Twenty-First Excursus: Twin-Earthability and Internalism

David J. Chalmers

In chapter 7, I introduced the notion of Twin-Earthability. In this excursus I discuss a number of subtleties regarding this notion and connect it to issues about internalism and externalism.

On the definition given earlier, an expression *E* is Twin-Earthable if there can be a nondeferential utterances of *E* for which there is a possible corresponding utterance by a twin speaker with a different extension. So 'water' is Twin-Earthable since a nondeferential utterance of 'water' by Oscar (on Earth) may refer to H_2O while a corresponding nondeferential utterance by his twin Twin Oscar (on Twin Earth) may refer to XYZ. By contrast, 'zero' is not Twin-Earthable: while Burge's arguments suggest that twins in different linguistic communities might use 'zero' with different extensions, this requires that the utterances be deferential.

1. *Clarifying the definition.*

First, some clarifications and subtleties about notions used in the definition of Twin-Earthability.

(i) *Twin*: I defined twins as functional and phenomenal duplicates: beings whose cognitive systems have the same functional organization and are in the same functional states, and who have the same conscious experiences. We might also call these *fp-twins*, to disambiguate from a more standard notion of *i-twins*, or intrinsic duplicates. One might even define corresponding notions of Twin-Earthability$_{fp}$ (the notion defined above) and Twin-Earthability$_i$.

Why appeal to fp-twins rather than i-twins? One reason is that some key cases involve fp-twins that are not i-twins: cases involving speakers doubled in size, as well as the original case of Oscar and Twin Oscar. A second reason is that doing this also allows us to avoid a powerful anti-internalist argument involving i-twins that are not fp-twins, discussed shortly.

What exactly is functional duplication and functional organization? The details are not too important for present purposes. But as discussed in Chalmers 1996 (chapters 7 and 9), a functional organization is an abstract object akin to a combinatorial computational machine. It is determined by specifying a number of abstract components, a number of different possible states for each component, and a system of dependency relations whereby states of each component and outputs depend on prior states of each component and inputs. A physical system S *realizes* such an organization when it can be divided into an appropriate number of physical components with the appropriate number of states, such that the causal or counterfactual dependency relations among the components of the system, inputs, and outputs precisely reflect the dependency specified in the functional organization. Any physical system will realize many different functional organizations. We can say a *fine-grained* organization of S is an organization realized by S that is fine-grained enough to determine the behavior and behavioral dispositions of S. Two systems are functional duplicates when they share a fine-grained organization and are also currently realizing the same functional state. Then in effect the states of the system and their patterns of transitions will be isomorphic.

One obtains different notions of twinhood and therefore of Twin-Earthability depending on where one draws the boundary of a subject. For the most part I will draw this boundary at the standard place (roughly, the skin) for a human subject, although I will consider other boundaries when the issue becomes relevant.

(ii) *Extension*: As noted in chapter 7, the extension of a singular term or a kind term is its reference: the extension of 'John' (or an utterance thereof) is John, the extension of 'water' is water. The extensions of predicates and general terms are sometimes taken to be classes, but for the purposes of this definition, one should conceive of them as properties: the extension of 'yellow' is the property of being yellow (not the class of yellow things), and the extension of 'bachelor' is the property of being a bachelor (not the class of bachelors).

One consequence of defining Twin-Earthability in terms of extension is that descriptions composed entirely of non-Twin-Earthable expressions, such as 'The only conscious being', may be Twin-Earthable. Two twins, each the only conscious inhabitant of their own world, might use this description to refer to different beings, themselves. One could avoid this consequence by replacing extensions in the definition with an intuitive notion of meaning, or perhaps by secondary intensions. The twins in this case will use 'The only conscious being' with the same meaning and the same secondary intension, if a different extension.[1]

(iii) *Nondeferential*. As defined in chapter 6, an expression is used deferentially by a speaker when the referent of the speaker's use of the expression depends on how others in the linguistic community use the expression. Likewise, when an expression is used deferentially, the utterance of the expression is deferential. One can use 'deferential' in other ways, for example for certain

[1] We might call these notions extensional and intensional Twin-Earthability. Intensional Twin-Earthability is closely related to the notion of "semantic stability" developed by Bealer (19xx).

acts of deference to other speakers or for dispositions to defer, but this stipulative usage is most convenient here. The definition may allow in a few spurious cases of deference, such as 'The most common use of this expression', but these will make little difference for our purposes.

Let us say that an utterance (as opposed to an expression) is Twin-Earthable if there can be a corresponding utterance by a twin speaker with a different extension. Then we can say that an expression is *deferentially Twin-Earthable* if there can be a Twin-Earthable deferential utterance of it, and that an expression is *nondeferentially* Twin-Earthable if there can be a Twin-Earthable nondeferential utterance of it. Twin-Earthability as defined in the text is the same as nondeferential Twin-Earthability here. It is arguable that almost any expression (even 'zero') is deferentially Twin-Earthable, so the important notion for our purposes is nondeferential Twin-Earthability.

One interesting question is whether it is possible for a deferential and a nondeferential utterance to be corresponding utterances by twin speakers. I think not: deferentialness is determined by functional organization and phenomenology. But if someone thinks this is possible, perhaps for any expression, then it will be easy for any expression to be Twin-Earthable on the current definition. To exclude this possibility we can modify the definition to require that there can be a corresponding *nondeferential* utterance by a twin speaker.

I have heard it suggested that every utterance is deferential. The standard arguments for deference do not come close to establishing this claim, and I have given reasons for rejecting it in chapter 6. Still, someone who accepts this claim could use the following definition of (nondeferential) Twin-Earthability instead: an expression is Twin-Earthable if there can be an utterance of *E* that has a possible corresponding utterance by a twin speaker in a twin community with a different extension. Here, a twin community is the natural extension of the notion of a twin speaker. One could also drop "in a twin community" and say "with a different extension, not in virtue of a difference in usage in their community". Either of these definitions will serve to exclude cases in which a difference in extension arises only in virtue of a difference in usage within a community. If every utterance is deferential, the notion of non-Twin-Earthability that results will no longer go along with determination of meaning by the internal features of a speaker, but it will at least go along with determination of meaning by the internal features of a community.

(iv) *Possible*: The definition of Twin-Earthability has two modal elements: an expression *E* is Twin-Earthable if there *can* be a nondeferential utterance of *E* that has a *possible* corresponding utterance by a twin speaker with a different extension. What are the modalities here? And why not unify these: if there are two possible corresponding nondeferential utterances of *E* with different extensions?

One reason not to unify is that the unified definition may not apply to the 'water' case. It is arguable that Twin Oscar is not uttering the English word 'water' at all, but a different word that sounds the same. If so, then while we should require that one speaker (the "original" speaker) is uttering E, we should not require that the other speaker is uttering E.[2] A second reason is tied to the issue about modalities below.

Regarding modalities, one might think that both "can" and "possible" here should involve metaphysical possibility. Certainly the second occurrence should: it is not clear that Twin Earth is nomologically possible, for example, and we sometimes need to consider twin speakers in worlds with different laws. However, it is not obvious that the first should. One reason is that we may want Twin-Earthability of an expression to be tied to us or to our world.

Suppose it turns out that no humanly possible utterance of 'zero' is Twin-Earthable, but that there are alien beings whose cognitive system works in a very different, more externalist way than ours, so that their nondeferential utterances of 'zero' can be Twin-Earthable. Or suppose that the content of 'red' for humans is determined by internal states, but for beings in Eden, it is determined by the external Edenic properties that they are acquainted with. If so, understanding "possible" as "metaphysically possible" entails that 'zero' and'red' are Twin-Earthable, while understanding it as "humanly possible" allows that it is not.

Of course both notions are available, and the choice depends on our purposes. My purposes here are tied especially to human use, and to seeing whether the content of an utterance in humans is determined by internal, functional, or phenomenal factors. So for my purposes, we can understand "can" as invoking something like human possibility, while the occurrence of "metaphysically possible" invokes metaphysical possibility.

There may be polysemous or context-dependent expressions that can be uttered both Twin-Earthably (in one context or on one usage) and non-Twin-Earthably (in another context or on another usage). 'Law' might be one example, perhaps. If so, then these expressions are officially Twin-Earthable. However, for such an expression there will usually be a nearby possible expression that is constrained to be utterable only in the non-Twin-Earthable way. This nearby expression

[2] If we did require that both twins are using the same expression, Twin-Earthability would require nonconstant character in the sense of Kaplan (1989): that is, it would require that the same expression can pick out different extensions in different contexts of utterance. There are views on which 'water' and 'Hesperus' are indexical expressions with variable character, but the claim that these expressions are Twin-Earthable does not require this. The definition assumes that expressions are typed more strictly than by orthography. On the standard view, the Twin Earth expression 'water' on Twin Earth is orthographically identical to the English expression but has a different linguistic meaning; if so, then it counts for present purposes as a distinct expression.

will then be Twin-Earthable. When I say that terms such as ‘law’ are Twin-Earthable, in practice it is these more constrained expressions that I am appealing to.

(2) *Twin-Earthability and internalism.*

Twin-Earthability is closely tied to issues about internalism and externalism. Internalism about a mental property is the thesis that the property is an intrinsic property (or at least that it is an intrinsic property when instantiated by human subjects). This can be glossed as the thesis that possible i-twins never differ with respect to the property (or that human subjects and their i-twins never differ with respect to the property). Externalism about a mental property is the negation of internalism, requiring that the property be extrinsic, or that i-twins can differ with respect to it.

We can likewise say that an utterance is *narrow* if it is non-Twin-Earthable$_i$: that is, any corresponding utterance by an i-twin has the same extension. An expression is narrow if every (humanly) possible nondeferential utterance of it is narrow. Utterances and expressions that are not narrow are wide. The meaning (or at least the extension) of a narrow expression is in a sense determined by the intrinsic state of a subject using it. In this case, a sort of internalism is true of meaning. ‘Zero’ is a good candidate for a narrow expression, for example.

I will generally assume internalism about phenomenal properties; when phenomenal externalism becomes relevant, I will address it directly. Given this thesis, and given that i-twins are fp-twins, any non-Twin-Earthable$_{fp}$ expression will also be a narrow expression. So Twin-Earthability in this sense is also a reasonable guide to internalism and externalism.

Internalism defined in terms of i-twins is a *modal* internalism, concerning modal dependence on intrinsic properties. One can also define internalism as a *constitutive* internalism (Davies 1993), requiring constitutive dependence wholly on intrinsic properties, with constitutive externalism requiring constitutive dependence partly on external properties. I have heard it suggested that non-Twin-Earthability is a guide to modal internalism but not to constitutive internalism. Given that constitutive dependence entails modal dependence but not vice versa, constitutive internalism entails modal internalism but not vice versa. However, we can obtain the reverse entailment if we accept the plausible thesis that intrinsic properties are modally independent of external properties (purely extrinsic properties) and the thesis that mental properties constitutively depend wholly on a combination of internal and external properties. Then modal internalism (conjoined with the first thesis) entails that mental properties are modally independent of external properties, which entails that mental properties are constitutively independent of external properties, which (conjoined with the second thesis) entails constitutive internalism.

I have occasionally heard it suggested that every expression is Twin-Earthable and therefore wide, or at least that every expression outside logic and mathematics is Twin-Earthable and therefore wide. However, the Putnam and Burge arguments do not support these claims. The Burge arguments suggest that there are Twin Earth cases for every expression in cases of semantic deference, but these are irrelevant to Twin-Earthability in the current sense. Putnam-style arguments support Twin-Earthability for natural kind terms, names, and the like, but on the face of it they do not extend to expressions such as ‘bachelor’, ‘friend’, ‘action’, ‘conscious’, ‘cause’, ‘part’, ‘fundamental’, ‘two’, and ‘and’ (or at least, they do not extend to certain core uses of these expressions). So the thesis of universal Twin-Earthability requires support by a quite different sort of argument.

Someone might appeal to the causal theory of reference in support of universal Twin-Earthability, but there is little reason to think that the causal theory applies to every expression. The best case for the causal theory is grounded in precisely the paradigm Twin-Earthable terms discussed above (names, natural kind terms, expressions used deferentially), and for a number of the terms just mentioned, it seems to be hopeless. So in the absence of an argument that expressions such as these are Twin-Earthable, I will take it that at least some expressions are not Twin-Earthable.

At this point, Justin Fisher’s argument (2007) that internalism is false for all mental properties becomes relevant. Fisher considers species that have evolved in two radically different environments. One species (the “pulselings”) inhabit an environment irradiated by pulses of cosmic rays every 1/100 of a second, and the normal functioning of their brains depends on these pulses. The other species inhabits an environment without these rays, and their normal brain functioning requires the absence of the pulses. Fisher argues convincingly that a pulseling and a non-pulseling in these environments could have brains that are intrinsic physical duplicates for a moment while engaging in entirely different activities: one is driving down the road, the other is playing a saxophone.[3] Correspondingly, it is plausible that they will have entirely different mental and semantic lives.

Fisher’s argument can be applied to any mental property, suggesting that it is possible for a pulseling to have the property while a twin non-pulseling does not. One can argue further that we ourselves may have momentary physical duplicates in entirely different environments,

[3] By my own lights, the pulseling and the nonpulseling will be intrinsic physical duplicates but not intrinsic duplicates *simpliciter*, as I take them to have different intrinsic phenomenal properties. This move is not available to physicalists about the phenomenal, however, and I will not rest a response on my dualism. Even on my view, Fisher’s argument still bears on the nomological supervenience of phenomenal properties (suggesting that even if they are intrinsic properties, they do not nomologically supervene on intrinsic physical properties) and for the intrinsicness of nonphenomenal mental properties (suggesting that intrinsic physical and phenomenal duplicates might vary with respect to these properties).

with entirely different mental lives and mental properties. If so, this suggests a sort of global externalism: instantiating any mental property depends on the environment. For similar reasons, the extension of arbitrary terms—even ‘and’ and ‘zero’—will depend on the environment. Fisher’s argument is therefore a strong argument for global Twin-Earthability, or at least global Twin-Earthability$_i$.

In response, I think one can concede Fisher’s conclusion where intrinsic duplication is concerned, and instead stipulate a different notion of duplication for relevant purposes. For example, one might stipulate that two beings are intrinsic/dispositional twins (d-twins) when they have the same intrinsic properties and also the same local dispositions, where a local disposition is a disposition to have a certain intrinsic state when one is in another intrinsic state.[4] Local dispositions can themselves sometimes depend on the causally proximal environment, as the pulseling case illustrates. The pulseling is disposed to drive, while the non-pulseling is disposed to play the saxophone, because of the difference in the environments. To be sure, both systems will share other dispositions, such as the disposition to talk if hit by a pulse. But for many of the dispositions we are interested in, triggering conditions do not specify environmental circumstances such as whether there are pulses, and in these cases whether the disposition is present will depend on the environment.

Then even if Fisher’s subjects are i-twins, they are not d-twins. They are also not fp-twins: they have different functional organizations in virtue of their different local dispositions. So Fisher’s argument does not establish that all expressions are Twin-Earthable in the sense used in the text.

What is the relation between these notions? Two beings can be fp-twins without being i-twins or d-twins if they have microphysical differences that do not affect functional organization or phenomenology. D-twins will be functional twins, given that functional organization is a matter of dispositional relations among certain intrinsic states. For the reasons discussed in the last paragraph, i-twins need not be d-twins or functional twins. I think that phenomenal properties are intrinsic properties, so that i-twins are phenomenal twins, but those who hold that phenomenal properties are extrinsic will hold that i-twins need not be phenomenal twins. So the notion of an

[4] There are some subtle issues about just which dispositions count. The relevant dispositions will be dispositions to go into intrinsic state I_2 when one is in intrinsic state I_1. “Go into” should be understood as an immediate or short-term transition. Also, differences arising merely from the fact that the systems receive different perceptual inputs (one is about to see a red thing, the other is about to see a blue thing) intuitively should not count against their being d-twins. I will not try to make the notion of d-twin entirely precise. The move to fp-twin avoids some of these issues; any remaining indeterminacy in the notion may well be mirrored by indeterminacy in our judgments about mental states, as discussed below.

fp-twin is weaker in some respects than that of an i-twin (due to the coarse-grainedness of functional organization), and stronger in others (due to externality of local dispositions and also due to the possibility of phenomenal externalism).

Correspondingly, Twin-Earthability$_i$ neither entails nor is entailed by Twin-Earthability$_{fp}$. More generally, i-internalism, the thesis that intrinsic twins share certain mental or semantic properties, neither entails nor is entailed by fp-internalism, the thesis that fp-twins share those properties. Still, if we assume that phenomenal properties are intrinsic, then d-twins will be fp-twins. If so, then fp-internalism entails at least *dispositional* internalism (or d-internalism), which holds that the relevant mental and semantic properties are determined by intrinsic properties plus local dispositions.

In light of Fisher's argument, internalists can modify their core thesis to dispositional internalism. Dispositional internalism is compatible with some dependence of mental properties on the environment. But the role of the environment is screened off by its impact on local dispositions. So causally proximal (if sometimes spatially distant) features of the immediate environment matter: they must be closely enough connected to a subject to affect its local dispositions.[5] Mere differences in the distal historical or present environment will have no effect on local dispositions and will therefore be irrelevant to mental properties, given dispositional internalism. In view of the argument, dispositional internalism is the best sort of internalism we are going to get, and I think it is good enough for most internalist purposes.[6]

[5] Fisher argues that mental states depend on history by appealing to a principle of "mental inertia" saying that differences in current environment never suffice for a difference in mental states. I think that upon giving up internalism there is no longer much support for this principle. Still, it is an interesting question whether introducing a non-pulseling into a pulse environment (or vice versa) is enough to change (i) its mental states and (ii) its local dispositions even prior to being hit by the first pulse. If one says no to (i) and yes to (ii), then comparing the non-pulseling to a pulseling in the same environment suggests that dispositional internalism is false. My intuitions about both (i) and (ii) are unclear. Both belief attributions and disposition attributions seem compatible with yes or no answers here, depending on whether or not one gives a role to the recent past. Insofar as indeterminacy of mental states mirrors indeterminacy of dispositions, there is no problem here. Insofar as there are determinate facts about mental states here (as there may be for phenomenal states, for example), we can precisify the understanding of dispositions so that changes in dispositions reflect changes in mental states.

[6] Understanding internalism this way also helps to deal with a related problem for Humean views raised by Hawthorne 2004: because Humeanism entails externalism about dispositions and mental properties depend on dispositions, Humeanism cannot be reconciled with internalism about the mental. Fisher's argument suggests that the worry here arises equally for Humean and nonHumean views, and that the best way to retain internalism is to move to dispositional internalism.

Also relevant here are the “extended mind” arguments of Clark and Chalmers (1998), in which it is argued that beliefs (although not phenomenal states) may depend on aspects of the environment that are tightly coupled to cognitive processing in a subject’s brain. For example, the beliefs of Otto, an Alzheimers’ patient who uses his notebook as a memory, might be partly constituted by what is written in his notebook. Then Otto and Twin Otto might be intrinsic twins with different (dispositional) beliefs, because of what is written in the notebook. A dispositional internalist may suggest in response that Otto and Twin Otto are not dispositional or functional twins. One could perhaps argue that the notebook is analogous to the pulses, so that the twins have different local (body-internal) dispositions. The matter is not clear, though, as the dispositional differences here involve less immediate transitions than in the pulseling case (and they also arguably turn on a difference in perceptual inputs). Alternatively, one can argue (as Clark and Chalmers do) that the notebook is part of the cognitive system here, so that the two subjects are not fp-twins, and perhaps are not even intrinsic twins once one draws the boundaries appropriately. I will not try to settle this issue, but insofar as one holds that relevant mental states are external, one can also assume a notion of twinhood so that Otto and Twin Otto count as d-twins and fp-twins.

Henceforth, I will understand internalism as dispositional internalism. That is, internalism about a property requires that necessarily, d-twins (intrinsic/dispositional twins) do not differ with respect to that property. Narrowness will be defined in a similar way: a narrow expression is an expression that is non-Twin-Earthable$_d$, where this is defined in the obvious way. I will generally assume internalism about phenomenal properties (when phenomenal externalism becomes relevant, I will address it directly). Given this thesis, and given that d-twins are functional twins, any non-Twin-Earthable$_{fp}$ expression will also be a narrow expression. So our main notion of Twin-Earthability is a reasonable guide to internalism and externalism in this sense.