Metaphysical Realism and Thought

Sanford C. Goldberg
Department of Philosophy
1880 Campus Drive
Northwestern University
Evanston, IL 60208-2214
s-goldberg@northwestern.edu

Forthcoming in *American Philosophical Quarterly*

> It is … conceivable that human cognitive limitations prevent us from ever discovering certain of nature's joints. In such cases, I believe, we can still pick out kinds whose borders are defined by such joints. – Jesse Prinz (2002: 5)

ABSTRACT
In this paper I identify some unexplored implications of content externalism. The implications can be traced to the role that metaphysical realism plays in some of the standard arguments for content externalism. I do not regard these implications as undermining the case for externalism. I identify them, rather, as part of an attempt at honest accounting: the revisionary nature of content externalism may take us even further from some received views than many have heretofore assumed.

1.

Metaphysical realism holds that what is the case may outstrip our best (and ultimate) attempts to know what is the case. To a first approximation, we can formulate the position thus:

MR Some empirical truths are not knowable through even ideal human inquiry (performed under ideal conditions, including ideal cognitive conditions for the inquirer).

(MR) is modal claim: it asserts that some empirical truths cannot be known through even ideal human inquiry. In this paper, I will be interested in what might be regarded as a weaker version of this thesis, as follows:

> MR-P It is (metaphysically) possible that some empirical truths are not knowable through even ideal human inquiry (performed under ideal conditions, including ideal cognitive conditions for the inquirer).

In defense of (MR-P), it might be noted that nothing in the nature of the empirical world, or of the cognitive processes through which humans (aided by scientific technology) aim to acquire knowledge of that world, guarantees that all empirical truths are knowable. Many philosophers, of various ideological stripes, express an inclination to endorse some version of metaphysical realism entailing (MR-P).

Consider next a position often designated as 'content externalism.' This is a thesis regarding the individuation of the attitudes, to the effect that

> CE Some of a subject *S*'s attitude-types depend for their individuation on features of *S*'s environment – features that can be varied even as the intrinsic (non-relational) features of *S*'s body remain fixed.

(The basis for formulating a thesis regarding attitude individuation as an externalism about *content* is that attitudes are individuated in part by their content, and it is the content properties of the attitudes that depend for their individuation on external features.) Deriving from the seminal and well-known work of Hilary Putnam and Tyler Burge, (CE) (or something like it) now appears to be endorsed by a majority of those working in the philosophy of mind.

The aim of this paper is not to defend either (MR) or (CE), but to articulate several implications that obtain when we combine these two doctrines. Since the standard arguments for (CE) depend on one or another form of realism, the question then arises whether the sort of realism needed to argue for (CE) implies (MR-P). If so, the present line

of argument would establish that anyone who endorses the standard arguments for (CE) must accept these implications.

2.

I begin with the claim that standard arguments for (CE) depend on one or another form of realism. To a first approximation, we will call a position 'realist' with respect to a particular domain *D* when it holds that

R The objects in *D* exist, and have at least some of their properties, independent of anyone's intentional states, language, conceptual scheme, etc.[1]

The realism of (R) differs from that of (MR) and (MR-P). (R) is officially neutral on the completeness of subjects' cognitive abilities: for all (R) says it is impossible for there to be in-principle unknowable facts regarding the existence or properties of objects in *D*. At the same time, endorsing (R) does put pressure on one to endorse (MR) or at least (MR-P). For one thing, once one endorses (R) one eschews all sorts of views – idealism, conventionalism, positivism – that one might use to resist (MR) and (MR-P). For another, one who endorses (R) but continues to hold out against (MR) and (MR-P) faces the following question: what guarantee is there that the subject-matter itself – which (R) assumes to be (at least partially) independent of one's beliefs, language, conceptual scheme, etc. – is such that all facts in the domain can be known by the inquiring subject, in principle if not in practice?[2]

[1] This gloss borrows from the entry for 'Realism' at the Stanford Encyclopedia of Philosophy; see http://plato.stanford.edu/entries/realism/. (The entry was cited on October 3, 2007.) No doubt this gloss is crude in the extreme; but I will not waste time trying to make it precise.

[2] Of interest here is what Peacocke (1999) calls the 'Integration Challenge.' He articulates the challenge thus: "The concept of truth, as it is explicated for a given subject matter, must fit into an over-all account of knowledge in a way that makes it intelligible how we have the knowledge in that domain that we do have" (pp. 1-2). Of course, as Peacocke points out, one way of responding to the Integration Challenge in a given domain is to surrender realism regarding that domain. But if one remains committed to realism, then (MR) and (MR-P) begin to exert their pull. I do not claim that this is inevitable. I only claim that the burden of proof is on those who would resist the pull. I return to this theme in **4**.

As stated here, (R) leaves room for further clarification: what, exactly, does the notion of independence come to?[3] But it is still clear enough for us to see that something like (R) is used in the standard arguments for (CE).

Consider first Hilary Putnam's familiar (1976) argument involving Twin Earth. Putnam's thought experiment[4] was meant to illustrate the claim that life-long denizens of Earth mean something different by their use of 'water' than do life-long denizens of Twin Earth. What is more, Putnam argued that this difference in meaning was present even prior to the discovery that water is H_2O. Now prior to the discovery that water is H_2O, no one would have known that water is a different liquid from the superficially similar watery liquid on Twin Earth (= XYZ). It would thus appear that Putnam's conclusion, asserting a difference in meaning in the respective uses of 'water' on Earth and Twin Earth, is advanced on the basis of the fact that H_2O is a different liquid than XYZ. That no one at the time knew of, or even suspected, the difference in liquids is irrelevant: Putnam's view is that what people on Earth meant by 'water' is determined in part by the nature of the liquid with which they were interacting, even before anyone knew of that nature.

Nor is Putnam's argument the only externalist argument that depends on a form of realism. Consider one of Tyler Burge's (1986a, 1988) arguments for (CE).[5] Burge imagines two creatures, intrinsically alike, both of whom perceive shadow-like "entities," but where the entities in question are different kinds of thing: cracks in the case of one of the creatures, shadows in the case of the other. This difference, Burge argued, is enough to regard their

[3] Nor is this the only question. (R) clearly fails as a statement of realism regarding any domain whose objects and kinds are themselves mental or linguistic entities.

[4] Putnam's thesis in his (1976) concerns linguistic meaning, not mental content or the attitudes. However, he has subsequently agreed that the significance of his argument holds for mental content and the attitudes as well (see Putnam 1996). So I will move back-and-forth, without comment, from talk of meaning to talk of content.

[5] The argument I am citing here, of course, are not the only arguments Burge has for (CE). There remains his famous argument for an anti-individualistic version of (CE), as found in his (1979, 1982b, and 1986b), among other places.

perceptual states as differing in representational content: one creature's perceptual states represent the presence of a crack (as such), while the other creature's perceptual states represent the presence of a shadow (as such). Once again, it is the difference between cracks and shadows (and the history of the creatures' interactions with these), rather than anything that the creatures might believe (or otherwise represent in their language, or encode in their concepts, independent of how the world is), that underwrites the difference in the representational content of their respective perceptual states.

It is worth underlining that there are familiar variants on the Burge crack-shadow argument for (CE). If sound, these variants would establish that the difference in kinds (cracks; shadows) underwrites a difference in representational content even if the creatures themselves could *never* register the difference between these kinds. The variants I have in mind are those offered by proponents of the causal-informational account of (CE), as found in e.g. Dretske 1981 and Stalnaker 1993, and the teleology-based version of (CE), as found in e.g. Millikan 1984 and Papineau 1993.

Consider how proponents of these versions of (CE) might describe a familiar case involving the frog. Frogs reliably token a particular perceptual state-type when looking at a fly. But it turns out (we are told) that tokens of the same state-type can be elicited by the presence (in the frog's visual field) of BBs and other things that are smallish and blackish in the manner of flies. The question is: what do tokens of that perceptual state-type represent? Proponents of the causal-informational and teleology-based version of (CE) will agree that, despite the fact that tokens of this state-type are elicited by the presence of both flies and BBs, these tokens represent the presence of the one and not the other. To be sure, it is a difficult matter to determine the precise representational content of the frog's perceptual experience: does the experience represent *the presence of a fly*? *food nearby*? etc. What is more,

proponents of these two versions of (CE) might disagree about the proper explanation for the fact that the perceptual experience has the content it has. The proponent of the causal-informational version of (CE) might ground this ascription of representational content in the state's asymmetric dependence (in normal or ideal situations) on the presence of flies,[6] whereas the proponent of the teleology-based version of (CE) might ground the ascription in this state-type's role in indicating the presence of flies to the frog's evolutionary ancestors (a fact which, they will hold, explains the persistence of this state-type in the species). On either view, though, the frog's inability to distinguish flies from BBs (even under "ideal conditions" for frog perception) is compatible with ascribing to the frog a state representing the presence of the one, and not the other. Here, it is the fact that frogs regularly interacted with – and so have internal states that depend informationally on – flies, rather than BBs, that underwrites the claim about the content of their perceptual representations. On this line of argument, it is not even necessary that the frog be able to distinguish flies from BBs, let alone that frogs "know the fundamental nature of" the properties involved, in order to represent the presence of flies as such.

The arguments provided for (CE) by Putnam's water-twater case, Burge's shadow-crack case, and the causal-informational and teleological variants on Burge's case all share a realist assumption: the environmental kinds that determine the content of the subject's mental representations do not depend for *their* individuation on anything the subject might know (or represent) about those kinds. This is perhaps clearest in our reflections on frog cognition: that the frog's perceptual experience represents the presence of food (for example) does not depend on the frog's representing e.g. the difference between food (fly) and food-look-alike (BB). Given its cognitive limitations, it is reasonable to think that the

[6] Dretske's version of the causal-informational view appeals to what goes on in the learning period; but this detail need not detain us further.

frog will *never* come to represent this difference.[7] And this, I submit, raises an intriguing question: might we be like the frog, in that some of the kinds we represent in thought have an underlying nature we could not in principle discern (given human cognitive limitations)?

3.

Consider the following scenario. Imagine – what (MR-P) allows as a possibility – that there is a world *w*, very much like our own, in which there are properties α and β whose fundamental natures are such that, given human cognitive limitations, we will never be in a position to tell these properties apart (and in particular we are not in a position to acquire individuative knowledge of these natures). On *w* humans regularly come across α, but only extremely rarely come across β. (Maybe one human in a hundred comes across β, and only once in her life.) Humans on *w* can easily distinguish α from every other property on *w* with the exception of β. What is more, such humans have an expression in their language that they use in the presence of α, form beliefs regarding α (to the effect that e.g. it is prevalent), and so forth. To be sure, on those extremely rare occasions when a human finds herself in the presence of β, she will use the expression normally used for α. But even so, most versions of (CE) will hold that the term in question is correctly applied to α, not to β,[8] and that the beliefs in question have as part of their content a concept that correctly applies to α, not to β. (Think of this as the one-world version of the water-twater thought experiment.)

[7] Of course, if BBs were as common as flies in the frog's local environment, then a failure to discriminate the two might well diminish the chance of survival of the frogs' species. Under such conditions there will be some evolutionary pressure on any adaptive advantage whereby future frogs would be able to make the discrimination in question. This, however, does not affect my present point, as the actual world is one in which BBs are a rare presence in the frogs' local environment.

[8] It is unclear whether Davidson's versions of (CE) has this implication (see various of his essays in Davidson 2001); and it is clear that Bilgrami's (1992) version of (CE) does *not* have this implication. But it is worth noting that neither Davidson's nor Bilgrami's version of (CE) is the standard version. In any case my argument is directed against the more standard versions (above all, the versions emerging from the work of Putnam and Burge).

In what follows I want to examine the implications of holding, with (CE), that the relevant concept correctly applies to α, but not to β.

I begin with the most salient of the implications, which is this: on our present hypothesis, no human will be in a position to offer a correct and informative characterization of the application conditions of the concept in question. The substance of this claim can be brought out with the aid of an expression that is stipulated to designate all and only those things with property α. Let 'bloofer' be such an expression; let 'BLOOFER' designate the corresponding concept. Anyone who speaks a version of English extended to include 'bloofer' would be able to offer the following characterization of the application conditions of this term:

B1 'Bloofer' is true of all and only bloofers.

(B1) would be a correct characterization of the application conditions of 'bloofer' and BLOOFER, but it would be uninformative. (Intuitively: one has to "know what bloofers are" in order to understand B1, and so in order to understand what 'bloofer' is true of.[9]) Now if we press for a more informative characterization, it would have to be of the form

B2 'Bloofer' is true of all and only …

where '…' is replaced by a description of conditions that pick out all and only bloofers. If these conditions are observational, then that completion of (B2) will be incorrect, since the observational conditions used to pick out bloofers will also pick out things with β, which (by assumption) are not bloofers. In fact, it would seem that the only strategy available, if one wants a more informative characterization that succeeds in picking out all and only those things with α, is to resort to a description involving something like a theoretical place-holder, as follows:

[9] See Goldberg 2002 and 2006, where I characterize the relevant notion of informativeness at greater length.

B3 'Bloofer' is true of all and only those things of the same underlying nature as the things we (typically) pick out with …

where '…' is replaced by a description of the conditions used to pick out bloofers. Unlike the proposed completion of (B2), there might well be some completion of (B3) that is correct (or so it may be granted). But such a completion would also be uninformative: so completed, (B3) would give no theoretical information that would enable a human to effectively discriminate those things with α from those with β. Since it seems that no human could do better than some completion or other of (B3) in her attempt to offer a correct and informative characterization of the application conditions of 'bloofer', it would seem that the unknowability (by humans) of the fundamental nature of bloofers would prevent humans from offering offer a correct and informative characterization of the application conditions of the concept in question.

Arguably, proponents of (CE) have acknowledged an implication in this vicinity all along, without appealing to anything as fanciful as bloofers. They illustrate the relevant claim – that human subjects are often unable to offer correct and informative characterizations of their concepts' application conditions – with cases involving ordinary speakers and ordinary concepts like WATER, SOFA, and CUP.[10] Two externalist points are familiar here. The first is that ordinary subjects typically fail to know the individuation conditions of their concepts. The second is that ordinary subjects often have only an incomplete grasp of their own concepts. Both of these claims appear to support the contention that human subjects are often unable to offer correct and informative characterizations of their concepts' application conditions.

It is important to see how my claim above differs from these familiar externalist points. To see this, compare (i) the subject who entertains a BLOOFER-thought with (ii)

[10] I defend a version of this claim in Goldberg 2002.

the chemically ignorant subject who entertains a WATER-thought. It is a well-worn point that a chemically ignorant speaker of English can think and mean WATER-involving propositions[11] even though she is ignorant of the nature of water, and so even though (without further empirical knowledge) she cannot give a correct, informative characterization of the application conditions of the English word 'water'. Even so, she can defer to experts who are not ignorant of that nature (and so who can offer such a characterization). Of course, there are cases in which the relevant knowledge is not possessed by anyone, and so (by extension) there is no relevant expert to whom appeal can be made regarding the knowledge in question. Indeed, as noted in section **2** Putnam himself explicitly raises such a scenario in his discussion of Twin Earth. But it is worth noting that even in this case the relevant fact – the fact needed to individuate the relevant 'water'-concept (= the fact that water is H_2O) – was there to be discovered.[12] My present claim is that, given the standard arguments for (CE), there can be cases in which *even this is too much to hope for*. The cases in question are ones in which subjects employ a concept that depends for its individuation on the nature of the kind to which it applies, yet there is no humanly-discoverable fact about that nature that will enable any human at any time to play the role of the relevant expert. While I don't think that this is an implication that should make the externalist blush, it is an important one nevertheless, and one that has yet to be acknowledged.

Building on this, we can identify another implication of the standard arguments for (CE), as follows. Since BLOOFER correctly applies to things with α but not to things with β, there will be determinate empirical judgments that are such that no human would ever be in a position to rule out certain empirical challenges to the hypothesis that the actual world

[11] I am supposing propositions to be the contents of speech acts as well as the contents of thoughts.
[12] A similar point might be made in connection with cases involving deference to experts, where the experts in question disagree among themselves about the nature of the relevant kind. The hope is that these disagreements can be resolved in the long run.

verifies (or falsifies) the judgment[13] – and this, even assuming that all of the humanly-discernible facts are known.

To bring this out, I need to introduce some terminology. First, let us say that a fact *f* counts as 'humanly-discernible' just in case (i) *p* is the proposition corresponding to *f*, and (ii) there are cognitive processes possessed by humans such that, through the use of such processes, a human subject could come to know that *p*.[14] We can then extend this to characterize a notion of a humanly-discernible *state of affairs* as one which, were it to obtain, it would be a humanly-discernible fact. Next, let a scenario be any metaphysically possible state of affairs, and let us say that a judgment that *p* is verified by a scenario just in case the scenario's obtaining guarantees the truth of *p*.[15] (A judgment that *p* is falsified by a scenario when the scenario verifies ~*p*.) Now suppose that *F* is some property that can be meaningfully ascribed to bloofers; and suppose that (after all of the humanly-discernible facts are in) several apparent bloofers have been found to be non-*F*. If such a scenario is to falsify the hypothesis that

UA All bloofers are *F*

it must guarantee that

PN Some bloofers are not *F*.

Our scenario will guarantee (PN) if, but only if, the following condition holds:

---

[13] As will emerge below, the notion of verification as I am using it is a *semantic* notion, not an epistemic one: it is used to partition the sets of possible worlds so as to identify the proposition that is the content of the judgment (= the set of all possible world-scenarios that verify the judgment). Epistemic matters come in when we say that there are unanswerable empirical challenges to the claim that the actual world verifies the judgment in question. To say this is to say that no human will be in a position to know which, of the various possible worlds compatible with everything that any human can ever know, is the actual world. I bring this out below.

[14] I'd hope that there are variant characterizations that do not depend on a 'correspondence' between proposition and fact, but I will not bother trying to give one.

[15] The guarantee in question might be based on a logical entailment – as when some (canonical?) description of the scenario entails a sentence expressing the verified proposition – or it might be based on something else. We need not settle this here.

@ Some apparent bloofers that are non-*F* are *actual* bloofers that are non-*F*.

The 'if' part of this claim is obvious. For the 'only if' part, suppose (@) is false. Then none of the apparent bloofers that are non-*F* are actual bloofers. Since all bloofers have the appearance of bloofers (we stipulate that there are no 'bloofers in disguise' as part of the thought experiment), the result would be that (PN) is false. And if (PN) is false, then (UA) is true – in which case it follows trivially that the scenario fails to guarantee the falsity of (UA). In this way we see that ~(@) presents an empirical challenge to the claim that (UA) has been falsified. Of course, ruling out ~(@) requires being able to distinguish bloofers (things with α) from non-bloofer look-alikes (things with β). And by hypothesis this is something no human can do. And so we reach the conclusion that there are empirical judgments for which no human will ever be in a position to rule out certain empirical challenges to the hypothesis that the actual world verifies/falsifies the judgment in question – even assuming that all of the humanly-discernible facts are known.

One might concede this implication but wonder about its significance. To bring out this significance I highlight a corollary: no human could offer a correct, informative characterization of the truth conditions of such judgments. The argument for this corollary would be as above in connection with (B1)-(B3): any correct characterization that a human could offer will be uninformative in the manner of (B1) or (B3), and any informative characterization that a human could offer will be incorrect (for not being extensionally adequate) in the manner of (B2). Given this corollary, it would appear that if humans do have knowledge of the truth conditions of their BLOOFER-thoughts, this knowledge could only consist in relatively insubstantial knowledge like that expressed in (B1) above – knowledge to the effect that the truth of a BLOOFER-thought depends on the characteristics of bloofers. And herein lies the significance of the implication noted above.

For if our knowledge of the truth conditions of BLOOFER-thoughts consists in this relatively insubstantial sort of knowledge, then in effect we face a dilemma – one that lies at the intersection of semantics and epistemology.

Before presenting and arguing for the dilemma itself, some background is needed. It is uncontroversial that we often fail to know whether our thoughts are true. The natural account of such cases is that we know *what it would take* for the thought to be true, but fail to know whether what it would take has, in fact, obtained. What is more, there are many thoughts for which we know what it would take for the thought to be true, but we are no longer in a position where it is physically possible for us to determine whether the thought is, in fact, true. Consider the hypothesis that there were between *n* and *m* hairs on your head for your fifth birthday (for some realistic *n*, *m*). It is clear what it would take for this to be true, but given the inaccessibility of the past, the truth-value of the hypothesis may well be physically impossible to determine. But even in such cases we can frame what it would take for that thought to be true; and this enables us to determine, for any proposed specification of how the world's history might have gone, whether the thought in question would have been true, or false, on that specification of the world's history.[16] It is precisely this sort of determinability-in-possible-scenarios that is under pressure on any view that combines (CE) with (MR-P).

To bring this out, let us consider how matters stand regarding 'bloofer'-judgments from both the semantic and the epistemic point of view. Start with the semantics. Getting the semantics of 'bloofer'-judgments right requires rendering the truth conditions of such judgments as turning on the features of things with α. Suppose we introduce an expression,

[16] Here I am ignoring the complications that arise in connection with the combination of (CE) and the notion of 'truth on a specification of the world's history.' These complications are the sort that give rise to two-dimensionalist proposals (and to the critics of such proposals). I ignore these complications as not central to my concerns here.

e, for the purpose of (rigidly) designating the nature of the relevant kind. In that case, it is easy to specify possible worlds so as to get the semantics right: '*a* is a bloofer'[17] will be true at all and only those worlds *w* at which ***a*** (the denotation of '*a*' in *w*) is of kind *e*; 'All bloofers are *F*' will be true at all and only those worlds in which all things of kind *e* are *F*; 'Some bloofers are non-*F*' will be true at all and only those worlds in which some things of kind *e* are non-*F*; and so forth. For any 'bloofer'-judgment we might then effect the right partition in the space of possible worlds, and so identify the proposition that is the content of the judgment in question.

The trouble is that when we get the semantics right in this way – and there would appear to be no other way to get the semantics right – there is *in principle no way that any human could ever determine which of the various possible worlds, so described, is the actual world.* By hypothesis we are allowing possibilities involving merely apparent bloofers – that is, things with all of the properties by which humans identify bloofers, but which are not, in fact, bloofers (being things with β). So in order to know which possible world is the actual world, one would have to know, of the possible distributions of properties α and β consistent with the entirety of one's knowledge, which was the distribution in the actual world.[18] Even if we allow that one's knowledge is knowledge *of all humanly-discernible facts*, this will not enable one to know

[17] Or, if one prefers, the *judgment expressed* by '*a* is a bloofer' is … . (I will disregard this in what follows.)

[18] One might wonder about the very demanding sort of discrimination in play here: discrimination of the actual situation from *all other* possible alternatives. This is much more demanding than anything used in contemporary epistemology: e.g., the relevant alternatives theory of knowledge employs a notion of discrimination on which what is required is only discrimination from *relevant* alternatives. But it must be kept in mind that the current discussion concerns the *semantics* of 'bloofer'-judgments, not their epistemology. If one addresses the semantics of these judgments in the rubric of a possible worlds framework, then for any 'bloofer'-judgment we had better be able to partition the entirety of the set of worlds, into those in which the statement is true, and those in which it is false – and this requires the very demanding discrimination requirement I am employing. In this respect semantics is more demanding than epistemology: the sort of discrimination relevant to semantics must encompass *all* possible worlds, on pain of leaving the semantics of our expressions indeterminate (and hence inconsistent with how (CE) would have us treat the relevant semantics).

which of these distributions is the actual one. To do so there would have to be a way for humans to discriminate things with α from things with β – and this is something which, by hypothesis, no human can do. The upshot is this: in the semantics described above one would be able to tell, for each possible world so specified, whether the BLOOFER-thought in question is true at that world; but this knowledge will not determine whether the thought is actually true – and this, even assuming that all humanly-discernible facts are known.

In effect, this is the dilemma at which I hinted above. On the one hand, we can model BLOOFER-propositions (in the standard way) as sets of possible worlds. But if we do so, it is at the cost of postulating distinct possible worlds within which *in principle* no human could ever know which is the actual world. On the other hand, we can avoid this unhappy epistemic result by restricting the possibilities we countenance.[19] But if we do so then we will fail to be able to represent all possibilities, since such a restriction will leave us unable to distinguish things with α from things with β (and so we will not be able to represent the various possible distributions of these kinds). On this horn of the dilemma, we cannot represent BLOOFER-propositions in the standard way (as sets of possible worlds). In sum, it seems that we can get the semantics of BLOOFER right, but *only* at the cost of postulating a range of possible worlds within which no human could ever know which is the actual world.

The foregoing was meant to clarify the significance of the two implications I have identified for any position that combines (CE) with (MR-P). I want to round out my discussion by noting one final implication. On any view endorsing the envisaged combination, there can be cases in which the logical relations holding between determinate

[19] We might countenance only those possibilities π such that knowledge of all humanly-discernible facts would suffice to determine whether π is actual, in the sense (roughly!) that the scenario consisting of all such facts verifies or falsifies the judgment corresponding to π.

thoughts *are cognitively closed to us*. I will illustrate this with a schematic example. Let subjects *S* and *S** be doppelgängers who occupy different planets in a single universe. *S* grows up on a planet with kind *K*, where *S** grows up on a planet with kind *K**. Both *S* and *S** can discriminate the relevant kind from other kinds on her own planet; and each has thoughts purporting to pick out the kind in question. But both kinds, *K* and *K**, have underlying natures that are unknowable to creatures like *S* and *S**. Now imagine that *S* and *S** use the same word-form to pick out *K* and *K**, respectively, and that the kinds in question are exactly alike as far as the discriminatory powers of creatures like *S* and *S** could ever tell. Finally, imagine that both *S* and *S** represent the relevant kind as such – that is, *S* represents *K* as *K*, and *S** represents *K** as *K** – where *S* thinks that the kind in question is *F* (for some characteristic *F*), and *S** thinks that it is not the case that the kind in question is *F* (for the same characteristic *F*). Given the indistinguishability (to humans) of *K* and *K**, *S* and *S** would reasonably think that the kinds in question, *K* and *K**, are the same.[20] Because of this, *S* and *S** would conclude that their respective thoughts – *S*'s thought that *K* is *F*; and *S**'s thought that it is not the case that *K** is *F* – contradict one another. But they would be wrong on both accounts: *K* and *K** are not the same kinds, and *S*'s and *S**'s respective thoughts do not contradict one another. Yet these errors would forever be closed to *S* and *S**: *they could not know better.*[21] (Related points could be made about cases involving other logical relations.[22])

---

[20] One might deny this, on the grounds that in the context in which they are interacting with one another, *S* and *S** will no longer think with their original concepts (picking out *K* and *K**, respectively), but instead with some other concept, which will be common to both *S* and *S**. (See Gibbons 1996 for a version of this move, in the intrapersonal context of memory judgments.) However, I have argued elsewhere that construals of this sort will be less plausible – and in many cases, significantly less plausible – than the construal on which *S* and *S** continue to think with their original concepts (see Goldberg 2005, 2007a and 2007b).

[21] One might try to argue that their error here is not a logical one, but rather involves the falsity of an implicit premise that each endorses, to the effect that *K* = *K**. (See e.g. Burge 1998 for a discussion of a related sort of case, in an intrapersonal case involving reasoning.) But this move is deeply implausible:

However unacceptable positivist strictures are in general, there would appear to be something strange in the idea of in-principle indetectable errors *regarding the layout of the realm of thought*. To extend the familiar metaphor: it is one thing to say that *the world's* joints could be more fine-grained than anything we humans could ever discern; it is quite another to say that *how we represent the world's joints in thought* could be more fine-grained than anything we humans could ever discern. Some might think that the latter 'possibility' is of dubious intelligibility. For her part, however, the externalist must acknowledge this possibility, and so must set her teeth against the allegation of unintelligibility – perhaps explaining it away as a remnant of an unacceptably internalist conception of the mental.

4.

The foregoing argument would appear to establish that any version of (CE) that is combined with a version of realism entailing (MR-P) will have some far-reaching implications regarding the layout of the realm of thought. Although I suspect many proponents of (CE) are or would be willing to live with these implications, it is worth asking whether they *have* to live with them. I mentioned above, in **3**, that standard arguments for

---

among other reasons, it would have us ascribe implicit identity premises to most, if not all, of our arguments. It is hard to see what independent support could be offered for such a far-reaching proposal. (This is no criticism of Burge 1998, though, since he only offered this analysis as one possible analysis, and did not commit himself to using it in all cases of the sort described above.)

[22] An application of the argument from Boghossian (1994) or Goldberg (2007a and 2007b) could be used to establish this. I should note, though, that in claiming that there might be logical relations between thought-contents that are forever closed to the human mind, I am going beyond the point made in Boghossian (1992). Boghossian (1992) aimed to show that the assumption of (CE) jeopardizes "the *a priority* of logical abilities" (1992: 22). In one sense my claim is weaker than this. If logic is the study of the entailment relation, and logical abilities are abilities to compute whether such relations hold (given examples involving sentences and sets of sentences of some formal language), then I deny that (CE) has any such implication. Its implications are for *the employment* of one's logical abilities in one's attempt to discern the logical features of the thought-contents we express in natural language. (See Goldberg 2007a and 2007b.) But in another sense, my claim is stronger than Boghossian's. With respect to the logical relations between thought-contents expressed in natural language, I claim that there can be cases in which these features will be *forever unknowable* to the subject; Boghossian's claim was only that (given (CE)) such features would not be knowable *via reflection* by the subject (e.g. in cases in which she is ignorant of the relevant empirical facts).

(CE) appeal to some version or other of realism, understood as an independence thesis in the manner of (R). The question before us, then, is whether the content externalist's commitment to (R) can be combined with a denial of (MR-P), while still preserving the case for (CE). Here my claim will be that it cannot – with the result that those who endorse one of the standard arguments for (CE) *must* live with the implications recently noted.

One reason to suppose that the externalist's commitment to (R) *cannot* be combined with a rejection of (MR-P) is this: it appears unmotivated to argue for (CE) in the manner of the externalist arguments above, yet deny (MR-P). Suppose that we assume with Putnam (1976) that the concept WATER was possessed by subjects even prior to the discovery that water is H_2O. The point holds of subjects who lived in times when the techniques for exploring the empirical world did not allow them to engage in any chemical analysis whatsoever. Such subjects possessed the concept WATER in virtue of the fact that they were interacting with what in fact were samples of H_2O, discriminating such samples from other (liquid and non-liquid) kinds, and forming beliefs regarding such samples. Presumably it is their ability to distinguish the kind in question by its superficial features that enabled them to represent the kind as such in thought, even as the underlying nature of the kind was not something they would ever be in a position to know in their lifetimes (or, indeed, in the lifetimes of subsequent generations). It is clear, then, that if (CE) is motivated in this way, the externalist accepts that

> I1 A subject can think a determinate thought even under conditions in which the environmental factors that individuate that thought are inaccessible to anyone living at the time.

Now one might suppose that if a conception of thought allows (I1), the same conception of thought will allow that

> I2 a subject can think a determinate thought even under conditions in which the relevant environmental factors are *forever* inaccessible to humans.

After all (one might think), it is a contingent matter – one that reflects the contingencies of the human cognitive system – whether a particular kind has a nature that is knowable to humans. Even if one's metaphysics and epistemology has it that it *couldn't* have turned out the fundamental nature of water was humanly inaccessible, surely it is metaphysically possible, given just (R), that there are *other* kinds whose underlying natures are forever inaccessible to creatures with our cognitive endowment, which kinds we regularly interact with, and which kinds we are able to discriminate from other kinds of our acquaintance by their superficial properties. Unfortunately, (I2) entails (MR-P): for in effect (I2) asserts that there can be determinate thoughts even under conditions where the fundamental natures of the individuating (worldly) properties are humanly inaccessible; and if there can be cases where the fundamental natures of some properties are humanly inaccessible, then (MR-P), which asserts this possibility, holds.

In this way we see that the proponent of (CE) who hopes to repudiate (MR-P) must deny (I2). But for precisely this reason such a proponent is left with an awkward asymmetry. The asymmetry is this: given any kind whose underlying nature is presently unknown, (CE)'s proponent will have to hold that the kind in question can be represented as such in human thought *if but only if* the nature of the kind is humanly accessible.

The asymmetry in this position is clear; its awkwardness calls for further development. It can be seen by considering what these two types of case can have in common. Whether we are dealing with a case in which the presently unknown nature will at some time in the future be known, or with a case in which the presently-unknown nature is forever inaccessible to human knowledge, the subjects who purport to mentally represent those kinds as such might well purport to single out the relevant kind *in precisely the same sort of way*. Take the case involving a kind whose nature is presently unknown but ultimately

knowable. In this case, the subject explicitly brackets the underlying nature in question. That is, by her own lights she intends to represent that kind, *whatever it is*, of which this example presently before her is an instance (or of which standard examples with such-and-such superficial properties are typically instances). Now she does not know what that kind is (in the sense that she does not have individuating knowledge regarding the kind's nature); nor can she appeal to the existence of such knowledge among the experts in her community (for there is no such knowledge). But presumably she can still refer to the underlying nature nevertheless, in a specification in the manner of (B3) above. Indeed, we might suppose that it is this fact, together with the subject's history of discriminating the kind in question, that enables her to mentally represent the kind as such. But notice that *exactly the same thing* can be said for the subject in the case involving a kind whose nature is unknown-*because-unknowable*: she too can single out the kind by bracketing its underlying nature, can distinguish the kind from other kinds of her acquaintance by their superficial properties, and so forth. So it would seem that if these sorts of fact warrant ascribing a determinate thought *as of the relevant kind* to the one, they warrant ascribing a determinate thought *as of the relevant kind* to the other as well. But this would show that the difference between kinds whose natures are unknown but knowable, and those whose natures are simply unknowable, would not bear on the question whether one can mentally represent the kind as such. And it is precisely this difference that the 'awkward asymmetry' insists upon.

Perhaps the proponent of (CE) will offer the following reply. Insofar as our subject-matter is *human* thought, then as theorists we ought to constrain our realism to kinds whose fundamental natures are in principle accessible *to humans*. The principle is: the individuation of human thought must appeal only to features that are humanly accessible. Equipped with

such a principle, we could allow Putnam's (1976) implicit endorsement of (I1) to stand, even as we deny the stronger (I2).

But three points can be made in response.

First, it is unclear that such a response is compatible with an endorsement of (R). According to (R), what exists, and at least some of the features of what exists, exist independent of the intentional states, language, and conceptual schemes of human beings. It is not clear that such a view is compatible with a view that restricts kinds to those that are humanly accessible. The issue is vexed, however, so I will not hang very much on this initial point.

A second point to be made against the proposal to constrain our realism to humanly-accessible kinds is this: the proposal threatens to undermine some of the motivation for (CE). For (it might be wondered) once we agree that the possibilities envisaged by the industrial-strength realism of (MR-P), even if actual, are not relevant to thought-individuation, by what right do we maintain that, nevertheless, factors which are inaccessible to a whole community (at a given time) can nevertheless be relevant to the individuation of thoughts entertained by subjects at that time? To speak in vivid if somewhat obscure terms: if it is worrying to suppose that some thought-contents are individuated by features that are humanly inaccessible, isn't it *also* worrying (and in precisely the same way) to suppose that some thought-contents are individuated by features that are *practically* inaccessible to us, given the limitations of present technology? Take the folks who represented the kind *water* as such (i.e., via the concept WATER), prior to the discovery of that kind's nature. If told that they were representing a kind that is more determinate than anything they (given the concepts, techniques, and tools of the science of their day) are in a position to discover about the world, they would be uneasy. They might well wonder: how can the kinds we

represent in thought be more fine-grained than the distinctions made in contemporary science's best theories of those kinds? This sort of uneasiness does not seem to me to be fundamentally different from the uneasiness that one would have in the face of an alleged instance of (I2). For given such an instance we might well wonder: how can the kinds we represent in thought be more fine-grained than any distinctions that could *ever* be made in human science's best theories of those kinds? Of course, if the uneasiness in question is not fundamentally different in the two cases, then we are owed a justification for treating the cases asymmetrically.

There is yet a third reason for being suspicious of the move to defend the awkward asymmetry by appeal to the principle that only humanly-accessible properties should be appealed to in a theory of human thought: the principle itself would appear to be objectionable. Such a principle appears to undermine the causal-information and teleological arguments for (CE). (Insofar as Burge's crack-shadow argument was a version of one of these sorts, it would be undermined as well.) What was interesting about these arguments is that they appeared to give us a basis for ascribing a determinate representation e.g. as of a fly, to a creature who cannot differentiate flies and BBs (and so, presumably, who does not mentally represent this difference in any way).[23] Notice, though, that if this sort of argument for (CE) is correct, then the following general principle would be false: for all creatures of type *D*, the features in virtue of which the representations employed by *D*s are individuated must in principle be accessible to (and discriminable from look-alikes by) *D*s. This begs the

[23] I should make clear that Burge himself does not present his crack-shadow argument in this way. On the contrary, in his (1986) he notes that his argument is compatible with the assumption that "given [the subject] P's actual abilities and the actual law of optics, P would be capable, in ideal circumstances, of visually discriminating some instances of C's (cracks) from instances of O (relevantly similar shadows)" (1986: 42). His point is rather that in *ordinary* circumstances P cannot do so. My claim in the text is thus not about Burge's own presentation of the argument, but about a variant on that presentation – though one that will be familiar to proponents of the causal-informational and teleology-based version of (CE).

question why we should accept a version of this principle in the case of humans. At a minimum, proponents of (CE) would owe an answer.[24]

I can't pretend that this all-too-brief excursion into the options available to the proponent of (CE) are compelling: perhaps the various arguments for (CE) that employ one or another version of realism can employ a version of realism that stops short of entailing (MR-P), without undermining the motivation for (CE) itself. But the foregoing should be enough to show that doing so will not be easy. This points to what I would contend is an unpaid debt: proponents of (CE) must either establish that there are versions of realism strong enough to support the case for (CE), but not so strong as to have the implications that arise on the combination of (CE) with (MR-P); or else endorse the implications identified in section **3**, and (where necessary) explain away an intuitions to the contrary.

5.

Given (MR-P), what is the case might outstrip our methods of inquiry into such matters: the worldly realm might be (partially) unknowable, in that there are worldly truths we cannot know. The burden of the present paper has been that if (MR-P) is combined with (CE), a similar point can be made for the realm of thought: what is *thought* (or mentally represented) to be the case might be (partially) unknowable, in that there are truths (regarding how a particular subject represents the world in thought) we cannot know.

[24] Perhaps it will be said in reply that the principle, that a theory of human representational kinds ought to appeal only to features that are accessible to humans, is not a special case of the more general principle cited in the text. Perhaps it will be said that our principle here is an instance of another sort of generalization – one according to which, for *any* type of creature (whether frog, wombat, human, or what-have-you), a theory of that creature's representational kinds ought to appeal only to humanly accessible features. But in response I say that the proposed generalization is a sort of methodological species-chauvinism, and we would need to be told why we should be chauvinistic in this way. This question is all the more pressing, giving that such proponents aim to endorse some (weaker) version of realism.

While the suggestion of unknowable *worldly* truths is a familiar one, the suggestion of unknowable truths *regarding the content of one's thoughts* is not; contemplating such a possibility raises new issues that I think need to be explored further. It is worth emphasizing that the issues raised by the envisaged possibility are different from the issues raised in connection with the compatibility of (CE) and first-person authority.[25] In the debate regarding the compatibility of (CE) and first-person authority, it is *assumed* that the facts that go into determining what one is thinking – the particular content being thought – are accessible to human inquiry. Very roughly put, the question there concerns how we should make sense of one's having first-person (reflective or armchair) access to one's own thoughts, when the conditions that individuate these thoughts are only *a posteriori* knowable – they are only knowable through ordinary, empirical (third-person) methods of inquiry. This issue, of course, has been discussed at great length in the literature (see the references in the footnote above). The present issue, by contrast, is different, and in any case cuts much deeper: the matter before us concerns the very idea of determinacy in thought, when the conditions that individuate one's thoughts might not be humanly accessible *at all*.

I conclude by noting how this issue bears on one traditional assumption regarding the layout of the mental. On the traditional assumption in question, no sense can be made of the suggestion of an in-principle unknowability in the domain of thought. While one who endorses such an assumption could allow for the possibility that the *non*mental world is, *au fond*, unknowable, she holds that this *can't* be right for the domain of thought: any distinctions between thought-contents must, in principle, be discernible by at least some

[25] The seminal discussion of this sort of skepticism is Burge 1988b. See also various articles in the collections Externalism and Self-Knowledge, Martin and Ludlow, eds. (Palo Alto: CSLI Publications, 1998); Knowing Our Own Minds, Wright, Smith, and Macdonald, eds. (Oxford: Oxford University Press, 1998); New Essays on Semantic Externalism and Self-Knowledge, Nuccetelli, ed. (Cambridge: MIT Press, 2003); Meaning, Basic Self-Knowledge, and Mind. Frapolli and Romero, eds. (Palo Alto: CSLI Publications, 2003); and The Externalist Challenge. Schantz, ed. (Berlin: Walter de Gruyter, 2004).

human subject capable of entertaining those contents. The overarching conclusion of this paper is that such an assumption is foreclosed to those who endorse (CE) on the basis of the standard arguments.[26]

Work Cited

Bilgrami, A. 1992: Belief and Meaning. (Oxford: Basil Blackwell.)

Boghossian, P. 1992 : "Externalism and Inference." In. E. Villanueva, ed., *Philosophical Issues* 2: 11-28.

Boghossian, P. 1994: "The Transparency of Mental Content." In *Philosophical Perspectives* 8: *Language and Logic*, ed. J. Tomberlin. (Ridgeview: Atascadero.)

Burge, T. 1979: "Individualism and the Mental." Reprinted in The Nature of Mind, D. Rosenthal, ed. (Oxford: OUP, 1991), 536-67.

Burge, T. 1982a: "Other Bodies." In Thought and Object, A. Woodfield, ed. (Oxford: Oxford University Press), 97-120.

Burge, T. 1982b: "Two Thought Experiments Reviewed." *Notre Dame Journal of Formal Logic* 23:3, 284-93.

Burge, T. 1986a: "Individualism and Psychology." *Philosophical Review* 95, 3-45.

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

Burge, T. 1988a: "Cartesian Error and the Objectivity of Perception." In R. Grimm and D. Merrill, eds. Contents of Thought. (Tucson: University of Arizona Press).

Burge, T. 1988b: "Individualism and Self-Knowledge." *Journal of Philosophy* 85: 649 63.

Burge, T. 1989: AWherein is Language Social?@ In Reflections on Chomsky, ed. A. George. Oxford: Basil Blackwell.

[26] I would like to thank James Beebe, Kamper Floyd, David Hershenov, **Dien Ho**, Michael Horton, **Scott James**, Mike McGlone, Matthew Mullins, and Neil Williams, who have given me feedback on earlier versions of this paper; an audience at a colloquium at the University of Buffalo Philosophy Department, where I have given a version of this paper; and two referees from this journal, for helpful comments on an earlier draft.

Davidson, D. 2001: Subjective, Intersubjective, Objective. (Oxford: Oxford University Press.)

Dretske, F. 1981: Knowledge and the Flow of Information. (Cambridge: MIT Press.)

Goldberg, S. 2002: "Do anti-individualistic construals of the attitudes capture the agent's conceptions?" *Noûs* 36:4, pp. 597-621

Goldberg, S. 2005: "(Non-standard) Lessons from World-Switching Cases." *Philosophia* 32:1, pp. 95-131.

Goldberg, S. 2006: "An Anti-Individualistic Semantics for 'Empty' Natural Kind Terms." *Grazer Philosophische Studien* 70: 55-76.

Goldberg, S. 2007a: "Semantic Externalism and Epistemic Illusions," in S. Goldberg, ed. Internalism and Externalism in Semantics and Epistemology (Oxford: Oxford University Press, 2007), 235-52.

Goldberg, S. 2007b: "Anti-Individualism, Content Preservation, and Discursive Justification." *Noûs* 41:2, 178-203.

Falvey, K. and Owens, J. 1994: "Externalism, Self-Knowledge, and Skepticism." *Philosophical Review* 103: 107-137.

Fodor, J. 1983: The Modularity of Mind. (Cambridge: MIT Press.)

Millikan, R. 1984: Language, Thought, and other Biological Categories. (Cambridge: MIT Press.)

Nagel, T. 1986: The View From Nowhere. (Oxford: Oxford University Press.)

Papineau, D. 1993: Philosophical Naturalism. (Oxford: Basil Blackwell.)

Peacocke, C. 1999: Being Known. (Oxford: Oxford University Press.)

Prinz, J. 2002: Furnishing the Mind. (Cambridge: MIT Press.)

Putnam, H. 1976: "The Meaning of 'Meaning.'" In H. Putnam, Mind, Language, and Reality: Philosophical Papers, Volume 2. (Cambridge: Cambridge University Press.)

Putnam, H. 1996: "Introduction." In A. Pessin and S. Goldberg, eds., The Twin Earth Chronicles. (Armonk, NY: M.E. Sharpe).

Stalnaker, R. 1993: "Twin Earth Revisited." *Proceedings of the Aristotelian Society* 43: 297-311.