Color Terms
and Semantic Externalism

ÅSA WIKFORSS
Department of Philosophy,
Stockholm University

The paper discusses whether the color terms should be given an externalist semantics. In the literature on the semantics of color terms externalism is standardly taken for granted, and Twin Earth style arguments play a central role. This is notable given that few people would claim that semantic externalism applies across the board, to all types of terms. Why, then, should the color terms belong with this group of terms? I argue that the standard externalist strategies, introduced by Tyler Burge and Hilary Putnam, do not apply to these terms: The color terms do not function like natural kind terms, and the idea of semantic reliance on others does not apply to them. I conclude that the externalist arguments fail and that a version of internalism, more properly called 'individualism', applies to the color terms.

The color properties, it is often pointed out, occupy an interesting position between the fully subjective and the fully objective. Unlike properties such as being gold or having mass, they are visually presented to us in the sense that we can (normally) determine the color of an object simply by looking. There is thus a close connection between looking red and being red, between how we experience the property and the property itself. At the same time, colors are (prima facie) properties of external objects: Unlike subjective properties such as pain, colors are not properties of our mental states but of objects in the external world. Moreover, there is an indisputable gap between appearance and reality. Although looking red is good evidence that an object is red, there are situations (familiar to all of us) when an object looks red without

being red, and when an object is red without looking red.

This mixture of objective and subjective elements constitute the focal point of the contemporary debate over the nature of color. Can colors be understood as ordinary physical properties of objects, or do we need to appeal to the role of the subject, making color a relational property?[1] The same mixture underlies competing theories about the semantics of the color terms: Should we take these terms to denote external, objective properties, or do they rather pick out subjective properties, or perhaps some combination of both? In this paper I shall address one particular aspect of this debate, one that I do not think has been sufficiently discussed: Should the meaning of the color terms be construed externalistically? In the literature it is commonly assumed that this is the case, and Twin Earth style arguments play a central role in the debate. An early example is Ned Block's thought experiment involving Inverted Earth where the assumption that the individual's "physical and linguistic environment" determines the meaning of the color terms plays a central role. [2] At the same time, it is far from clear whether the standard arguments for semantic externalism apply to the color terms, and little effort has been made in supporting the externalist conclusions in the case of these terms. Instead, meaning externalism is commonly taken for granted, and from this conclusions are drawn concerning the content of color experiences (as well as, in some cases, their phenomenological qualities). This is notable given that few people would claim that semantic externalism applies across the board, to all types of terms. Why, then, should the color terms belong with this group of terms?

In what follows I shall examine whether, indeed, the color terms could plausibly be said to have an externalist semantics. Although I am on the record for being skeptical about semantic externalism in general, I will set the skepticism aside in this paper and assume that externalism does apply to certain classes of terms (such as the natural kind terms). My question, then, is whether the color terms should be included in this category. I shall argue that the standard externalist strategies, introduced by Tyler Burge and Hilary Putnam, do not apply to these terms. The paper is divided into four main sections. The first spells out the necessary background, characterizing the relevant externalist theses. The second section considers the proposal that the color terms can be treated along the lines of natural kind terms, and that a version of physical externalism applies to them. In the third section I discuss a complication that arises in the case of the color terms when it comes to using Twin Earth thought experiments. The final section discusses the proposal that the color terms are deferential terms

[1] For physicalist accounts see for instance Byrne & Hilbert 2003, and Jackson 1998. For a defence of relational accounts see Cohen 2004.

[2] Block 1990. See also Bain 2007, Boghossian & Velleman 1991, Macpherson 2005, Spohn 1997 and Tye 2000.

and that, therefore, social externalism applies to them. I conclude that the externalist arguments fail and that a version of internalism, more properly called 'individualism', applies to the color terms.

1. *What is Semantic Externalism?*

Since 'semantic externalism' was first introduced in the 1970's it has come to refer to a wide variety of positions. If one were to spell out the common denominator of these positions, one's best bet would be in terms of some kind of 'failure of supervenience' thesis, according to which meaning facts fail to supervene on internal properties of the speaker. Meaning facts are facts about the meaning of a word as used by a speaker at a time, for instance:

(i)   'Water' means *water* for speaker S at time t.
(ii)   'Arthritis' means *arthritis* for speaker S at time t.

The purpose of the various Twin Earth scenarios is to trigger the intuition that there are indeed such failures of supervenience: An individual on Earth and her twin on Twin Earth are 'internally identical', and yet their respective terms 'water' ('arthritis') have a different meaning. How precisely the externalist thesis is to be understood, therefore, depends crucially on how the relevant internal facts are to be spelled out. Unfortunately, there are difficulties drawing the internal-external distinction in a way that avoids begging any questions. I shall return to this issue below since it causes particular problems when it comes to the color terms. However, if we leave all these complications aside we can say, as a first approximation, that semantic externalism (in all its versions) implies that meaning facts such as (i) and (ii), fail to supervene on internal properties of individual speakers. And the question to be pursued in this paper, consequently, is whether meaning facts involving color terms, such as (iii), fail to supervene on internal properties:

(iii)   'Yellow' means *yellow* for speaker S at time t.

To set the stage, let us briefly recall how Putnam uses his famous thought experiment. Putnam's target is what he calls 'the traditional theory of meaning', spelled out in terms of two central theses:

(1)   Knowing the meaning of term is just a matter of being in a certain psychological state. This psychological state determines the intension/meaning of the term that, in turn, determines the extension. Hence, sameness in psychological state implies sameness in meaning and extension.

(2)   The meaning/intension of a term determines its extension.

The purpose of the thought experiment is to show that "these two assumptions are not jointly satisfied by *any* notion, let alone any notion of meaning" (1975: 219). Oscar on Earth, and his Twin on Twin Earth both use the term 'water' competently and are psychologically type-

identical. And yet, intuition would have it, their respective terms have different extensions, they pick out different properties. This is the evidence that our semantic theory is supposed to account for.

Putnam argues that the best way to account for this evidence is by rejecting (1) and 'externalizing' meaning. He considers the possibility of rejecting (2) instead, thereby retaining the traditional idea that meanings are 'in the head', but rejects this option. This, Putnam argues, would be the right move for an *absolute* indexical like 'I', but "it seems incorrect for the words we have been discussing" (1975: 245). It is therefore preferable, according to Putnam, to reject (1) and to conclude that Oscar and his twin mean something different by 'water'.[3] However, he suggests, even if natural kind terms cannot be construed as indexicals, there is an 'indexical component' in how the meaning of natural kind terms is *determined*, since ostensive definitions play a central role. For example, we point to a puddle and say 'this liquid is water', intending 'water' to pick out everything which stands in the relation of being 'the same liquid' to the initial samples. In the case of natural kind terms, Putnam argues, the relevant sameness relation is microstructural and to be determined by science. Since on Twin Earth the liquid pointed to by Oscar's twin, Toscar, has a different chemical composition, Putnam argues, their terms have a different extension and, as a result, a different meaning (1975: 231).[4]

That Putnam chooses this strategy implies that he accounts for the Twin Earth intuitions by appealing to foundational semantics, considerations having to do with the *determination* of meaning, rather than by appealing to assumptions about the semantic content of these terms, their 'descriptive semantics'.[5] The conclusion is that the meaning of natural kind terms is not fully determined by the speaker's associated descriptions etc. but also by features in her external environment, such as facts about chemical composition. For this reason I have elsewhere labeled this type of externalism *foundational externalism*.[6]

---

[3] It should be noted that some people disagree and have suggested precisely that Putnam's thought experiment shows that natural kind terms function like indexicals (see for instance Donnellan 1993 and McKinsey 1987). The upshot is a mixture of meaning internalism (where meaning is construed along the lines of Kaplanian character) and content externalism (since the full propositional content is contextually determined).

[4] There is a complication here derving from the fact that natural kind terms are general terms. It is clear that the extension of a general term may vary from one world to another without there being any reason to say that there is a difference in meaning. For instance, there are worlds in which 'tiger' has a much smaller extension than in the actual world (because tigers have been hunted to the point of extinction say). The claim can therefore not simply be that 'water' has a different extension on Twin Earth; rather, the claim must be that it predicates a different property. For a discussion, see Wikforss (forthcoming).

[5] For the distinction between foundational semantics and descriptive semantics see Stalnaker 1997.

[6] Wikforss 2008. It should be stressed that foundational externalism, being a

This is also the type of externalism defended by Burge, although he appeals primarily to the role of the speaker's *social* environment when it comes to meaning determination.[7] Burge draws attention to the fact that individuals in a linguistic community intend to use their terms in accordance with how other members of their community use them. Evidence for this, he argues, comes from the fact that people stand corrected when they deviate from standard use. Burge's central claim is that such correction should not be seen as involving a *change* in meaning, but as involving *semantic* deference. If it did involve a change in meaning we would have to say that the terms used by a deviant individual, such as Bert who utters 'I have arthritis in my thigh', have a deviant meaning and that the individual has said something *true*. According to Burge there is a strong intuition that what Bert says is false. This shows, Burge argues, that the term 'arthritis' as used by Bert had its standard meaning, also prior to correction, even though he had an incomplete grasp of this meaning, and believed that 'arthritis' applies to rheumatoid diseases of the joints as well as the ligaments. The meaning of the term is therefore not determined by facts about Bert alone, but by the larger practice of the community, the experts. Consequently, Burge concludes, meaning facts do not supervene on facts about individual use. In the counterfactual world, where the community uses 'arthritis' to apply in accordance with Bert's use (to rheumatoid diseases of the joints and the ligaments), the meaning of Bert's term would differ, and he would have said something true when he utters 'I have arthritis in my thigh'.[8]

Underlying Putnam's and Burge's versions of externalism, therefore, are two distinct ideas:

(Nk) If a term is used as a natural kind term, then its meaning is determined by underlying physical features of the samples involved in ostensive definitions of the term.

(Sd) When a speaker defers, semantically, the meaning of her term is determined not by facts about how she uses the term, but by facts about other people's use of the term.

The meaning determining facts in the first case are physical, in the second social, but the upshot, is the same: Meaning facts are determined by facts external to the individual speaker; consequently such facts fail to supervene on internal facts.

As is well-known, Burge also goes a step further and suggests that the same holds for *content* facts, such as (ii)*:

(ii*) Bert thinks a thought with the content *I have arthritis in my thigh*.

---

theory about meaning determination, is in itself neutral on how we are to construe the semantic content of the relevant term.

[7] Burge 1979.

[8] Burge 1979: 77–85.

Although Putnam initially was unclear on the implications of meaning externalism for mental content, he soon came to concur with Burge: If the meaning of a term is determined externally, then so is the corresponding thought content.[9] As a result, the original externalist considerations were taken to show not only that meaning facts fail to supervene on internal properties, but also that content facts fail to do so. This is of particular importance for the color terms, since it has implications for the content of color experiences. If meaning externalism implies content externalism, and if experiential content is construed as involving full-fledged intentional content, then an externalist account of the meaning of the color terms, has the surprising implication that the content of color experiences does not supervene on internal properties of the subject: Two subjects may be internally identical, and yet their experiences have a different content. It has even been suggested that we can take this a step further, and extend the externalism not only to the content of experience but also to its phenomenological character. Thus, according to representationalists, phenomenological character depends on representational content. Consequently, two subjects may be internally identical, and yet have experiences with different contents and a different phenomenology, simply as a result of differences in their external environment.[10]

In what follows I shall focus on meaning externalism, since the standard route to content externalism proceeds on the assumption that meaning externalism is true. If this assumption does not hold for the color terms, this particular route to content (and phenomenological) externalism is blocked. First, let us consider the role of the physical environment.

2. *Physical Externalism*

The most obvious strategy if one wishes to defend an externalist account of the color terms would be to pursue the idea that these terms are natural kind terms, just like 'water' and 'gold'. Prima facie such a proposal does not seem wholly implausible. As noted above, there is a strong intuition that color terms pick out objective properties, and there is a sophisticated science concerned with the nature of colors and color perception. Indeed, after having presented his case for externalism, against traditional theories of meaning, Putnam suggests that

---

[9] Putnam 1988. It should be noted that Burge simply identifies the linguistic meaning of a term with the concept expressed and, moreover, takes mental content to be composed of concepts. On Burge's view, thus, meaning externalism trivially entails content externalism.

[10] See for instance Tye 2000. It is common to distinguish between strong representationalism, which identifies phenomenological character with experiential content, and weak representationalism, according to which phenomenological character merely supervenes on experiential content. On either view, externalism about phenomenological character relies on the assumption that color contents are externalist.

a similar account can be applied to 'other parts of speech', including adjectives like 'red' (1975: 244). Kripke, likewise, suggests that color terms resemble natural kind terms in that they rigidly designate certain external properties.[11]

Let us start, therefore, by trying to design a color thought experiment that runs parallel to Putnam's experiment, involving Color Twin Earth. Imagine Oscar and his twin on Color Twin Earth, Toscar, both using the term 'red' competently. In the original Twin Earth scenario, they are supposed to apply their respective terms in the same way 'superficially', to the wet, thirst-quenching, odorless liquid that comes out of taps and fills lakes and rivers, etc. To make a parallel case for color terms, then, we would have to imagine that people on Twin Earth apply 'red' to tomatoes, fire engines, roses and blood, just like we do.[12] Moreover, we shall have to assume that if Oscar were to visit Color Twin Earth we would not notice the difference, just as in the original Twin Earth case. In an intuitive sense, therefore, the objects 'look' the same on both planets (there are complications here, and we shall return to them shortly). However, after a while scientists discover that on Color Twin Earth what causes these experiences have different properties than what causes them on Earth. For instance, the causes are not reflectance properties, but some other surface property.[13] Do we have the intuition that 'red', on Twin Earth expresses a different property than 'red' on Earth? That there are no red objects on Twin Earth?[14]

In this case, I venture to guess, most people would say no. After all, the term 'red' is applied to the same set of objects that we apply it to (tomatoes, fire engines etc.), and no one can tell the difference (just by looking) between the planets. That the objects in question have different physical properties does not seem to warrant the conclusion that 'red' picks out a different property in Twin-English. Indeed, it is a well-known fact that right here, on Earth, there is not one physical property that is common to the objects we would normally classify as 'red'. Thus, there is the problem of metamers, the fact that objects with different reflectance properties can look the same to observers in given conditions, as well as the simple observation that the objects we clas-

---

[11] Kripke 1980: 128, fn 66, and134.

[12] It is worth noting that Block's Inverted Earth differs from Color Twin Earth in this respect. Block simply suggests that because of the physical differences between the two planets, everything has the complementary color on Inverted Earth, and the inhabitants have an inverted color vocabulary. Block's thought experiment can therefore not be used to test intuitions concerning whether the color terms function like natural kind terms (indeed, Block's thought experiment *presupposes* physicalism). Of course, this is not Block's intention either. Rather, his intention is to set up a scenario that shows intentional content and phenomenological content to be distinct (1990: 61–62).

[13] For a defense of the claim that the colors are identical to certain reflectance types see Armstrong 1991, Byrne & Hilbert 2003, and Tye 2000.

[14] Waving irrealist intuitions according to which there are no red objects on Earth either.

sify as 'red' ('blue', 'green', etc.) include a class of varying reflectances.[15]
Colors, it is therefore often said, are multiply realizable: Although all
red things have the property of redness in common, this property can
have different physical realizations.

This suggests that the color terms do not function like natural kind
terms. What is distinctive of the natural kind terms, according to the
standard Kripke-Putnam story, is that they are used with the inten-
tion of picking out underlying micro-physical properties, which serve
to explain the macrophysical behavior of instances of the kind. This
allows for inductive generalizations from one sample of the kind to oth-
ers: Studying the distinctive behavior of samples of water, we expect
instances of the kind to have the same set of causal properties in virtue
of the shared underlying microphysical property. Moreover, it implies
that the observable properties associated with the kind are merely con-
tingently related to it. Although water on Earth is a wet, thirst-quench-
ing, potable liquid, there are possible worlds where there is water that
does not have these properties (a world in which H2O is pink and solid
for instance), and there are worlds where there is a liquid that has all
these properties without being water (as on Twin Earth). For this rea-
son there is no such thing as multiple realizability of a natural kind: If
it is discovered that there is not one underlying, explanatory property
picked out by the term, but two, the conclusion is not that the kind is
multiply realized, but that the term is ambiguous and is used to pick
out two distinct kinds.[16]

In the case of the color terms, by contrast, we do not seem to have
the intention to pick out underlying properties. For instance, we do not
expect all instances of red to share causal properties that can be used for
inductive generalizations concerning all red objects.[17] This is also evi-
denced by the fact that we allow for multiple realizability. When we find
out that the objects classified as 'red' (by normal observers in normal
circumstances) do not share a certain reflectance property, we do not
conclude that the term is ambiguous but, rather, that the property of be-
ing red can be multiply realized. Similarly, the notion of 'fool's red' is dif-
ficult to make sense of, the idea that we have mistakenly thought that
there is a class of objects (tomatoes, say) that are properly classified as

[15] See Byrne & Hilbert 2003 for a discussion of these and other cases. They put
it: "Surfaces with grossly different reflectances can perceptually match even under
fairly normal circumstances" (2003: 11).

[16] A famous example is 'jade' which picks out two distinct minerals, jadeite and
nephrite. See Soames 2002: 281–284, for a discussion of these cases. Soames suggests
that if it is discovered that there is not one underlying microphysical property but a
multitude, the conclusion may even be that the term fails to refer. See also Jackson
1998: 105–112.

[17] See Cohen 2007 who argues that the color concepts are not used to pick out
underlying essences. Our inferential treatment of *beeing red*, he argues, "seems
not to involve making inductive generalizations about red things" and this shows
that we are not committed to the existence of any shared constitutive ground of red
things (2007: 8).

being red, without being red. Of course, something may look red without being red if the observation conditions are not favorable, but if an object looks red under normal conditions then we conclude that it is red. It is therefore to be expected that the Color Twin Earth scenario would not provide intuitive support for externalism but, rather, be treated as an illustration of the fact that colors are multiply realizable.

It may be objected that this is too quick and that multiple realizability in itself does not settle anything since it is compatible with physicalism and, hence, with externalism. Thus, it has been argued by Frank Jackson that the physicalist can simply say that the property picked out by 'red' is a *disjunctive* property (1998: 108).[18] Physicalists, he suggests, should say that yellowness is the disjunction of all the physical properties that make things look yellow in the same way (for instance, both the property responsible for the canary looking yellow and the property responsible for the picture of the canary looking yellow). Jackson acknowledges there must be some limits on what these disjuncts can be; otherwise the position reduces to a version of dispositionalism, the idea that colors are second-order properties, properties of first-order physical properties to be disposed to cause certain sensations. There will therefore have to be "sufficient similarity between what typically makes things look red to allow us to identify red with a disjunctive property that is sufficiently unified to count as a cause" (1998: 108). If there is not, Jackson argues, then we would have a case similar to that of 'jade': "we should say that the red of sunsets is a different property from the red of tomatoes just as New Zeeland jade is a different kind from Chinese jade" (1998: 112). Jackson even suggests that if the assumption of physical unity fails more radically, if there are too many distinct physical properties involved, we would have to declare color "a pervasive illusion" (ibid.).

This, indeed, is what we should say if Jackson's physicalism is right. However, the reflections above suggest that we would not. Since the claim of multiple realizability is primarily guided by how we spontaneously classify objects (in normal viewing conditions) it is unlikely that we would treat 'red' like 'jade' in Jackson's scenario. Rather, we would say that we have found yet another physical realization of the property redness, which, in turn, would support a dispositionalist account of the property, rather than a disjunctive version of physicalism. Similarly, it is unlikely that we would conclude that colors are an illusion simply because the assumption of physical unity fails. Of course, there are theoretical arguments to the effect that colors are an illusion. But what we are looking for is evidence concerning the semantic intentions of ordinary speakers, and it is implausible that ordinary speakers would take the failure of the assumption of physical unity to show that the color terms do not pick out a property.[19]

[18] A similar line is defended by Byrne & Hilbert 2003.
[19] Compare with the case of 'sand', which does not pick out anything physically

Jackson, in fact, expresses a certain ambivalence on this point. Although he suggests that 'red' should be treated just like 'jade' in case the assumption of underlying unity fails, he also notes that there are important differences between natural kind terms and color terms. In the case of 'water' and 'gold' he argues, we take for granted that there is something important that might properly be regarded as a natural kind and it is therefore plausible he suggests, that it is part of the meaning of these terms that they denote kinds. In the case of color terms, by contrast, the "folk are too sensible to have presupposed something as risky as that there is a distinctive kind in common to things we call 'red'" (1998: 108). If so, however, why should we endorse even the disjunctive version of physicalism? Physicalism, also in its disjunctive version, does after all require the 'risky' assumption that Jackson speaks of.[20]

Jackson expresses a similar ambivalence about the modal cases, involving cross-worlds scenarios. The Kripke-Putnam claim, again, is that we use the natural kind terms with the intention of picking out a certain underlying microphysical property across all possible worlds. This underlies the thesis that the natural kind properties are *identical* to these underlying properties; for instance, that the property of water is identical to the property of H2O. While we may rely on superficial properties (being wet, transparent, thirst quenching liquid, etc.) in fixing the reference of the term, it is argued, these properties are only contingently related to the kind. Similarly, if we use the color terms like natural kind terms, we use them with the intention of picking out the physical property that, in this world, causes certain experiences. It follows that 'red' picks out the same physical property (or the same disjunctive physical property) across all possible worlds, and that there are worlds in which red objects do not look red (as well as worlds in which something looks red without looking red). This is what Jackson is committed to. As he notes, "it is indeed a consequence of the causal theory that redness, for instance, is not essentially linked to being red" (1998: 101). At the same time, Jackson grants that speakers of natural language 'vacillate' on this issue, and are not committed to saying that the manifest properties of the colors merely serve as reference fixers. If he is right about that, however, speakers do not treat the color terms as natural kind terms, contrary to Jackson's own assumption.[21]

unified. This was not taken to show that 'sand' fails to refer but merely that it does not pick out a natural kind.

[20] It is of interest to note that Kripke expresses a similar ambivalence. Although he does claim that the color terms resemble natural kind terms, he also suggests that they differ from these terms. Thus, after arguing that terms such as 'gold', 'tiger' and 'water' are natural kind terms that pick out non-manifest, underlying properties and that, therefore, Mill was incorrect to suggest that general terms differ semantically from names in that they express properties, Kripke suggests that *perhaps* Mill was right when it comes to some general terms, such as 'foolish', 'fat' and 'yellow' (1980: 127).

[21] Surprisingly, Jackson himself does not want to commit on the cross-worlds issue, but wants to remain a 'fence-sitter' on whether the relevant causal roles fix

There are therefore reasons to think that the color terms are not used like natural kind terms. When we consider how a color term applies across possible worlds what is decisive is not the physical properties of the objects, but how they are experienced, which suggests that the color terms are not used to pick out physical properties but, rather, second-order dispositional properties. If this is correct it is difficult to see how physical externalism could apply to these terms. The reason many people accept physical externalism in the case of the natural kind terms, again, is precisely the idea that since we use these terms with the intention of picking out an underlying physical property, causal interaction with instances of the property allow us to pick it out even before we have full knowledge of it. For instance, it is because we intended 'gold' to pick out such an underlying property, that our interaction with instances of gold serve to determine that 'gold' applies only to instances of the element Au with the atomic number 79—long before we knew anything about the essential nature of gold.[22] If reflection on our use of the color terms shows that we do not use them with such an intention, then there is little reason to think that the meaning of these terms is determined externalistically: The physical properties of the objects we have causally interacted with when applying 'red', do not determine the correct application of the term.

At the same time the close connection between the color properties and how they are experienced poses certain difficulties when it comes to considering counterfactual scenarios, and relying on Twin Earth thought experiments. Before going on to discuss social externalism, let us consider these difficulties.

3. *Interlude: Twin Earth Thought Experiments and the Color Terms*

It was noted above that semantic externalism is standardly characterized in terms of a *failure of supervenience* thesis: As the thesis that meaning facts fail to supervene on internal properties. It is therefore essential that the relevant distinction between internal and external properties can be spelled out in a non-question begging way. However this turns out to be more difficult than anticipated.

In Putnam's original set-up Oscar and his twin are said to be internally identical in the sense of being in the same narrow psychological states: They may differ only in their wide psychological states, where a 'wide' state is a factive state or an otherwise object-dependent state. The wide states, then, do not include ordinary *de dicto* attitudes and hence Oscar and his twin are said to have the same *de dicto* beliefs, de-

the reference or give the meaning of the color terms: "I say nothing about whether things with these properties count as red in these worlds…" (1998: 101). This is not compatible with his physicalist position since that requires the assumption that the descriptions function as mere reference fixers.

[22] This is quite explicit in Putnam 1975: 235. See also Kripke 1980: 118–119.

sires, etc. This, as is well known, presupposes limiting the externalism to meaning, since once it is extended to content the twins will not be in this sense psychologically identical: Oscar will have beliefs about water, his twin about twater, etc. Since most contemporary externalists endorse both meaning and content externalism an alternative characterization of the sense in which Oscar and his twin are 'internally identical' is required. However, it is far from clear that there is a distinction to be drawn that is both useful in eliciting intuitions, and avoids begging any central questions.[23]

This is a problem for content externalism in general, but when it comes to color terms the problem seems particularly pressing. The trouble derives from the close connection between the color properties and how they appear to the subjects. As Jackson points out, it is a truism that colors are 'visually conspicious' in the sense that if an object is red then it is disposed (under normal conditions) to look red (1998: 88–89). The question then arises how we are to analyze the content of a color experience, such as *that x is red*. Starting from the truism, we get that this content is to be analyzed as *that x is disposed to cause experiences with the content that x is red*, which is obviously circular.[24] It is much disputed how to avoid this problem. What is of interest from the point of view of externalism, is that it poses difficulties employing Twin Earth thought experiments as evidence for or against externalism.

The problem should be obvious: When considering a counterfactual scenario, such as Color Twin Earth above, we cannot specify the content of the subject's experiences without begging the central question. For instance, we cannot say that that when Toscar looks at a tomato he has the same color experiences as Oscar on Earth when he looks at a tomato, since if we say this we have already presupposed that tomatoes have the property of being red on Color Twin Earth. Instead, we must characterize the sense in which Oscar and his twin are 'internally the same', without appealing to how they experience the objects. But what we need to know in order to draw any conclusions from the thought experiment concerning the color terms, it would seem, is precisely how Oscar and his twin *experiences* things. In this respect the difficulties involved in drawing a useful distinction between internal and external properties go beyond the difficulties encountered in the case of natural kind terms. While we could provide a neutral description of Oscar's experiences of the manifest properties of the liquid called 'water' in his environment (as being wet, transparent, thirst-quenching, etc.) no such neutral description can be given of Oscar's experiences of the manifest properties of the color called 'red' on Color Twin Earth: Any such description will already have settled the very question to be settled.

[23] For an interesting discussion see Farkas 2003 and 2008.

[24] See Boghossian & Velleman 1991: 83, and Glüer 2007. As Glüer makes clear the problem is not simply circularity but that any such attempt to specify the content of color experiences sets off a vicious regress.

The question, therefore, is whether we can spell out the sense in which the twins are internally identical without begging any questions. One proposal would be to appeal to internal *physical* factors: Oscar and his twin are internally identical, it might be said, in the sense of being physically identical 'from the skin in'. The trouble with this, however, is that it is very difficult to see how describing Oscar and his twin this way could serve to trigger any intuitions at all about the meaning of 'red'. If we are told that Oscar and Toscar are physically identical in this sense, how could it provide us with sufficient information to draw any conclusions at all about the color properties and the color terms? By describing the twins as internally identical in this sense, it seems, we have deprived the thought experiment all value as an 'intuition pump'.

A more promising option is to characterize the twin situation in terms of 'subjective indistinguishability': If Oscar were moved to Twin Earth (unknowingly) he would not notice the difference. This is how I described the thought experiment above, and it captures the idea (stressed by Putnam) that Oscar's and Toscar's situations are indistinguishable from the first person point of view. There are complications here as well. A natural way of thinking of subjective indistinguishability is in terms of phenomenological character: Oscar will not notice the move since it does not affect the phenomenological character of his color experiences. [25] While this does not beg any questions with respect to content externalism, it does beg the question against representationalism according to which the phenomenal character of a color experience is identified with its intentional content.[26] Even if representationalism is a very controversial thesis, it would clearly be preferable not to assume its falsity in the very characterization of Twin Earth scenarios.

To avoid these difficulties, I propose, we should think of 'subjective indistinguishability' in more objective terms, appealing not to the sameness of phenomenological character but rather to Oscar's behavior. Thus, Oscar does not notice the switch in the sense that he does not in any way manifest a surprise, and does not change his categorizations of objects as being 'red', 'blue' etc. This does not beg any questions against the physicalist, since the claim is not that Oscar experiences the tomatoes as red or, even, that the phenomenological quality of his tomato-experiences stays the same. The question, then, is whether 'red', as used by Oscar, has a different meaning simply in virtue of the physical differences between the two planets: Physical differences

[25] For a defence of a version of internalism based on such a characterization see Farkas 2008.

[26] The question is only begged against strong representationalism, though, since according to weak representationalism the dependence relation is merely one of supervenience: Supervenience implies that there cannot be a difference in phenomenological character without a difference in intentional content, but not that sameness in phenomenological character entails sameness in intentional content.

that do not lead Oscar to change his categorizations.[27] This provides a sufficiently theory-neutral test of whether the term is a natural kind term. If 'red' were a natural kind term, like 'water' in the original Twin Earth thought experiment, the sameness in Oscar's categorization dispositions would not support the conclusion that there is a sameness in meaning, given the fact that there is a difference in the underlying physical properties of the objects classified as 'red' on the two planets. My contention, again, is that in the case of Color Twin Earth intuition goes the other way: The sameness in categorization dispositions does support the conclusion that that there is sameness in meaning. Hence, 'red' is not a natural kind term.

If this is correct it suggests that meaning facts involving the color terms supervene on facts about the individual's use of these terms, more precisely, on the individual's categorization dispositions. Such a view qualifies as internalist in one important sense of the word: It takes meaning to be determined by facts about the individual, not by facts about her environment.[28] Still, it is not internalist in the stronger sense of taking meaning facts to supervene on that which is merely subjectively available. Thus, if one rejects representationalism (as internalists tend to do), and one wishes a purely subjective notion of indistinguishability, one can hold that meaning facts supervene on phenomenal character with the consequence that meaning facts are as accessible from the first-person point of view as are phenomenal facts (such as the phenomenological character of a pain). Even waving the issue of representationalism this type of internalism is problematic, I believe, since there is little reason to think that linguistic meaning supervenes on phenomenological character. Meaning facts do not have a phenomenological character in themselves, and there are meaning differences that are not detectable simply via introspection but require considering how one uses a term in actual and counterfactual scenarios.[29] For these reasons the type of internalism defended here is perhaps better labeled *individualism* (following Burge 1979).

[27] A natural explanation of the fact that Oscar categorizes the objects in the same way is that there is no difference in the phenomenological character of Oscar's color experiences. The person who adopts physical externalism combined with strong representationalism, notice, is unable to provide such an explanation, since it follows from this combination of views that Oscar's color experiences on Twin Earth have a different phenomenological character.

[28] Unless one adopts representationalism, this account of meaning and content determination is compatible with inverted spectra scenarios, where two subjects have the exact same set of categorization dispositions (and, by implication, the same color experiences) while the phenomenological quality of their experiences is reversed. I think that allowing for the possibility of such a scenario is necessary consequence of taking meaning to be determined by publicly available facts, but I will not be able to discuss the issue further here.

[29] The same, arguably, holds for content facts, and the idea that we can tell the sameness and difference of thought contents purely introspectively. I argue against transparency of thought content in "The Insignificance of Transparency", submitted (MS).

Naturally, people sometimes make mistakes about the color of an object, and this would have to be taken into account in the full story of how facts about individual use determines meaning. However, this holds for all terms, from theoretical terms to artifact terms: Under certain conditions we make mistakes. Any plausible theory of meaning will have to allow for this. This requires, at a minimum, that the function between use and meaning is construed as a many-one function, such that differences in use are compatible with sameness of meaning. It also requires appealing to some meaning determining principle, telling us how to map facts about use on to meaning facts.[30] For example, one might appeal to some version of the principle of charity, such as the principle that one should minimize inexplicable error. What is characteristic of the color terms, arguably, is that the mistakes derive primarily from facts having to do with the subject's perceptual situation: Her perceptual apparatus (such as the lenses) and the lighting conditions. If these perceptual conditions are normal, and if she applies 'red' in a deviant way, the proper conclusion is that her term has a different meaning. This is another way of illustrating the point often made, that although the colors are visual or manifest properties, they are not manifest under all visual conditions. Our use of the color terms displays awareness of this. We know what to do to determine the color of an object, such as holding it up in the light, holding it against different backgrounds, manipulating it in various ways (rotating it, cleaning it, drying it, etc.).

However, it is notoriously difficult to say anything very specific about the nature of 'normal conditions' or, even, about the nature of normal observers.[31] This raises difficulties concerning how to think about the meaning of the color terms in counterfactual cases where the perceptual conditions are stipulated not to be normal. Consider Block's Inverted Earth, where we are supposed to form intuitions about the color experiences of an individual with inverted lenses on Inverted Earth. On Inverted Earth, recall, everything is stipulated to have the complementary color of Earth (the sky is yellow, grass is red, fire hydrants green, etc.) and the color vocabulary is inverted in the same way ('yellow' means *blue*, 'green' means *red*, etc.). Block then imagines a switching scenario in which a subject from Earth has inverted lenses inserted (unknowingly) and moved to Inverted Earth. Waking up on Inverted Earth the subject does not notice any difference. Initially, Block suggests, the individual's words retain their old meaning, so that when he utters 'The sky is blue' he has said something false. Block then appeals to externalism to argue that after a while, there is a switch in the meaning of the subject's terms, and in the corresponding contents, as a result of the difference in the physical and linguistic environments.

---

[30] For a discussion of these matters see Pagin 1997.

[31] See Cohen 2004 and Hansen 2011. As Hansen notes, what conditions are relevant for determining the color of an object may vary depending on what one is looking for and why (2011: 212)

It follows that the individual's utterance 'The sky is blue', etc. will be true, not false. The conclusion, according to Block, is that there can be differences in intentional content without a corresponding difference in phenomenological character.[32]

One trouble with this thought experiment, I think, is that it is far from clear what we are to say about the meaning of the color terms as used by an individual with inverted lenses. On Earth, clearly, such an individual would deviate from our use of the color terms (he would classify the sky as 'yellow' and the grass as 'red'). Would we conclude that his terms have a different meaning, or would we rather conclude that he is massively mistaken? On the one hand his categorization dispositions differ radically from ours, which provides a reason to say that there is a meaning difference. On the other hand, we know that his perceptual conditions are abnormal: He has had inverted lenses inserted and if those are taken out he will go back to his normal usage. The same holds after he is moved to Inverted Earth, only the other way around. The individual's categorization dispositions cohere with those of the speakers on Inverted Earth, which provides a reason to say that his color terms has the meaning of their language, but at the same time his perceptual conditions are not normal: He has inverted lenses and if those are taken out he will go back to his normal Earth use, which is to say that he will use his color terms in a way inverted to the speakers on Inverted Earth. It is to be predicted that we will have conflicting intuitions about such cases, and this makes the thought experiment of dubious use in settling our externalist intuitions. If anything, I think, Block's thought experiment illustrates the dependency of our color terms on normal background conditions and the indeterminacy that results when those conditions are altered.

I therefore think that physical externalism does not apply to the color terms and that a version of individualism can be defended. Next, let us consider another challenge to individualism, provided by considerations having to do with the subject's *social* context. Whereas physical externalism of Putnam's sort holds only for the natural kind terms, social externalism is meant to hold across the board: For artifact terms such as 'sofa', for social kind terms such as 'contract', and for a variety of other non-natural kind terms ('mud', 'sand', 'air'), including most adjectives. Perhaps, then, it can be applied to the color terms as well.

4. *Color terms and the social environment*

According to Burge the considerations that support social externalism in the case of 'arthritis', apply equally to the color terms. The argument

---

[32] As often pointed out, Block then begs the case against strong representationalism since he presupposes that the difference in contents does not imply a phenomenological difference. Block's reasoning assumes that if there were a phenomenological difference, the switched subject would notice this, but the representationalist does not accept this assumption.

can get under way, Burge argues, in any case where it is intuitively possible to attribute a mental state whose content involves a notion that the subject incompletely understands. Just as we may have an incomplete understanding of the meaning of a medical term such as 'arthritis', we may have an incomplete grasp of the meaning of a color term. For instance, Burge suggests, people make mistakes about color ranges, they insist on applying 'red' not only to the standard cases (blood, fire engines, etc.) but also beyond the conventionally established range. When other speakers confidently correct them people give in and adjust their use accordingly (1979: 82). However, we can imagine a counterfactual world, Burge suggests, in which this person's use would be deemed correct by the community. If so, the meaning of 'red' would differ, and yet the subject would be internally identical to the subject in the actual world: "Holding their non-intentional phenomenal experience, physical history, and behavioral dispositions constant, we can imagine that 'red' were applied as they mistakenly apply it. In such cases, we would no longer ascribe content-clauses involving the term 'red' in oblique position". Hence, there would be a difference in content that does not supervene on internal properties (1979: 82).[33]

How compelling is this? Let us grant that when it comes to certain terms, such as 'arthritis', Burge is right and their meaning is determined not by individual use but by expert use. Does this apply in the case of color terms? One important difference, it should be clear, is that the idea of a *color expert* is problematic. Of course, some people have a more sophisticated color vocabulary than others, knowing the names of many more shades. But this is just to say that some people have a richer language than others, not that the color terms people do in fact use are such that their meaning is determined by the experts's use. Indeed, the very idea that some people are experts on how to use the term 'red' seems odd. In the case of 'arthritis' it is not implausible to think that some people (the experts) have a better grasp of the term in that they can provide a proper explication of it, spell out a correct theory about the disease. However, this reasoning is difficult to apply to a term like 'red', which is used to pick out visual properties. No doubt, there are several aspects of the meaning of the color terms beyond their use in categorizing objects. For example, to grasp the meaning of 'red' one has to know a variety of things concerning the relation among the colors, that colors are not sounds, that the perceived color of an object depends on the lighting conditions, etc. But there is nothing here that resembles a definition that someone may have a more or less grasp of and that may serve to explain the subject's deviant categorizations. Rather, when the subject deviates the only plausible explanation is that she uses the term differently because she responds to the reflec-

[33] For a similar line of reasoning see Ball 2009 and Williamson 2007: 124. Ball argues that the concept red is deferential in the sense that "the concept can be possessed even by those who have an inaccurate conception, in virtue of their relations to their linguistic communities" (2009: 951).

tance properties of these objects differently. The difference between the deviant individual and the 'expert', that is, consists entirely in their visual reactions to certain stimulus, and it is very difficult to see why the reactions of some individuals should be singled out as meaning determining for individuals who do not share those reactions.[34]

This is related to Burge's second point, about correction. It is not at all clear that speakers who are 'off-range' will in fact stand corrected. Sometimes of course we make linguistic errors involving the color terms, as when someone thinks that 'marron', in English, applies to brown. But people who are off-range in Burge's scenario are not making this type of mistake. On the contrary, they confidently apply 'red' to everything we apply it to. It is just that they take it to apply beyond the standard range; arguably, because they experience these cases as being sufficiently similar to objects within the standard range. They are therefore unlikely to stand corrected when told that their application is mistaken. Indeed, since the reason they deviated from the community in the first place was not (as in the case of 'arthritis') that they had failed to fully grasp the definition of 'red', but that they were *perceptually* deviant, it is hard to see how they *could* stand corrected. That is, even if they have the desire to conform to the community practice, they will have difficulties doing so precisely because their perceptual systems incline them to categorize objects differently.

There are therefore reasons to think that the standard cases employed by social externalists, involving medical terms such as 'arthritis' or terms for social kinds such as 'contract', are importantly different from the color terms. Terms like 'arthritis' and 'contract' do not pick out visual properties, and subjects who fail to apply them in accordance with the standard practice need not deviate perceptually. Burge may well be right about the empirical fact that people intend to use their color terms in accordance with the community practice, but this does not suffice to show that the meaning of these terms are determined by this practice. As noted above, that an individual stands corrected is in itself a neutral datum that can be accounted for by the individualist by saying that the subject adjusts the meaning of her word. And in the case of the color terms this is a more plausible description: The color terms employed by the deviant individual have a slightly different meaning and extension, even if individuals will stand corrected and adjust their use accordingly. Indeed, the reason given for why the counterfactual community uses 'red' with a different meaning, should apply equally to the deviant individual: 'Red' in the counterfactual community has a different meaning since it is applied to a distinct set of color shades.[35]

---

[34] Also, if Burge were right it would seem that someone could have our concept of red while being color blind and being unable to experience red, simply in virtue of belonging to our linguistic community. See Bain 2007: 359, for a critical discussion of this idea.

[35] It is of interest to note the connection between the type of individualism defended here and an important methodological constraint employed by psychologists who do

Now, it is often suggested that the color terms are vague, that they have indefinite borders. If so, one should expect there to be disagreements on the application of the terms in the border line cases (coupled with agreement on core cases), without this necessarily involving any meaning differences—this disagreement would simply be a reflection of the vagueness of the terms. It is possible that this is how we should describe the types of disagreements Burge considers, in which case there would be no reason to say that there is any meaning difference between the deviant individual and the other speakers. Nevertheless, when the categorization differences are significant, in particular when there is disagreement concerning the core cases, then the proper conclusion is that there is a difference in extension and meaning: You use 'red' to pick out a class of objects that is relevantly different from the class I use the term to pick out.[36]

It might be thought that this is unacceptable since it introduces meaning differences that would undermine the possibility of successful communication: If deviant individuals can be said to mean something different by their color terms, then how can these terms be used to communicate? However, it is far from clear that successful communication requires a complete identity in meaning. Arguably, what is required for the purposes of coordinating actions is sufficient similarity, not identity, allowing for by-and large overlap in how the terms are applied. In the case of the deviant individual considered by Burge, there will be such overlap since the individual is assumed to apply 'red' to the standard cases of fire engines and tomatoes. Moreover, saying that there is a meaning difference does not introduce any difficulties that are not already there. After all, if you and I disagree on the range of objects that properly fall under the extension of 'red', there will be difficulties coordinating our actions and there will be a danger of misunderstanding. The externalist 'shotgun wedding' between the meaning of the term as used by the deviant individual, and the meaning of the term in the communal language, does not do anything to remove these difficulties—it just involves a redescription of them.[37]

Concluding remarks

I have argued that the standard routes to semantic externalism cannot be employed in the case of the color terms: Color terms are neither plau-

experiments on the categorization dispositions of individuals: Relevant differences in categorization behavior indicate differences in concepts.

[36] There is empirical evidence that there are disagreements on the core cases. For instance, what some people take to be a paradigm of pure green, with no other colors mixed in, others do not. See Hurvich et al. 1968.

[37] Thus, instead of saying that there is miscommunication, Burge would say that there is successful communication, only the subject does not fully grasp the communicated content (which explains why coordination fails). It is not clear what is gained by this redescription.

sibly construed as natural kind terms, required for physical externalism
to apply, nor can they be said to involve a deferential element of the sort
required for social externalism. If so, foundational externalism does not
apply to meaning facts involving color terms, such as (iii) above:

(iii)   'Yellow' means *yellow*, for *S* at *t*.

Facts such as these do not hold in virtue of facts about *S*'s physical
environment, or facts about her social environment, but in virtue of
facts about *S*'s individual use of the term. If *S* uses 'yellow' to categorize
objects differently from others in the community, the proper conclusion
is that her term has a different meaning. This means that the standard
route towards content externalism is blocked: If foundational external-
ism does not apply to the color terms it cannot be argued, as it often is,
that because the meaning of the color terms is determined externalisti-
cally, so is the content of the color experiences.[38]

An individualist account of how the meaning of the color terms is
determined, as indicated above, is naturally combined with a disposi-
tionalist account of the colors. The purpose of this paper has not been to
defend a specific account of the nature of the colors, and I will not claim
that dispositionalism is the only theory about the colors that is compat-
ible with this type of individualism. However, *if* colors are dispositional
properties, second-order properties to cause certain experiences under
normal conditions, then which such property 'red' latches on to, should
depend on which objects an individual (in normal conditions) classifies
as 'red'. After all, how a subject categorizes objects in terms of their col-
ors (in normal conditions) is a direct reflection of how she experiences
them. Just like dispositionalist theories try to combine the subjective
aspect of the colors, their effects on us, with their objective aspects
as properties of objects, so individualism combines the subjective and
objective aspects of the color terms by appealing to how subjects are
disposed to categorize objects.[39]

[38] This is not to deny that there are other routes to content externalism, less
commonly explored in the literature. As noted in footnote 3, one option would be to
construe color terms along the lines of indexicals, combining meaning internalism
with content externalism (for such a proposal in the case of the color terms see
Spohn 1997). It should be stressed, however, that this alternative strategy requires
defending a specific, and controversial, semantics of the color terms. It is therefore
very different from the strategy discussed in this paper, which appeals to general
considerations about the role of the physical and social environment in meaning
determination, without committing to any specific semantics of these terms.

References

Armstrong, D. 1968. *A Materialist Theory of the Mind*. Humanities Press, Routledge and Kegan Paul.

Bain, D. 2007. "Color, Externalism, and Switch Cases", *The Southern Journal of Philosophy* 45: 335–362.

Ball, D. 2009. "There Are No Phenomenal Concepts", *Mind* 118: 935–962.

Block, N. 1990. "Inverted Earth", *Philosophical Perspectives* 4: 53–79, ed. J. Tomberlin.

Boghossian, P. and Velleman D. 1991. "Physicalist Theories of Color", *The Philosophical Review* 100: 67–106.

Burge, T. 1979. "Individualism and the Mental", *Midwest Studies in Philosophy* 4: 73–121.

Byrne, H. & Hilbert D.R. 2003. "Color Realism and Color Science", *Behavioral and Brain Sciences* 26: 3–64.

Cohen 2004. "Color Properties and Color Ascriptions: A Relationalist Manifesto", *The Philosophical Review* 113: 451–506.

_____2007. "Color, Variation, and the Appeal to Essences: Impasse and Resolution", *Philosophical Studies* 133: 425–438.

Donnellan, K. 1993. "There Is a Word for That Kind of Thing: An Investigation of Two Tought Experiments", in *Philosophical Perspectives, 7. Language and Logic*, ed. J.E. Tomberlin, Atascadero, Ca: Ridgeview.

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

_____2008. *The Subject's Point of View*, Oxford: Oxford University Press.

Glüer, K. 2007. "Colors without Cicrles", *Erkenntnis* 66: 107–31.

Hansen, N. 2011. "Color Adjectives and Radical Contextualism", *Linguistics and Philosophy* 34: 201–221.

Hurvich, L.M., Jameson D. & Cohen J. 1968. "The Experimental Determination of Unique Green in the Spectrum", *Perception and Psychophysics* 4: 65–68.

Jackson, F. 1998. *From Metaphysics to Ethics. A Defence of Conceptual Analysis*. Oxford: Clarendon Press.

Johnston, M. 1992. "How to speak of the colors", *Philosophical Studies* 68: 221–263.

Kripke, S. 1980. *Naming and Necessity*. Oxford: Blackwell.

Macpherson, F. 2005. "Colour Inversion Problems for Representationalism", *Philosophy and Phenomenological Research* 70: 127–152.

McKinsey, M.1987. "Apriorism in the Philosophy of Language", *Philosophical Studies* 52: 1–32.

Pagin, P. 1997. "Is Compositionality Compatible with Holism?", *Mind & Language*, 12: 11–33.

Putnam, H. 1975. "The Meaning of 'Meaning'", in *Minnesota Studies in the Philosophy of Science*, vol. 2, ed. K. Gunderson, Minneapolis: University of Minnesota Press.

_____1988. "Representation and Reality", Cambridge, MA: MIT PRess.

Spohn, W. 1997. "The Character of Color Terms: A Materialist View", in *Direct Reference, Indexicality and Propositional Attitudes*, ed. W. Künne, Stanford: CSLI, 351–379.

Stalnaker, R. 1997. "Reference and Necessity", in *A Companion to Philosophy of Language*, B. Hale & C. Wright eds. Oxford: Blackwell: 534–554.

Tye, M. 2000. *Consciousness, Color and Content.* Cambridge, MA: MIT Press.

Wikforss, Å. 2001. "Social Externalism and Conceptual Errors". *Philosophical Quarterly* 51: 217–231.

_____2008. "Semantic Externalism and Psychological Externalism", *Philosophy Compass*, 3/1, 2008: 158–181.

_____2012. "Bachelors, energy, cats and water. Putnam on kinds and kind terms", forthcoming in *Theoria*.

Williamson, T. 2007. *The Philosophy of Philosophy*. Oxford: Blackwell.