (2), or something very close to it. And he is at least attracted to the dualism expressed in (1): see especially Burge 2003a and 2003b.

[7] The Modified Spatiophysical Criterion also faces the usual difficulty with the "water" example: it classifies *inhabiting an environment in which the watery stuff is* H_2O as intrinsic to the thinker, since that property satisfies condition (i) of the criterion.

[8] This was Fodor's view in Fodor (1980).

[9] See Jackson (2003). This marks a change from Jackson's earlier dualism.

[10] I have in mind here the familiar idea that the predicate dualism advocated by Davidson (1970/1980) is really a kind of property dualism.

[11] This answer retains the desired ontological neutrality only on the assumption that being spatially located does not entail being physical. If this assumption is false, the objections to the Spatiophysical and Modified Spatiophysical criteria may also defeat the Spatial Criterion.

[12] The epistemic criterion also fits with the plausible idea that, as Farkas puts it, "what it is to have a mind" is tied to the epistemic.

[13] If Farkas' proposal is not purely exegetical—e.g., if she is instead proposing a subtle reorientation of the debate—our evaluation of the Epistemic Criterion will not constitute an objection to it.

[14] While the incoherence of phenomenal externalism followed from the idea that *all* phenomenal differences '*make a difference* to the way things appear', the second horn follows from the idea that *only* phenomenal differences make a subjective difference.

[15] This follows from their so-called Parity Principle: "If, as we confront some task, a part of the world functions as a process which, were it to go on in the head, we would have no hesitation in accepting as part of the cognitive process, then that part of the world is (for that time) part of the cognitive process." (Clark and Chalmers 1998, 8)

[16] Some versions of vehicle externalism, including the version advanced in Clark and Chalmers (1998), claim only that *dispositional* states—such as standing beliefs—are sometimes "extended". But most versions of content externalism concern *occurrent* states. I will ignore this complication.

[17] Whether vehicle externalism qualifies as externalism, using the Epistemic Criterion, hinges on whether thinkers enjoy privileged access to "extended" mental states. Clark and Chalmers argue that it would be question-begging to deny that consulting a notebook to ascertain what one believes, say, is an introspective process. Be that as it may, vehicle externalism is most plausible as regards dispositional (non-occurrent) attitudes. Insofar as there is good reason to think that we enjoy privileged access only to occurrent thoughts and attitudes (as I argue in Gertler 2011, ch. 3), vehicle externalism counts as an externalist view, according to the Epistemic Criterion.

[18] According to the Vehicle Construal, the internalist is committed to denying that content properties consist in relations to factors outside the thought but intrinsic to the thinker. While Descartes' view seems amenable to this position, it's not clear to me that it is committed to it.

[19] A total realization may suffice for the property's being instantiated only on the assumption that certain background conditions are in place (Wilson 2004). I ignore this complication.

[20] This point brings out an obstacle faced by content externalism, one which is not faced by internalism or by vehicle externalism. Both internalism and vehicle externalism can draw the boundary of the thinker, and her thoughts, at the boundary of total realizations. So neither of these views depends on some *other* way of delineating thinkers. By contrast, the standard version of content externalism construes thought contents as relational features of the thinker. So it must delineate thinkers in some other way. (In Gertler 2007, I argue that this obstacle is insurmountable,

as there is no way of delineating the thinker that meets externalist requirements while preserving our basic conception of thinkers.)

[21] I expect that the resulting conception of the thinker and her boundaries will be some sort of epistemic conception. In other words, I agree with Farkas that "what it is to have a mind ... is a thoroughly epistemic notion", understanding "epistemic" as encompassing the phenomenal (as she does). The arguments of Section 3 show that that result will constitute a significant departure from the current debate.

[22] I presented an ancestor of this paper at the Australian National University, in January 2010, and received helpful feedback. For discussion or comments on earlier versions of this paper, I thank Anita Avramides, David Chalmers, Katalin Farkas, John Maier, Lisa Shabel, Susanna Schellenberg, Daniel Stoljar, and especially Trenton Merricks.