**Varieties of externalism**[1] **comments welcome**

Forthcoming in Richard Menary, ed, *The Extended Mind,* Ashgate.

Susan Hurley
July 2006

Externalism comes in varieties. While the landscape isn't tidy, I offer an organizing framework within which many of the forms it has taken (though perhaps not all) can be located. This taxonomy should be useful in itself. I'll also use it to survey and compare arguments for different kinds of externalism, while probing related intuitions.[2]

## 1. Taxonomy and preliminaries.

*1.1. What vs. how and content vs. quality.* My taxonomy consists in a two-by-two matrix: 'what' externalism contrasts with 'how' externalism, and content-related versions of each contrast with phenomenal quality-related versions. (I often say 'quality' as short for 'phenomenal quality'.)

Some forms of externalism invoke external factors to explain the 'what' of mental states, while other forms invoke external factors to explain the 'how' of mental states. The 'what'/'how' distinction isn't always sharp; how a thing works can determine what type of thing it is. For present purposes, the distinction is used as follows. 'What'-explanations explain the mental types of mental states--their personal-level content-types or phenomenal quality-types. For example, they explain why an intention is an intention to look inside the box on the left, rather than to look inside a different box, or to do something else entirely. Or why an experience is one of how something looks rather than of how it feels or sounds, or is an experience of red rather than of green. 'How'-explanations explain how the processes or mechanisms work that enable mental states of a given content or quality type (see and cf. McDowell 1994 on the 'enabling' language). They explain, e.g., what processes or mechanisms enable a given intention to look inside the box on the left, or a given visual experience of a certain surface as blue. If token mental states of the same type (content or quality) are implemented in different ways (reflecting neural plasticity), we can ask whether their 'how' explanations are fundamentally different, or also display commonality at the right level of description (Hurley and Noë 2003, and see below).

'What' externalism is also called 'taxonomic externalism' (Wilson 2004); the content externalism familiar to philosophers from traditional debates about wide vs. narrow content and twin earth, is

[1] For discussion and comments, I'm grateful to Fred Adams, Ken Aizawa, Jan Bransen, Andy Clark, Ron Chrisley, Jan Degenaar, Hans Dooremalen, Barbara de Ruijter, Mark Greenberg, Fred Keijzer, Menno Lievers, Alva Noë, Don Ross, Rob Rupert, Mark Rowlands, Richard Samuels, Nick Shea, Mike Wheeler, and an anonymous referee.

[2] Internalism claims to characterize all mental states, and externalism denies this claim must hold without itself claiming to characterize all mental states. Externalism thus has a lower burden of proof than internalism: externalism is vindicated by providing counter-examples to internalism, but internalism is not vindicated by providing counterexamples to externalism. Externalism can accommodate examples of internalist explanation with equanimity, since it denies that internalism's universal ambitions are justified without adopting comparable universal ambitions of its own. This assumption applies to all the varieties considered.

Externalism also has a lower burden of proof than internalism in a second way. Internalist explanations can appeal only to internal factors. But externalist explanations can appeal to internal as well as external factors; they are typically externalist in virtue of extending the explanans to include external factors that interact with internal factors.

one variety of it; another variety applies to phenomenal quality rather than (or as well as) intentional content (Dretske 1996; Hurley and Noë 2003).

'How' externalism is a more recent arrival; I christened it 'vehicle externalism' (1998), but here I'll also call it 'enabling externalism'.[3] 'How' explanations can be given at different or mixed levels of description, including subpersonal neural, information processing, dynamical systems, and ecological descriptions. Talk of 'vehicles' of content here does not imply that vehicles must be subpersonal representations--that only representational accounts of enabling processes are in the running; my use of 'vehicle' is neutral between representational and any non-representational accounts there may be of enabling processes. Some dynamical accounts of enabling processes may not count as representational, for example, while others do (Wheeler and Clark, 1999; Clark 1997; Wheeler 2001; Wheeler, forthcoming in Menary, ed.; Van Gelder 1999b, 1995, 1998, etc.).[4] My usage of the term 'cognitive' (as in 'cognitive processes' and 'cognitive science') is similarly neutral between representational and any non-representational accounts there may be of how mental states are enabled (see Wheeler and Clark, 1999). No assumptions about these empirical issues are implied by my terminology. Of course, the content of my arguments doesn't depend on this labelling, if someone prefers another.

In the rest of this section I sketch the landscape my taxonomy generates, to orient readers and introduce some general issues; the following sections focus on each category of externalism in turn.

*1.2. What-content externalism* is the most well-established variety: externalism about the intentional content of mental states. Just as the meaning of 'water' is determined in part by the external world, on this view the content of mental states about water is determined in part by the external world. Arguments for content externalism typically derive from intuitions that content does not supervene internally in 'twin earth' thought experiments, supplemented by various positive externalist accounts of content that explain such intuitions. These intuitions are supposedly widely shared (though I'm not aware of empirical work verifying this) and widely accepted as appropriate data for philosophical theorizing and thus as support for externalist theories of content that predict such intuitions, such as causal or teleosemantic theories.

*1.3. What-quality externalism and the magical membrane problem. What-quality externalism* is analogous to what-content externalism, but applies to the phenomenal quality of mental states. This is less widely accepted, but has proponents, especially among those who link the phenomenal qualities of experience to its intentional content. If content is partly externally determined, and content determines phenomenal quality, then phenomenal quality is partly externally determined (Dretske 1996, Harman 1990).

Note two contrasts between what-content externalism and what-quality externalism. First, widespread intuitions favour what-content externalism while what-quality externalism is often regarded as counterintuitive. The results of supervenience thought experiments for phenomenal character are controversial, but probably evoke more phenomenal-internalist than phenomenal-externalist intuitions. (Again, these are merely informed impressions about intuitions; I'm not aware of empirical work confirming them.) Rather, what-quality externalism is often regarded as a bullet to

---

[3] It's also referred to as 'active externalism' (Clark and Chalmers 1998), 'environmentalism' (Rowlands 1999), 'locational externalism' (Wilson 2004), and 'process externalism' (Keijzer, work in progress; see also and cf. his 2001).

[4] Wheeler 2001 argues that the related conditions of arbitrariness and homuncularity are needed for representational explanations, and these may not be met by neural processes where continuous reciprocal causation makes for non-trivial causal spread of enabling explanatory factors. See also Clark 1997.

be bitten, a price to be paid, in order to have the courage of one's what-content externalist convictions.

A second, meta-intuitive contrast should be distinguished from the first, intuitive contrast; it concerns the status rather than the content of intuitions. Do intuitions provide competent data to resolve the internalism/externalism issue, or merely expressions of opinion or predictions? Meta-intuition says that intuitions in supervenience thought experiments provide appropriate data to resolve what-content issues, but could simply be wrong about what-quality issues.

To elaborate: Regardless of what intuitions say about content in supervenience thought experiments, they provide competent data for theories of content. What-content meta-intuitions say that intuitions about content under various hypothetical suppositions provide the right kind of data, to which the issue between internalist and externalist what-content explanations are ultimately responsible. Content-intuitions provide the observations or evidence that what-content explanations must explain.

What-quality meta-intuitions contrast strikingly. Regardless of what intuitions say about phenomenal quality in supervenience thought experiments, it is not at all clear that such intuitions provide data competent in principle to resolve the issue between internalist and externalist what-quality explanations. Oddly, although intuitions here may be more strongly internalist, they are also meta-intuitively less competent to determine the issue. The question whether phenomenal qualities can vary with external factors when internal factors hold constant appears to be an empirical question, ultimately responsible to experience itself rather than our intuitions about thought experiments. Such intuitions merely express our opinions or predictions about what experience would be like in such cases; but we could simply be wrong. This *autonomy meta-intuition for phenomenal qualities* is an expression of the intuition that there is an intractable explanatory gap between physical or functional properties and phenomenal qualities. Autonomy meta-intuitions provide resistance to views that tie phenomenal qualities tightly to intentional contents, since qualities are ultimately autonomous in relation to data to which contents are ultimately responsible. Although autonomy and explanatory gap intuitions are often held alongside internalist intuitions, the combination is paradoxical: If someone really has no conception of how neural or internal functional properties--or indeed any others--could explain phenomenal qualities, then how can he be so confident that *if* phenomenal qualities can be explained, it must be internal factors that do the job? What is so magical about the boundary around internal factors? I discuss this 'magical membrane problem' below.

Why are intuitions favoring what-quality internalism so prevalent? Why, that is, it is so widely assumed that qualities of experience cannot vary with external factors, when internal factors are fully controlled for? I'll reflect below on two responses, appealing to hallucinations and to brains in vats. First, specific localized hallucinations and illusions can have phenomenal qualities that do not depend on external factors. It's tempting to generalize from this point, by postulating neural twins in different environments, to what-quality internalism. Second, suppose it were technically possible to transfer a brain from vivo to vitro in such a way that neural processes continue undisrupted in established patterns during and after the envatting process, relying on computer-generated inputs and feedback. It's widely assumed the transition would be phenomenally seamless: the subject/brain would experience no global or local phenomenal changes simply as a result of the envatting process, despite changes in external factors and/or intentional contents. This supposition may also seem to support what-quality externalism.

*1.4. Enabling externalism for content and for quality.* So far we have two varieties of *what*-externalism, concerned respectively to explain the content and the phenomenal quality of persons' mental states. Other varieties of externalism aim to explain *how*—by what processes or mechanisms or 'vehicles'--mental states are enabled. Enabling processes can be explained in terms of computation, neural

networks, dynamical systems, and so on. What are the boundaries of the relevant enabling processes? Can enabling processes extend beyond exclusively internal neural processes into the body and its environment? Enabling externalism (or how-externalism, or vehicle externalism) answers 'yes'.

We should distinguish externalism about processes that enable intentional content from externalism about processes that enable phenomenal quality. Arguments for *content-enabling externalism* have often proceeded under the headings of 'extended mind' or 'embodied, situated cognition'. For example, an Alzheimer patient's cognitive processes arguably extend to a notebook he uses in place of reliable neural memory processes (Clark and Chalmers 1998; see also Clark forthcoming in Acero and Rodriguez, eds); an accountant's cognitive processes may include her use of pen and paper in complex calculations. Arguments for *quality-enabling externalism* have tended to appeal to embodied, situated interactions with natural environments, often under the heading of 'sensorimotor dynamics'.

*1.5. What/how relations and the 'causal/constitutive error' error.* 'What' and 'how' explanations needn't coincide, of course. What-externalism doesn't require how-externalism; indeed, most what-content externalists are probably internalists about enabling processes. On the other hand, how-externalism may require what-externalism (see Wilson 2004, 179). More generally, externalist what- and how-explanations can overlap significantly, or constrain one another.[5] In particular, externalist what-quality explanations and externalist quality-enabling explanations tend to converge in their worldly portions, since both appeal directly to dynamic sensorimotor interactions with natural environments (e.g., Noë 2004).

The distinction between 'what' and 'how' explanations in philosophy of psychology should not be confused with a distinction between explanations of something's constitution as opposed to of its causes. The 'what'/'how' distinction doesn't align cleanly with either a causal/constitutive distinction or an internal/external distinction.

In philosophy of psychology, explanations tend to be treated as causal or constitutive with no independent justification[6], in accord with prior assumptions or intuitions about boundaries, which often themselves have no clear basis and do not illuminate the distinction. For example, prevalent externalist what-content explanations appeal to causal relations between an organism and its environment, understood to provide constitutive rather than merely causal explanation of mental content-type, even though in many illusions and hallucinations, tokens of content-types do not participate in the type of causal process that is nevertheless taken to explain, constitutively, their content. By contrast, externalist what-quality explanations and enabling explanations may be accused of committing a 'causal/constitutive error' if they regard extended explanations as constitutive rather than merely causal (Block 2005), and illusions and hallucinations are widely taken to support what-quality internalism and enabling internalism. Why are externalist explanations allowed to be constitutive in the former case but assumed to be 'merely causal' in the latter? The answer

---

[5] An interesting analogy is that between explanations of what phenotype a gene is for and how the processes work that enable a gene to express a given phenotype (see Wheeler and Clark, 1999; Wheeler 2003).

[6] Mark Johnston gives constitutive internalist 'what' explanations of the "primary objects of hallucination" (2003, 166-168) and of the phenomenal qualities that perceptions and hallucinations can share, in terms of "qualitative sensible profiles" that are instantiated in the case of perceptions but not hallucinations (133, 135, 140). He also asserts that externalist 'what' explanations for the intentional contents of perceptions (138-140) are constitutive, as against conjunctivist views that wrongly regard external causal processes as causing rather than partly constituting perceptions. However, his underlying account of the causal/constitutive distinction, which would explain why he holds that the distinction falls just where it does in these cases, is not clear. Johnston, like many discussants of arguments from illusion, does not explicitly distinguish what-quality issues from quality-enabling issues.

presumably turns on some theoretical account of content, or phenomenal quality, or their enabling processes—but this is just what is at issue between internalism and externalism. The *'causal-constitutive error' error* is the error of objecting that externalist explanations give a constitutive role to external factors that are 'merely causal' while assuming without independent argument or criteria that the causal/constitutive distinction coincides with some external/internal boundary. To avoid thus begging the question, we should not operate with prior assumptions about where to place the causal/constitutive boundary, but wait on the results of explanation. I understand externalism as in the first instance a claim about explanation rather than about metaphysics or constitution.

Some internalist critics of the 'causal/constitutive error' do provide a criterion of the mental that doesn't evidently beg the question against externalism, so don't commit the causal-constitutive error error. For example, the criterion of underived content motivates Adams and Aizawa to argue that extended processes are not constitutive but merely causal. My reply to them is different but related. Criteria of the mental or the cognitive vary widely (if not wildly) across theorists; it isn't even clear what agreed work such criteria should do. Yet psychology continues on its way with a rough and ready sense of what it wants to explain, generating good explanations. The issues between internalism and externalism should be resolved bottom up by such scientific practice, not by advance metaphysics: by seeing whether any good psychological explanations are externalist, not by deciding on a criterion of the mental and using it to sort explanations as constitutive or not. In this context, I'm aware of no appropriate criterion independent of good explanations; to the extent good explanations reveal constitution, a criterion of the constitutive cannot be used to select among good explanations. As I understand it, externalism predicts that some good psychological explanations of the 'what' or 'how' kinds will be externalist.

*1.6. The intuitive landscape.* Widespread intuitions resist the extension of externalist explanation from content to quality and from 'what' to 'how'. As we've seen, intuitions tend to favour what-content externalism, but to resist what-quality externalism. Content-enabling externalism, about the vehicles of intentional content, is also regarded as counter-intuitive (Adams and Aizawa 2001). And quality-enabling externalism, about the vehicles of phenomenal quality, seems to be the most counterintuitive of all (when it is even registered as a possibility).

| **Varieties of externalism** | *Concerning intentional content (more intuitive)* | *Concerning phenomenal quality (less intuitive)* |
|---|---|---|
| *'What' externalism (more intuitive)* | what-content externalism (most intuitive) | what-quality externalism (less intuitive) |
| *'How' externalism (less intuitive)* | content-enabling externalism, about vehicles of content (less intuitive) | quality-enabling externalism, about vehicles of phenomenal qualities (least intuitive) |

Why do intuitions about varieties of externalism differ in these ways, resisting moves from 'what' to 'how' and from content to quality? Quite different arguments have been offered for these different forms of externalism. Some of the arguments may be more plausible than others, though intuitions about the different forms of externalism may be influenced by unexamined assumptions as well as by the plausibility of arguments. Continuing to take a lofty view of the landscape, let's compare the arguments and assumptions at work.

## PART I: 'WHAT'-EXTERNALISM

### 2. What-content externalism and supervenience thought experiments.

Arguments for what-content externalism typically take the form of supervenience thought experiments (STEs) in which the supervenience of mental content on internal factors intuitively fails.[7] Internal supervenience requires that when internal factors are constant—duplicated or twinned---across some range of cases, then mental content is also constant. Hence, the supervenience counterfactual: if mental content were different across such cases, internal factors would have to differ also.

The duplication or 'twin earth' thought experiments that test supervenience requirements are so familiar that it is worth holding them up with a pair of tweezers and taking a detached look at them, to notice the assumptions that are being made. In this section I will first explain the sense in which STEs are controlled thought experiments that seek to separate out the explanatory roles of internal and external factors. Explanatory separability requires that internal factors can be unplugged from external factors. Second, I will distinguish the truth of a supervenience claim from the possibility of a corresponding STE. Internal supervenience can hold even though the relevant STE is not possible because internal factors cannot be unplugged from external. Third, the mere truth of internal supervenience provides no support for internalist explanation, if the relevant STE not possible. Internalist explanation requires explanatory separability.

[7] See Greenberg, forthcoming, on why supervenience thought experiments are an insufficient basis for externalism.

*2.1. Supervenience thought experiments as controlled thought experiments.* Let's begin with the general idea of a controlled experiment. Suppose we want to explain X. The method of controlled experimentation requires us to hold certain potentially explanatory factors constant while we vary others, in a systematic effort to separate out the factors that actually do the work of explaining X. So we divide potentially explanatory factors into two sets, A and B, and hold the B factors constant while manipulating the A factors. If we then observe that X varies if and only if the A factors vary, this suggests that the A factors are needed to explain X. On the other hand, if X holds constant with the B factors when the A factors vary, this suggests that A factors are not needed to explain X.

The method of controlled experiment seeks factors that are *explanatorily separable*. If the A and the B factors are explanatorily separable, then either the contribution made by A factors to explaining X is independent of the level of or relations among B factors, or vice versa. But if the contribution of A factors to explaining X depends on the level of or relations among B factors and the contribution of B factors to explaining X also depends on the level of or relations among A factors, then A and B factors are not explanatorily separable. In coupled dynamic systems, for example, the parameters of one system are the variables of the other system and vice versa. Or, consider the nonseparability of bodily phenotype and extended phenotype in explaining the presence of a certain genotype (Dawkins 1982). If X depends not just on the factors that are varied but also on the levels of and relations among the factors that are 'controlled', whichever way around we allocate variation and control to the A and B factors, then explanatory separability fails. Explanatory separability also fails if the A and B factors vary together in the relevant possible worlds, so that the factors in one set cannot hold constant while the others vary and thus their contributions to explaining X cannot be separated.

Perhaps we can reindividuate sets of potentially explanatory factors, so that they are explanatorily separable. But perhaps not. X may depend nonseparably on all the potentially explanatory factors and relations among them; they may be interdependent so as to form an explanatory unit.[8]

The ideas of control and explanatory separability can be extended from actual experiments to thought experiments about hypothetical cases. Supervenience thought experiments are in effect controlled thought experiments, which seek to separate out explanatory factors. STEs in philosophy of mind usually divide potentially explanatory factors into an internal set and an external set, relative to some boundary such as the skull or the skin; internal factors (neural or functional) are held constant by supposition while external factors vary. STEs thus assume that internal and external factors do not vary together in relevant possible worlds: that internal factors can be unplugged from one array of external factors and plugged into another. If internal factors are not unpluggable, but rather internal and external factors vary together across the relevant worlds, then they are not explanatorily separable.

Under the suppositions of a STE, does intentional content vary? What-content externalists answer 'yes'. This intuition provides evidence that suggests external factors are needed to explain content, though it assumes unpluggability. The explanatory role of external factors is characterized in different ways by different versions of what-content externalism. Some appeal to direct causal interactions, some to causal history, others to teleology and evolutionary function, others to expertise in the social community; different versions can apply to different content-types. Supervenience thought experiments provide the evidence such what-content explanations aim to explain.

---

[8] In complex nonlinear dynamical systems, nonseparability is common. Arguably, even though the system's behaviour might be explicable as evolving according to certain deterministic dynamical laws, nonseparability may undermine the sense in which certain factors causally explain the system's behavior, while other are merely background conditions. If so, causal explanation is arguably not the general form of explanation but a special case, just as intentional explanation is, and should not be overgeneralized.

Note that externalist what-content explanations explain the content-type of mental states; they are thus *type-explanatory* (Hurley 1998b). Not all tokens of the type need engage the external factors that explain the type in the same way. For example, an externalist what-explanation, in terms of certain normal causal interactions or normal functions, can be given of the content of a token illusory experience, even though the token itself does not participate in those type-explanatory processes (as in the well-known cracks and shadows example in Burge 1986, or in Millikan's account of how tokens can fail to perform their proper function, which determines their type). So what-content explanations of token mental states can be parasitic on absent causal processes, in which the token is not engaged.

*2.2. Supervenience on internal factors is necessary but not sufficient for internalist what-explanation.* The truth of internal supervenience claims should be distinguished from the possibility of controlled STEs. Internal supervenience merely requires that mental content does not vary *if* the relevant internal factors are duplicated across different environments: unplugged and replugged. But the truth of this conditional claim does not require that the relevantly controlled STEs are possible: it may not be possible for the internal factors to remain constant while external factors vary, to be unplugged from one environment and replugged into another. They may vary together in the relevant possible worlds, so that they are not explanatorily separable.

Internalist 'what'-explanation requires not merely the truth of internal supervenience, but that controlled STEs are possible; internal supervenience is necessary but not sufficient for internalist 'what'--explanation. Internal supervenience without unpluggability provides no support for internalist explanation; mere supervenience is compatible with the explanatory nonseparability of internal and external factors. In effect, there are two generic ways for internalism to fail: because *given* unplugging and replugging external factors are needed to explain intuitions about content, and because unplugging and replugging are not possible in the first place. What-content externalism has focussed on the first type of argument against internalism, but largely ignored the second (but see Hurley 1998a, ch. 8; Wilson 2004).

The possibility of unplugging internal factors from one environment and plugging them unchanged into another is normally taken for granted when discussing internal supervenience. It shouldn't be. Whether it's possible depends on several matters:

*A. On the range of possible worlds across which the corresponding supervenience claim operates.* The modal strength of supervenience claims varies with the range of cases at issue: they may extend through this world only, through near possible worlds, or through all possible worlds. For example, the near-worlds counterfactual reading says that, in all near worlds in which mental content differs from that in the actual world, internal factors also differ; any worlds in which mental content differs while internal factors don't are far-out worlds, such as worlds where the laws of nature differ. A stricter modal reading of supervenience applies not just to near worlds but to all worlds.

Suppose near possible worlds don't permit unplugging and replugging, though far worlds with different laws do.[9] If so, a near-worlds supervenience counterfactual is trivially true: in near worlds where internal factors are the same, external factors are also the same, hence mental content is also the same. However, the relevant STE, controlled across near worlds, is not possible; this is what could provide evidence for internalism, understood as a naturalistic explanatory claim. STEs that rely on far, merely logically possible worlds where laws differ can't provide evidence for naturalistic internalist explanations. So, from the truth of the near-worlds supervenience counterfactual it doesn't

[9] Though if the laws of nature that govern internal factors differ across worlds, internal factors are arguably not constant across the worlds; see Hurley 1998a, ch. 8, for discussion.

follow that external factors are not needed to explain content. The supervenience counterfactual could be true even though internal and external factors were not explanatorily separable.

*B. On the specific contents and environmental variations in question.* Unplugging and replugging may be possible in some specific cases but not others, with different implications for internalist explanation in those cases (for an example in which unplugging and replugging is not possible, see my discussion of El Greco worlds, 1998a, ch. 8).

*C. On whether the mental 'states' are dynamic and extend through time.* Consider the perceptual 'states' of an agent who moves body, hands, head, and eyes continually as she probes and samples her environment through multiple informational channels, generating multiple feedback loops both wide and narrow. Unplugging and replugging is less likely to be possible for such dynamic cases than for static 'snapshot' cases. Temporal extension leads to spatial extension; Dennett famously made the intracranial version of this point in his arguments against a Cartesian Theatre (1991), but the point extends promiscuously across the boundaries of skull and body.

*D. On how the boundary between internal and external is understood.* If the boundary is understood functionally, the bodily or environmental scaffolding (e.g. compensating lenses, or computers controlling brains in vats) needed to duplicate neural factors in a different environment may themselves count as functionally 'internal' factors, so that the supervenience boundary should include them. If so, factors inside the relevant boundary are not duplicated after all (see Hurley 1998a, ch. 8).

In this subsection I've distinguished the truth of a supervenience claim from the possibility of a corresponding STE, with its further separability and unpluggability assumptions. I've argued that a supervenience claim is true if the corresponding STE is possible and supports it; but it can be trivially true (or at least not false) if the corresponding STE is not possible, since there will then be no relevant case in which the antecedent of the supervenience claim holds and the consequent fails. Moreover, when internal and external factors vary together across relevant possible worlds, mental content can supervene on internal factors even though it requires explanation in terms of nonseparable internal and external factors. The truth of an internalist supervenience claim is thus necessary but not sufficient for internalist explanation; internal supervenience per se does not *provide* an explanation of intentional content in terms of internal factors[10], or even entail that there must *be* such an explanation. When internal factors are not unpluggable, external factors may be needed for explanation despite internal supervenience.

*2.3. Reflective equilibrium between boundary intuitions and explanation: The explanatory role of supervenience claims depends on boundaries that are neither too wide nor too narrow.* Supervenience claims owe much of their significance to the explanatory credentials of the boundaries they draw. They provide evidence favoring internalist explanation if and only if they express the results of a controlled STE. This requires drawing a supervenience boundary that avoids two errors, of redundancy and trivialization.

On the one hand, the boundary shouldn't be too wide. The supervenience of content on the global physical state of the world tells us nothing about what specific factors explain content. Many factors in the global state may be redundant: may do no work in explaining content. Content might be unaffected if they were allowed to vary, while holding other physical factors constant. To avoid redundancy, the supervenience boundary be narrow enough should separate out nonexplanatory factors.

---

[10] As Kim has emphasized, though for different reasons (1993); see also Greenberg, forthcoming.

On the other hand, the boundary shouldn't be too narrow. It should be wide enough to permit potentially explanatory factors on one side of it to be held constant while those on the other side vary, across relevant possible worlds. Not all boundaries do so, since some factors cannot be separated or 'unplugged' from others for explanatory purposes. An overly narrow boundary runs a danger of trivializing supervenience: the danger is that when factors outside the boundary vary, nonseparable factors within the boundary will also vary, and when factors within it are constant, nonseparable factors outside it will also hold constant. As explained, supervenience cannot be falsified if the two sets of factors cannot come apart over relevant worlds. Supervenience boundaries should be wide enough to avoid cutting across explanatorily nonseparable factors.

For example, suppose we're trying to explain color experience, and draw a boundary around a subset of the cells in the color-processing area of the brain, within which a certain process sometimes occurs: process 1. Whenever process 1 occurs in these cells, color A is experienced; whenever color A is not experienced, process 1 is not occurring in these cells. So color experience supervenes on the state of the smaller group of cells. It doesn't follow that experience of color A can be explained by the occurrence of process 1 in these cells. For suppose process 1 can only occur in these cells when process 2 occurs within a wider group of cells that includes the first. The smaller group of cells may be so intricately embedded in the dynamics of the wider group that the narrower process and what is happening outside the narrower boundary are not explanatorily separable. Holding the narrower process constant while varying what is happening in surrounding cells, or vice versa, may not be possible. We lack the control needed to support explanation of experience of color A in terms of the narrower process. The explanation should instead be sought within a wider boundary. This point would not be disarmed if experience of color A could be induced by stimulating specific cells, since such stimulation could well induce processes in other cells that contribute to explaining experience of color A.

STEs provide controls for explanations of mental types; supervenience boundaries should be adjusted as needed for these purposes. The right supervenience boundary is the one that captures the factors that explain what we're interested in, avoiding false separability and unpluggability assumptions. To avoid the dangers of redundancy and trivialization, STEs should be part of a process of seeking reflective equilibrium, in which boundaries are revised—loosened or tightened in light of explanatory progress--rather than assumed exogenously.

More generally, I see no basis independent of explanatory success for regarding factors within some pre-specified boundary as deeply or constitutively explanatory, while those outside it are explanatory only in some shallower or 'merely causal' way. I take issues about internalism and externalism to be issues about explanation. Some boundaries, like the skin, are intuitively salient. But they may not capture the explanation we seek. Intuitive boundaries can cut between factors that are not explanatorily separable.

I'm largely in sympathy with STE arguments for what-content externalism. This section isn't intended as a general challenge to such views. Rather, its aim is to induce critical awareness of the unpluggability and boundary assumptions made by STEs, and to place them into a broader explanatory context, for purposes of comparison to other arguments for externalism.

## 3. What-quality externalism and supervenience thought experiments.

In the context of what-content externalism, the previous section distinguished supervenience claims from STEs, and made claims about separability, unpluggability, boundaries, and explanation. These points apply to issues about what-quality externalism also. In particular, the supervenience of quality

on internal factors is compatible with externalist what-quality explanation. This section does not rehearse the application of these points to quality as opposed to content. Rather, it pursues further issues, concerning the two dimensions of intuitive difference between what-content externalism and what-quality externalism. Recall: 1) STEs yield internalist intuitions for quality but externalist intuitions for content, while 2) meta-intuition says that intuitions about quality in STEs have lesser standing, as evidence for internalism or externalism, than do intuitions about content.

*3.1. Supervenience thought experiments for phenomenal quality type: the magical membrane problem.* Consider a STE for phenomenal quality instead of content. Let's place the supervenience boundary around the central nervous system (CNS), so the STE postulates CNS twins in different environments. Moreover, it postulates dynamic CNS twins, not merely snapshot CNS twins: their CNS processes continue to match over time, as they interact with their different environments. The STE assumes that even so, unplugging and replugging are not problematic. Finally, the laws of nature are the same for both twins (if they weren't their CNS processes arguably wouldn't be either). Granting all this, the STE asks: could these dynamic CNS twins experience different phenomenal qualities?

I hypothesize that a widespread intuitive response would be: "The CNS twins *could* experience different qualities, as a conceptual matter, but of course they *won't*." Intuitions about such cases typically combine strongly internalist predictions, that phenomenal qualities would in fact supervene on internal factors, with autonomy meta-intuitions, that phenomenal qualities are ultimately independent of our intuitive internalist predictions. Our intuitions could simply turn out to be wrong, if qualities of experience just did differ across the CNS twins.

The *autonomy meta-intuition for phenomenal quality* says: "Phenomenal qualities are ultimately independent of intuition, in a way that intentional contents are not. If our intuitions about content in STEs were to support internalist explanation, no unexpected brute facts about mental content in CNS twins could overturn this result. But even if our intuitions about phenomenal quality in STEs do strongly support internalist explanation, unexpected brute facts about the experiences of CNS twins could in principle overturn internalism. Qualities of experience aren't responsible to intuitions in STEs the way contents are."

The autonomy meta-intuition allows that phenomenal qualities could resist explanation in terms of internal physical or functional factors, including neural processes. It's an expression of the widespread view that there's an explanatory gap between phenomenal qualities and such internal factors, since we have no idea how they could explain phenomenal qualities. Yet at the same time, widespread intuitions strongly favor what-quality internalism.

This prevalent combination, of strong internalist intuitions with the autonomy meta-intuition, is puzzling, even paradoxical. Given the strength of internalist intuitions, this combination is more than just a matter of hedging one's empirical bets. Why are intuitions favoring what-quality internalism so strong, given the autonomy meta-intuition? If we have no understanding of how phenomenal qualities *could* be explained, why is the conditional intuition so strong that *if* phenomenal qualities can be explained at all, it could only be in terms of internal factors? Why does the internal/external boundary sustain fiercely internalist intuitions about *what, if anything, must* explain phenomenal qualities despite the general admission of bafflement about *how anything could possibly* explain phenomenal qualities, including neural properties? This is what I call the *magical membrane problem.*[11]

[11] I've introduced this problem in the context of 'what' explanations, but it also arises for 'how' explanations. Intuitions strongly favor internalism about enabling processes, despite the widespread bafflement Maudlin expresses as follows: "How pulses of water in pipes might give rise to toothaches is indeed entirely incomprehensible, but no less so than how electro-chemical impulses along neurons can". (1989, 413.)

I suspect internalist intuitions gain part of their strength from aversion to the perceived alternative: dualism. The qualitative inscrutability of internal material factors, including neural processes, yields autonomy and explanatory gap intuitions. These are usually interpreted to admit the conceptual possibility of dualism, as the relevant alternative to materialism. But the scrutiny that produces autonomy and explanatory gap intuitions is usually internally focused. So we should also consider alternatives that reject internalism instead of materialism. Boundary-crossing interactionist explanations of phenomenal qualities are usually overlooked, but they provide a more promising, nondualistic response to the qualitative inscrutability of purely internal processes.

What-quality externalism avoids the mysteries of dualism without incurring the magical membrane problem of internalism. On this view, how far what-quality explanations extend is an empirical matter, case by case. In principle what explains phenomenal qualities can be distributed within the brain, among brain and body, or among brain, body, and embedding environment, depending on the explanatory dynamics. We're familiar with the idea that explanatory processes can be distributed across disparate areas within the brain instead of localized. But no magical membrane contains distributed processing; brains are in continuous causal interaction with their bodies and their environments. Why should dynamics distributed within a pre-specified boundary be capable of explaining qualities, while those beyond in are in principle ineligible? The logical basis for externally extended explanation is no different in principle from that for internally distributed explanation (Hurley 1998).

As explained in section 2, processes on either side of any given boundary can in principle vary together, whether we are thinking about internally distributed or externally extended processes. Changes in one area of the brain can induce changes in another; changes in the environment can induce changes in the brain. When local or internal factors vary with distributed or external factors in near possible worlds, it is trivially true that phenomenal quality supervenes on local or internal factors. But it doesn't follow quality can be explained solely in terms of local or internal factors.

Why is the combination of intuitions that generate the magical membrane problem so widespread, if there's a better alternative to both internalism and dualism in the form of what-quality externalism? In the rest of this section I'll consider two responses on behalf of internalism: one appeals to local illusions and hallucinations, the other to brains in vats. I'll argue that neither succeeds. What-quality externalism deserves further attention.

*3.2. Why local illusions and hallucinations don't support what-quality internalism.* The first response generalizes a claim based on local illusions and hallucinations, which seems to support internalism in a CNS twin STE. Note the nod to explanatory gap intuitions at step 2:

1. Local illusions and hallucinations can share specific phenomenal qualities with veridical experiences, despite differences in the environment.

2. Given the external differences, there must be some purely internal way to explain the sameness in specific quality of experience (even if we can't at present understand how such an explanation would go).[12]

[12] See for example Johnston 2003, who argues that 'there seems to be no obstacle to supposing that the kind of awareness involved in hallucination', individuated in terms of what he calls sensible profiles, supervenes on brain state, since the 'none of the familiar models of Externalism' seem relevant (166-167). He takes the familiar models to include Putnam's arguments for externalism concerning the thoughts of brains in vats, and Burge's arguments for social externalism. One can agree that these are not relevant, but note that Johnston assumes supervenience on brain state to be the default position, and overlooks entirely the dynamic

3. If external factors are not needed to explain specific qualities of experience, they aren't needed to explain global phenomenal state either.

4. If global phenomenal state can be explained internally, it must supervene non-trivially on internal factors, as for CNS twins in different environments.

This argument appears to support internalist intuitions in STEs despite explanatory gap worries because it fails to probe the assumptions it makes about internalist explanation. It doesn't succeed because the claim made at step 2 is false. Sameness of quality despite external differences does not require purely internal what-quality explanation. So the antecedents of steps 3 and 4 aren't justified (there are also further problems with the generalization made at step 3, discussed below).

The problem with step 2's claim is that neural correlates can differ between an illusory (or hallucinatory—I won't keep adding this) experience and a veridical experience with the same specific quality. Sameness of phenomenal quality does not ensure sameness of neural correlate. But if internal as well as external factors can differ between illusory and veridical same-quality cases, it's an open question why these different combinations of factors are associated with the same quality. In some cases, the best what-quality explanation may be externalist, contra the claim at step 2.

The possibility of variable realizations of mental states in hypothetical aliens has traditionally been used as an argument for functionalism as against 'tissue' views. But variable neural correlates of given quality types aren't just for Martians—they begin at home. As well as varying across illusory and veridical experiences of the same quality type, neural correlates can vary across instances of the same quality type before and after perceptual adaptation, and over normal development within one brain. Neural plasticity extending from childhood well into adulthood is characteristic of human brains. As a result, the neural correlates of childhood mental states can be quite different from those of adult mental states of the same phenomenal quality type. In early development, some areas of a child's brain can generate as many as 100,000 synapses a second; this early synaptic exuberance is subject to interaction-driven pruning throughout later development. Children's neural processes tend to be more distributed, within brains that contain far more synapses, and adults' to be more localized, reflecting a long process of synaptic pruning (Huttenlocher 2002, 47 and *passim*). Such neural plasticity, yielding variable neural correlates of given types of experience, is part of the normal dynamics explanatory of human experience. It isn't an exceptional process that leads to a uniquely explanatory neural endpoint (see below on the 'internal endpoint error'). And neural plasticity is disciplined and directed largely by the interactions of embodied nervous systems with their environments.

To explain quality type, we should explain why variable neural correlates are associated with the same quality type (when they are). What-quality externalism allows (while internalism denies) that such explanations can turn on the extended dynamics that embed neural processes: extended dynamics

---

embodied/embedded explanations that I take to motivate what-quality externalism. Since the latter offers an empirical as well as a philosophical explanation, it may be at cross purposes with Johnston's conception of the territory. Johnston takes an example of a 'seamless transition' from a local hallucination to verdical perception (122) as his 'stalking horse' in explaining what is right and what is wrong about traditional arguments from illusion. In his own positive account the common explanatory factor in hallucination and verdical sensing (144) is 'at the level of experience (123), rather than a brain state per se, in contrast with the conjunctive view he describes (115-116) and rejects. Nevertheless, Johnston takes quality types to supervene on brain state. His reasons for doing so, as above, appear weaker than his arguments for his own positive account and are dissociable from it, if mere supervenience on internal factors is distinguished, as I have urged, from internalist explanation.

can have a characteristic underlying pattern that explains quality type, the neural components of which can be implemented or parameterised in different ways. In such cases, characteristic patterns of interaction between embodied nervous systems and their environments can explain what experience is like, not just the internal neural portion of such interactions.

I'll elaborate the argument from variable neural correlates against step 2 of the above internalist argument, in two steps. First, I'll give an example of variable neural correlates across illusory and veridical experiences of the same quality-type. Second, I'll explain how variable neural correlates can figure in an extended dynamical what-quality explanation of veridical cases, which I claim have explanatory priority.

*3.2.1. Variable neural correlates: example.* For an example of how the neural correlate of a local illusion can differ from that of a veridical experience of the same quality type, we can compare an illusion of environmental movement with a perception of environmental movement. But to do this, we need first to consider an experience that does not involve environmental movement. When you move your eyes sideways, motor signals are associated with resulting changes in actual visual inputs and with pre-created or simulated feedback in a certain dynamical pattern. This pattern correlates with your experience as of objects in your environment not moving, though your eyes have moved. By contrast, if your eye muscles are paralysed so they don't move sideways when you try to move them, and objects in the environment don't move either, you nevertheless have an illusory experience as of the environment moving sideways. On a standard view, the neural correlate of this illusion includes motor signals and simulated feedback similar to those correlated with the no-movement experience, but lacks the changes in actual visual input correlated with the no-movement experience.

So far we've got a veridical no-movement experience with one neural correlate, and an illusory experience of movement with another neural correlate. Now let's consider a veridical experience of movement, qualitatively the same as the illusory experience of movement. This veridical experience might be had during a sideways earthquake, during which you neither move nor attempt to move your eyes (apologies to J. J. Gibson!). In one case there is an earthquake, and in the other there is not; yet your specific experience of sideways movement is qualitatively the same. But the neural correlates of the experience are not the same in the illusory movement and earthquake cases. In the illusory movement case the neural correlate includes motor signals and simulated feedback relating to attempted eye movements, which are not part of the neural correlate in the earthquake case; in the earthquake case the neural correlate includes actual visual input signals that are not part of the neural correlate in the illusory case. It's tempting to explain quality in the illusory movement case in terms of its neural correlate alone, since no actual sideways movement occurs. But this would not explain why the same quality type is present in both the illusory movement and the earthquake cases, despite their varying neural correlates.

The internalist argument from illusion we're considering compares illusory and veridical cases, failing to recognize that neural correlates can vary despite sameness of quality. To explain this, we should focus in the first instance on veridical cases, and try to explain why neural correlates of a given quality can vary *across veridical cases.* Only then will we be in a position to explain sameness of quality despite variable neural correlates when we compare veridical and illusory cases. The argument from variable neural correlates for what-quality externalism gives *explanatory priority to veridical cases* in which neural correlates of a given quality can vary.

*3.2.2. Variable neural correlates within extended explanatory dynamics.* So let's now consider how neural correlates can vary across veridical cases, despite sameness of quality. We get examples of this across normal development, given normal neural plasticity, and when illusions induced by distorting lenses adapt away over time, as the agent interacts with her environment. For example, each lens of

Kohler's goggles were yellow to one side of the midline and blue to the other; they distort wavelengths reaching the eye as a function of eye movement or object movement across the midline, produces illusions about the colors of objects. But after a period of wearing Kohler's goggles, color experience adapts, so that veridical experience replaces illusory experience. However, the neural correlates of veridical experience of a given color presumably differ before wearing the goggles and after adaptation, as a function of, among other things, different wavelengths reaching the eye. So a given type of (veridical) experience can have different neural correlates before and after adaptation.[13]

What is the best explanation of sameness of experience type when neural correlates vary across veridical experiences? Does the explanans cross internal/external boundaries in some cases? This is an empirical question, to be settled by explanatory success, case by case. Moreover, while some illusions adapt away, deferring to reality, others do not. Why? To explain *what experience is like* we must explain why some experiences defer to *what the world is like*, while others do not (see Hurley and Noë 2003 on the dominance/deference distinction). What-quality externalism holds that, in principle, the needed what-explanatory factors can cross the internal/external boundary; it can do so in some cases, but not in others.

In general terms, explanations of mental types in terms of extended dynamic patterns go as follows. As an agent interacts with her environment, information flows from environment through nervous system along multiple sensory and motor channels and out into body, as embodied activity changes the environment and/or information flowing from the environment into the nervous system, along multiple channels of sensorimotor feedback. A complex multidimensional space results, which evolves through time in characteristic patterns. The nervous system also pre-creates or simulates feedback, adding further loops to these complex dynamic patterns. Such complex patterns carry information about both agent and environment; some of their dimensions are purely neural, while others extend beyond the neural. Feedback loops are what can rope in external factors; loopiness ('turbo drive', as Clark puts it in work in progress) is crucial to dynamic extension. Degree of extension is governed not by ultimate causal sources of organismic inputs but by the orbit of feedback loops whereby organismic outputs produce organismic inputs via external factors. Nervous system and embedding environment are informationally coupled, via the body, as each affects the other. The parameters of a system express the way its variables interact; in coupled systems, variables of each system act as parameters of the other. Not only can variables in human nervous systems change environmental parameters, but environmental variables can also reparameterize the plastic nervous system.

Relevant dynamic patterns are often multimodal as well as extended. Experience of a given type often depends not simply on a single channel of interaction between external and internal factors that can easily be interrupted or 'faked', but rather on relationships among multiple dimensions that develop in characteristic ways over time, as various modalities of sensory input and simulated feedback are followed by motor output with consequences that bounce off various features of the environment and generate multimodal feedback.[14] Such complex sensorimotor dynamics triangulate flexibly on

---

[13] Or indeed even simply after adaptation: consider Pappert, who wore left-right reversing goggles only half the time, until he could ride a bicycle while taking his goggles on and off, and experienced no visual reversal when doing so; a building on the right looked to be on the right to him, both with goggles on and with goggles off. For discussion and references, see Hurley 1998a, ch. 8, 9.

[14] The interactions of an active agent with her environment generate what I've called a *dynamic singularity* (1998a)*:* a tangle of causal and informational feedback loops centered on herself that moves with her and ropes in her brain, body, and elements of her environment. Dynamic singularities are extended in the same sense that phenotypes can be extended (Dawkins 1982); the skin is transparent to the dynamic feedback processes whose character explains what phenotype, or what type of experience, is in question.

environmental features. Single sensory channels can contribute to explaining experience type in the context of such extended patterns without being explanatorily separable from such context.

When a single sensory modality is distorted (e.g. by goggles), characteristic dynamic multimodal patterns are disturbed, and illusions result; multiple types of experience can be affected. Conflicts between experiential modalities can be resolved by veridical adaptation of one modality (e.g. vision) or nonveridical adaptation of another (e.g. proprioception); when illusory experienced distinctions adapt away, other illusory distinctions can arise (see Hurley 1998, ch. 9). Re-establishing coherence, between modalities and between experienced distinctions and constancies, can involve the partial reimplementation of extended dynamic patterns at new neural locations or with new parameterisations of intraneural aspects of the patterns. As an agent reacquires perceptual skills and experience of color constancy while wearing Kohler's goggles, the underlying sensorimotor pattern characteristic of certain colors is reimplemented, re-parameterized to reflect eye movements (see Gibson's account in Kohler 1964; Hurley and Noë, in press).

What-quality explanation is externalist if the dynamic pattern explanatory of an experience-type has boundary-crossing dimensions of embodiment and environmental embedding. Adaptive recovery of the same quality of experience can in some cases be best explained by the re-emergence of its characteristic extended dynamic, re-implemented in some internal dimensions by a variant neural process. For example, adaptation to Kohler's goggles restores color constancy as objects or eyes move across the midline. The neural correlates of experiencing an object moving across the mid-line as white differ, before wearing the goggles versus while wearing them after adaptation. Why are both neural correlates of the whiteness quality of experience? Because both participate in a certain extended dynamic characteristic of color, which reflects among many other things the fact that external objects do not change color systematically as they move, or as eyes move, across the midline.[15] The embodied adapting agent needn't explicitly represent this extended dynamic pattern or the reimplementation of its neural portion, but his embodied perceptual skills are part of what sustain it. With time, as a seamless result of neural plasticity and the agent's re-acquisition of such skills through interactions with his environment, the extended pattern characteristic of a certain experience type may re-emerge, relocated in certain dimensions of its multidimensional space so as to compensate for the imposed distortion. An underlying higher-order dynamic pattern can obtain across changes in neural implementation, as adaptation realigns what experience is like with what the world is like (see and cf. McDowell 1994, de Gaynesford 2004, on the openness of experience to the world).

This account of extended multidimensional dynamics reveals further problems with the internalist generalization argument above: with step 3's generalization from local to global internalist explanation, leading to step 4's postulation of CNS twins in STE. The best explanation of some quality types may be internalist, while others are best explained by extended multidimensional dynamics. Thus, it may be possible for some neural correlates to hold constant across different environments such that quality-types supervene on neural correlates, but in other cases it this may not be possible. In some cases internal factors may not be unpluggable and repluggable across near worlds, so that internal and external factors are not explanatorily separable. So we generalization from local internalist explanation in some cases to global internalist explanation is not warranted.

The argument from variable neural correlates in veridical cases does not assume that extended dynamical patterns, as opposed to purely internal dynamical patterns, must provide the best explanation. Even if neural correlates vary, they may both implement one purely internal functional pattern, rather than an extended functional pattern, which explains sameness of quality. The

[15] See O'Regan etc for more on the dynamics of color experience.

argument only assumes that extended patterns can provide explanations of quality type when neural correlates vary; it's an empirical question whether they do in any case.

However, it might be objected that there will always be an internal 'shadow', at a functional level of description, of any extended dynamic pattern, and this will always provide a better, internalist functionalist explanation of quality type even if neural correlates vary. In reply, we can suppose for the sake of argument that *some* internal functional shadow can be found to correlate with any extended dynamical pattern, even when neural correlates vary. But it doesn't follow that this internal functional pattern will have any independent explanatory role, let alone provide a better explanation. It may be seriously disjunctive. In the absence of the extended explanation, it may be one nonsalient functional pattern among many, with no nonarbitrary significance. Given the extended explanation, it may be a *mere* shadow, projected in the light of the extended dynamic that does the real explanatory work. Again, it is an empirical question, case by case, whether an internal functional pattern or an extended dynamic provides a better explanation.

Note that the answer to this question does not turn on the truth of internal supervenience, which is necessary but not sufficient for internalist explanation. An extended dynamic can provide a better explanation, because internal and external factors are not explanatorily separable, even though an internal supervenience claim is true.[16] The issue is one of explanation, rather than a prior metaphysical issue.

Note that the plausibility of externalist explanation depends on allowing that some qualities of experience may be best explained dynamically, rather than as a series of snapshots strung together. Internalist intuitions too often turn on snapshot assumptions. It is the dynamic character of experience that makes active, embodied CNS twins problematic and that knits internal and external factors together. As Dennett (1991) has argued, temporal and spatial extension go hand in hand.

The first internalist response to the magical membrane problem fails. It tries to support internalist intuitions in STEs by arguing from illusion. But we cannot get from specific illusions to what-quality internalism for global phenomenal state, as the internalist generalization argument tries to do. Nor does what-quality externalism depend on violations of internal supervenience. What-quality internalism applies too narrow a boundary to would-be twins in STEs, one that cuts across potentially explanatory extended dynamics in cases involving variable neural correlates resulting from environmentally driven adaptation and/or neural plasticity. Extended what-quality explanations shouldn't be excluded a priori. If there's no magical membrane, then it's an empirical question, case by case, whether they succeed. I predict they will for some qualities, in particular where neural correlates vary, and not others (see Hurley and Noë 2003 on deference vs. dominance). To explain quality type where neural correlates differ, we should give priority to comparison of veridical cases, and address illusory cases in the light of our understanding of veridical cases.

*3.3. Why brains in vats don't support what-quality internalism.* The internalist may be tempted to appeal next to virtual reality devices. But active, embodied agents can probe, manipulate, remove, smash, or walk out of such devices. Embodied action creates extended dynamic patterns that triangulate on qualities of the environment flexibly and reliably, sensitive to very small differences; it can outwit virtual reality as well as eliminate many illusions. The only action-proof virtual reality is a duplicate reality.

---

[16] For a somewhat different argument, see also the discussion of El Greco cases in Hurley,1998, ch. 8; and see Wilson 2004 on the inefficient redundancy of internalising the extended aspects of some processes. See also Noë 2004, and O'Regan and Noë's work on 'change blindness' phenomena, and the way active visual sampling of an environment by means of eye movements determines the contents and quality of visual experience.

A second, more radical internalist response to the magical membrane problem removes the very embodiment that mediates such extended dynamics: brains in vats seem to be internalism's ultimate weapon. But, I'll argue, brains in vats don't secure what-quality internalism in STEs either.

In my argument from variable neural correlates, embodied dynamic interactions make trouble for the internalist argument from illusion to internalist explanation of the experiences of CNS twins in STEs. Perhaps the trouble can be avoided by arguing for internalism from disembodied CNS twins—twin brains in vats. Unlike embodied brains, envatted brains are helpless. They can be unplugged from one environment and replugged in another as freely as technology allows. They can't manipulate or smash their vats or walk out of them to eliminate the phenomenal qualities the vat conjures up. They can't probe and sample their environments to induce illusions to adapt away. Nor can they wear distorting goggles to induce neural re-parameterizations that contribute to explaining quality type only as part of an extended dynamic. Duplicate neural processes in vats could in principle be sustained over time, despite being located in different environments—say, by means of computers that provides each brain with multimodal simulations of external input and of feedback in response to motor signals, and which cancel out any further influences their different environments might otherwise have on the brains. Of course, the envatted twin brains wouldn't actually be generating any movement or feedback from movement, but motor signals could be fed through their respective computers and generate sensory feedback along multiple channels, to create a simulation of the extended dynamic pattern resulting from active multimodal triangulation on external factors. But this simulated pattern needn't bear any relationship to the different further environments of each vat/computer pair.

If brains in vats have the last word, what do they tell us? Suppose internal supervenience holds; the duplicate brains 'experience' the same quality. (The scare quotes indicate noncommitment to the thought that brains, as opposed to animals or people, have experiences; having registered that point, I'll drop the scare quotes.) Would their duplicated neural processes explain the shared experience type?

Not necessarily. A version of the argument from variable neural correlates applies again. Brains in vats are highly nonstandard in being disembodied. But if they are normal brains, they should still display neural plasticity. Their computers could thus simulate the external feedback loops of wearing distorting goggles, inducing twin neural re-parameterizations in both twin-brains. Each twin brain could thus have variable neural correlates of that quality type. What explains the sameness of quality type despite varying neural correlates within each brain?

For present purposes, this question is no different from the question why the same quality can have different neural correlates for one interactive agent who sports an embodied, situated brain. The twinning and envatting are idle in answering the question about variable neural correlates within one brain, so don't support internalism. In the embodied case, I appealed to possibility that a complex environment-involving dynamic with variable neural implementations could explain quality type. In the disembodied case, computers simulate such a dynamic, independently of the environment beyond the computers; but by doing so they provide the external part of an extended dynamic. Here, an extended explanation of quality type despite variable neural correlates would be in terms each brain's interactions with its computer rather than with its further environment. But such an explanation would nevertheless appeal to something beyond the brains themselves—to the extended dynamics provided by their computer-environments. Again, it's an empirical question whether an extended dynamic provides the best explanation of quality type given variable neural correlates.

Conceding that twin brains in vats must share quality types thus doesn't support what-quality internalism. Neural supervenience is not the touchstone of what-quality internalism.

*3.4. Leaks in the magical membrane.* The magical membrane problem arises from combining strongly internalist intuitions with the autonomy meta-intuition for phenomenal qualities. If we're genuinely modest about our understanding of how quality type could be explained, we should remain open-minded about what-quality externalism, and consider externalist explanations on their empirical merits, case by case. We shouldn't assume that whatever could explain quality type must be located within a boundary that cuts between neural and external factors. What-quality internalism is not the only alternative to dualism. Neural processes are normally in continuous dynamical interaction with external factors; there's nothing magical about the boundary between them. In some cases it may be explanatorily transparent, so that internal and external factors make nonseparable contributions to explaining quality type. The qualities of the world we interact with may be part of what explains the qualities of our experience. Some of our bafflement about how to explain phenomenal quality may derive from boundary presuppositions that attempt to separate explanatorily inseparable factors and focus our scrutiny inward, when what is needed is a wider gaze, one that takes in extended dynamics with bodily and environmental *as well as* neural dimensions.

Here it may be objected that a 'causal/constitutive error' is being committed: that external factors are merely causally, not constitutively, related to quality type. If this objection helps itself to an unargued assumption that a causal/constitutive distinction coincides with an external/internal distinction, then it makes the 'causal/constitutive error' error. What non-question-begging criterion of constitutive explanation justifies this assumption? If extended multidimensional dynamic patterns provide the best explanation of quality type in some cases, why assume that external dimensions are merely causal while internal are constitutive? Moreover, it isn't clear how causal/constitutive talk can be mapped onto complex dynamical explanation, or even what work a criterion of the constitutive is supposed to do in this context. We don't have such a criterion here, and it isn't clear that the cognitive sciences need one in order to find good explanations. We should proceed by seeking good explanations of qualities of experience case by case, then noticing whether any are externalist, rather than by trying to apply a prior criterion of the constitutive to select among potential explanations.

The overall shape of my argument about what-quality externalism in section 3 has been this. Two internalist responses to the magical membrane problem were considered, both of which attempt to support internalist intuitions in STEs: one on the basis of illusions, the other on the basis of brains in vats. I've argued that variable neural correlates of given qualities make trouble for both responses: what explains why different neural correlates are collected by the same quality type? In some cases, extended dynamics in which internal and external factors are not explanatorily separable can provide plausible answers. Externalism competes case by case with internalism to provide a better explanation. There's no shortcut to internalism via the claim that only internalist explanations are constitutive.

## PART II: 'HOW'-EXTERNALISM

### 4. Content-enabling externalism.

I turn now from 'what'- to 'how'-explanations, or enabling explanations. Externalism about how mental states are enabled has been referred to as vehicle externalism; I'll speak here of *enabling externalism*. This section focuses on content-enabling externalism, and the next on quality-enabling externalism.

Most discussions of the 'extended mind' concern extended cognition--externalism about vehicles of intentional contents. They consider enabling explanations that cross internal/external boundaries,

including body, environmental objects, or both. Arguments for boundary-crossing vehicles of contents tend to be of two overlapping types, appealing to agents' dynamic interactions with cultural artefacts or tools in particular, or with natural environments more generally.

*4.1. Cultural extension: artefacts plus parity.* Cultural arguments for extended cognition invoke artefacts that extend the powers of the mind, often involving language, plus a principle of parity. Parity says that the location *per se* of a process doesn't determine whether it counts as part of how the mind works. If processes relying on silicon chips, or notebooks, do enabling work relevantly similar to work done by neural assemblies or synaptic settings—so that they would count as mental processes if they were in the head--, then they can count as vehicles of mental contents regardless of location.[17] If Otto's notebook "plays the right sort of role in driving cognitive processes"—i.e. does work similar to internal memory in reliably enabling him to go to the museum--, then, by parity, it's part of his extended mind, part of how it works (Clark and Chalmers 1998, 12). Continual interaction with artefacts isn't required for extended cognition; it could be enough for Otto automatically to check his notebook at critical points. But some cases of extended cognition do rely on continual dynamic interaction with artefacts, as when a skilled accountant performs complex calculations, her pencil flying across her notebook page, her eyes sampling just the pencil marks needed at each point of the calculation process. Accessible information doesn't need to be copied internally to be exploited in cognition (see Wilson 2004 on exploitative representation and wide computation).

*4.1.1. Cultural extension: objections and replies.* Cultural extension arguments face various objections. Slippery-slope objections urge that extended minds leak into the world uncontrolledly, with absurd consequences. In response, constraints are imposed on culturally extended cognition: artefacts should play their role fluently and automatically, and be available as and when needed (Clark 2005, 3).

Cultural extension arguments and the parity principle may seem to be in tension with another strand of 'how'-externalism, which holds that details of embodiment can be essential to how minds work. The tension isn't deep. Bodily details do contribute to how minds work: the distance between eyes and ears, the range of possible eye and head movements, the left-right symmetry and back-front asymmetry of the body, and so on (see Lakoff and Johnson 1999; Noë 2004). But it doesn't follow that embodied minds cannot be culturally extended. Rather, artifactual extensions of minds are informed and constrained by bodily mind-enabling mechanisms; mind-extensions cannot be body-neutral. Tactile visual substitution systems, e.g., are not body-neutral--though they capture only some aspects of normal vision's embodiment and the cognition they enable is correspondingly limited (see Hurley and Noë 2003). We can recognize the importance of embodiment in enabling minds without relocating the magical membrane accordingly: without assuming that only what's within a boundary around natural bodies could enable mental states. Distortions or deficits at skin-level can sometimes be compensated for by external artefacts, restoring an extended content-enabling pattern of brain/body/world interactions.

A prominent critique of cultural extension arguments objects that (1) cognitive states must have 'intrinsic' content and (2) it's empirically implausible that cognitive science will find extended states with intrinsic content (Adams and Aizawa 2001). Neither claim should be accepted.

As Clark argues, the idea of intrinsic content is not very clear (2005, 4). Intrinsic contents supposedly do not ultimately derive from other intrinsic contents. Social practices, conventions, and language

---

[17] Parity is named by Clark and Chalmers 1998; the same principle is independently invoked on behalf of vehicle externalism in Hurley 1998 e.g. 190-193, 325. More recently discussion in extended mind circles has shifted from the parity to the complementarity of internal and external processes; but this issue cuts across my purposes here, so I don't pursue it.

may be regarded as having nonintrinsic contents that derive from intrinsic mental contents. Various accounts of intrinsic content appeal to causal, historical, functional, or other relations, excluding social relations that presuppose intentional mental content. But content is no more intrinsic to brains in virtue of their relations to non-social environments than their relations to social environments. Underivedness is not the same as intrinsicness. We do better (as Adams and Aizawa now do) to focus on the underivedness rather than the intrinsicness of content.

But there are still problems with treating underived content as the mark of the cognitive. Consider artificially evolved robots; do they have only derived contents? If the artifice of evolvers deprives evolved robots of underived content, would a divine creator's intentions also deprive his creatures of underived content, hence of genuine cognition? Moreover, consider the way language transforms and enhances a child's cognitive capacities. Even if language builds on prior mental contents, so that linguistic contents are derived, further mental contents also build on language, so that their content is presumably also derived. Yet we do not therefore regard all such linguistically derived contents as not genuinely cognitive. Finally, the relations of derivation between mental contents and the content of language and other social practices are not clear; someone of a Vygotskian persuasion, for example, might argue that mental contents derive from linguistic contents and social interactions through a process of internalization (see Menary, forthcoming).

An alternative to the view that underived content is the mark of the cognitive is a view motivated by developments in dynamical cognitive science. On this view, the mark of cognitive processes is that, as well as being available on-line, in direct interaction with the environment, some version of a cognitive process is also available off-line, in simulative mode (Clark 1997, 465; Clark and Grush 1999, 12-13; Hurley 2006 on forward models and other simulations). Adams and Aizawa suggest that extended mind advocates largely ignore what's known about the brain and cognitive processes, and cast them as neo-behaviourist (2001, 47). But we should be wary of the dated dichotomy between classical computational and behaviourist conceptions of cognition. As Van Gelder comments, dynamics is arguably the single most widely used and powerful explanatory framework in all of science; we shouldn't be surprised to find it explaining cognition (1998, sect. 5).

What are the implications of this dynamically motivated 'availability off-line' criterion for extended mind hypotheses? This is a further question, about which there is disagreement (see Grush 2003; cf. section 5 below). But Adams and Aizawa (2001) display no recognition that extended mind views are motivated by boundary-crossing in contemporary dynamicist cognitive science, according to which what enables cognition is not bare brains but actively embodied and situated brains (Van Gelder 1999a, b; 1998; 1995, Clark 1997). This thriving body of work raises empirical and theoretical issues about whether cognition-enabling processes must be purely neural, excluding relations to social and natural environments--or indeed must even be representational. [18]

*4.2. Extended dynamics and cognition: A-not-B; acallosal integration.* Cultural examples of extended minds should be located within a broader dynamicist approach to cognition in terms of the dynamic coupling of brains, bodies, and environments. On this view, content-enabling processes can extend beyond the brain in the absence of cultural artefacts, although the coupling of brains via bodies to cultural artefacts can extend cognitive processes in further ways, distinctive of human cognitive. Without disputing the importance of cultural extension, I suggest that mind-extension arguments that appeal to dynamic coupling with natural environments in general are more fundamental than those

[18] Recall: I assume that processes that explain how minds work can be *cognitive*, whether or not they all turn out to be *representational*. Any non-representational dynamical processes that explain how minds work are not thereby disqualified from counting as cognitive. The point isn't how the label 'cognitive' should be used, but that it's an open question whether non-representational dynamical processes can explain how minds work.

that appeal to cultural artefacts in particular (Keijzer and Schouten, in press, make a similar claim; thanks for Fred Keijzer for discussion on this point).

Dynamical cognitive science has been well surveyed and referenced by those cited above; I won't repeat the job here. The general framework is one of a multidimensional space of possible states, developing over time, often in complex, nonlinear, and surprising ways. Variables in different dimensions can be interdependent, each changing in ways that can depend on values of and relations among other variables. From each point in multidimensional space, a trajectory develops over time in accord with system parameters, which can themselves change over time. The dynamical system can be expressed as a characteristically structured geometry of possible trajectories through this space, which may converge on certain attractors or avoid certain repellors in the space, or display other distinctive patterns of flow. Abrupt changes in flow structure can emerge from continuous changes in variables or parameters. In coupled dynamical systems, the variables of one system are the parameters of the other, and vice versa; they can be viewed as one system. The boundaries of dynamical systems are not exogenous to explanatory aims. In cognitive applications, the state space can extend to include dimensions whose variables are bodily and environmental as well as neural, as brain, body and environment interact in mutually shaping patterns. However, there's no ban on purely internal cognitive dynamics, in cases where it provides the best enabling explanation. In dynamical cognitive science it's debated whether some geometrical features of flow structure should be viewed as representations, or whether dynamic cognitive science can dispense with representations. Cognitive processes, however, are construed as features of the temporal evolution of a multidimensional space, not as static structures.

An example of dynamical cognitive science that contrasts nicely with traditional approaches is Thelen and Smith's account of the much-studied A-not-B error, made by infants of 7 to 12 months of age (Thelan and Smith 1994, ch. 10; van Gelder 1999a). A child faces two bins, bin A and bin B. If you hide an attractive toy in bin A, the child will reach for bin A. If you continue to hide it in bin A, he will continue to reach for bin A. If you then hide it in bin B instead, and responding is delayed a few seconds, the child will still reach for bin A; but he'll reach for bin B if responding is not delayed. Why? Various traditional approaches explain this error in terms of limitations in the child's conception of objects, representation of space, or memory. But they don't explain certain context effects in the experimental data. The error depends on length of delay in a way that changes with age, and the presence of more bins reduces the tendency to make the error. Thelen and Smith's dynamical model explains these wrinkles in terms of ongoing interactions between a 'what' system, for seeing toy, bin, and/or hand, and two 'where' systems, for looking and for reaching. Changing inclinations to reach in a direction at a time depend on position of the system in various interacting dimensions, including the direction of current reaching inclinations, general and specific features of the environment (such as number of bins present and their markings, and which bin the toy is currently hidden in), and memory-based habit. They find parameters for a complex equation which, when computationally simulated, produces the A-not-B error, including the subtle variations traditional approaches don't explain. Moreover, their model predicts further results that have subsequently been confirmed experimentally. Bodily and environmental features play essential roles in this dynamical account of how early cognition works. Such dynamical models motivate an extended view of cognitive processes, without relying on cultural artefacts to do the extending (assuming the bins could equally well be natural containers).

Another example of embodied, dynamically extended cognition that doesn't rely on cultural coupling is my hypothetical acallosal subject with extended mechanisms of integration via bodily movements (Hurley 1998, 2003). Although not set in a formal dynamical systems framework, it may be more intuitive. Information normally passes between the brain's two hemispheres via the corpus callosum. In commissurotomy patients this is surgically severed; as a result, information in their two

hemispheres is not integrated under various experimental conditions. In acallosal patients, the corpus callosum is congenitally absent; yet they show unified cognition under experimental conditions in which in commissurotomy patients do not (Jeeves 1965; Milner and Jeeves 1979; Diamond 1972, pp. 61-66). What enables the integration of information in acallosals? In principle, integration could be enabled by wholly internal processes, partly external processes, or both. It's likely that internal processes, relying on ipsilateral or subcortical neural paths, are at least partly responsible for integration. But, by the parity principle, partly external processes could also enable integrated cognition; these could rely on bodily movements that distribute or transfer information across the hemispheres. Access movements—automatic, habitual side-to-side movements of head or body--could give each hemisphere direct sensory inputs from an object that would otherwise appear in only one hemisphere's visual field. Cross-cuing by automatic facial expressions accessible to both sides could also function to transfer information across hemispheres (Bogen 1990).

The experimental tests of integration that commissurotomy patients fail and acallosal subjects pass are designed to exclude access movements and cross-cuing. That's why it's likely that acallosals actually have internal, neural mechanisms of integration (unless some extended mechanisms of integration are so subtle and automatic that they evade experimental control). Nevertheless, in ordinary uncontrolled circumstances, access movements and cross-cuing could also contribute to integrating information, along with secondary neural pathways; acallosals might only rely exclusively on the latter when deprived of the former. This could be an efficient, robust developmental solution to enabling acallosal integration. Marcel Kinsbourne remarks that absence of the corpus callosum is biologically trivial, since minor adjustments in orientation distribute the same information to both sides (1974); there's some evidence of motor habits in acallosal subjects that could serve this purpose (see Hurley 2003 for further discussion). Such extended mechanisms of integration would depend on bodily activity and feedback rather than purely neural factors. If they functioned when needed, reliably and automatically, by parity they would illustrate extended cognition.

*4.3. Diagnosing intuitions: explanatory relations between on-line and off-line processes in enabling cognition.* Why is content-enabling, vehicle externalism is less intuitive than familiar philosophical what-content externalism? Adams and Aizawa, e.g., regard the former as a "wild idea", at odds with common sense. Magical membrane assumptions may influence some intuitions, despite the causal congress of brains with bodies and environments (though Adams and Aizawa disavow such assumptions). Moreover, unfamiliarity with boundary-crossing dynamical cognitive science, as opposed to the traditional in-the-head computational variety, may wrongly make extended cognition seem empirically implausible.

But I want to consider a further possibility. The attention given to cultural cases of extended cognition, important and distinctively human as they are, may distract intuitions from a more basic point about the dynamics of extended minds, which doesn't depend on cultural artefacts. The more basic point concerns the explanatory relations between on-line processes and off-line simulations. In cultural extension cases, these relations are complicated by the way on-line processes involve external representations. We can make the more basic point salient by separating it from issues about relations between external and internal representations in cultural extension cases. In the rest of this section, I'll explain the distracting issues raised by external representations in the cultural cases. In the next section, I'll focus on noncultural cases and the more basic issue about explanatory relations between on-line processes and off-line simulations.

*4.3.1. Explanatory relations between on-line processes and off-line simulations: cultural versus non-cultural cases of extension.* Consider the distinction between on-line and off-line processes in cultural extension cases. The relevant cultural artefacts are themselves external representations, or work in ways that depend on external representations. External representations stand in for something else, which may not be

present for direct interaction. Recall the skilled accountant's fingers and pencil flying over the pages of her notebook; her eyes move to access just the information she needs just when she needs it. Such extended computation is a process of on-line sensorimotor interaction with an external medium of information storage and external symbols, pencil marks on paper. It involves direct interaction with symbols already at one remove from the items they stand for--such as bank balances and tax owed--, not direct interaction with these worldly referents.

However, the same work might be done by taking these on-line interactions with symbols off-line, using internal computations that simulate finger movements and symbol perception and relying on memory instead of pencil and paper to hold information for further use. Via internal simulation, an analogue of the extended process involving pencil and paper is available off-line. Note that such off-line simulations of interactions with symbols are at a second remove from the items the symbols stand for: they don't rely on direct interaction with external symbols, any more than with the items the symbols stand for. It's been proposed that such availability off-line is a mark of cognitive processes (Clark 1997, 465; Clark and Grush 1999, 12-13). It doesn't follow that only the off-line processes are cognitive, of course; the view is rather than on-line processes are themselves cognitive in virtue of availability off-line.

Issues now arise about relations between processes involving external representations and internal simulations thereof. Arguably, off-line simulations of interactions with external representations lack explanatory independence from the on-line interactions appealed to in cultural extension cases. For example, the off-line capacity for mental arithmetic arguably derives in normal development from long on-line practice with pencil and paper, so that the on-line version is explanatorily prior to the off-line version in an important ontogenetic sense (see Clark, in progress, for related discussion). More generally, the capacity for much off-line thought arguably continues to depend on on-line public language to maintain simulations. On the other hand, cultural extension may seem to enable cognition only because it presupposes symbols that can function to represent what's not present. And if this capacity is enabled by contentful internal processes, then the extended, on-line aspect of cultural cases is a detour (cf. Adams' and Aizawa's concerns about the derivativeness of content). Cultural extension cases seem to make extended cognition hostage to these issues about whether external representation derives from internal representation, and thus not to provide independent leverage for content-enabling externalism.

However, this set of issues about relations between internal and external representations distracts attention from a more basic underlying issue. The tangent develops because of the way cultural extension cases involve direction direct interactions with external representations but not with what they are about--since external representations are already at one remove from the items they're about, even before they're taken off-line. Even if external representations do *not* derive content from independent internal contents, nevertheless internal off-line simulations of interactions with external representations will inherit independence of the world represented from external representations. For example, even if interactions with external representations of bank balances enable thinking about bank balances, it doesn't follow that interactions with bank balances themselves enable thinking about bank balances. The more basic issue I want to separate out concerns relations between on-line interactions with the world—not external representations of it—and off-line simulations of such interactions. Can interactions with trees enable experiences of trees, or do only internal simulations of interactions with trees enable experiences of about trees?

To address this more basic issue, we should bracket issues about external representation and hence cultural extension cases. What's needed is a focus on explanatory relations between on-line interactions that *don't* involve external representations and internal simulations of such interactions. That is, we should focus on sensorimotor interactions among brain, body, and natural environment,

where the relevant on-line processes don't presuppose external symbols already at one remove from what they are about. The more basic issue concerns whether, in a Brooksian phrase, the world can be its own best representation—and in particular whether what the world is like can be part of what enables us to experience what it is like.

I gave two non-cultural examples of extended cognition, concerning the A-not-B error, and acallosal integration. One involves cognition in infants, the other pathology. The intuitiveness of mind-extension would be better served by cases involving normal adults (see Keijzer and Schouten, in press, on change-blindness). The adaptability and neural plasticity found in normal adults provided examples for what-quality explanations in terms of extended sensorimotor dynamics, in section 3 above. I'll return to such cases to bring into sharper relief the issue whether off-line processes are explanatorily independent of on-line processes, by separating it from the complications raised by cultural extension and external representation. I'll argue in the next section that on-line extended sensorimotor dynamics can provide quality-enabling explanations.

5. Phenomenal-quality-enabling (vehicle) externalism. So I now turn from content-enabling externalism to that most unintuitive and radical form of externalism, phenomenal quality-enabling (vehicle) externalism. Surprisingly, it's in this unpromising territory where extended mind intuitions can be run to ground.

*5.1. Preliminaries: what-quality externalism vs. quality-enabling externalism, and the middle-ground.* I'll shortly address explanatory relations between on-line interactions with the world and off-line simulations thereof, and consider whether the former can provide quality-enabling explanations. But first it will be helpful to make some preliminary points about what-quality explanations and quality-enabling explanations. It might be thought that even if my argument from variable neural correlates for what-quality externalism succeeds, the further, even more unintuitive step to quality-enabling externalism should be resisted. After all, most what-content externalists are content-enabling internalists. Why not similarly combine what-quality externalism with quality-enabling internalism? That would be a middle ground position that concedes some ground to radical externalism, contrary to initial intuitions, but is not *so* radically unintuitive as quality-enabling externalism. Why depart this middle ground to countenance extended vehicles of phenomenal qualities? To answer, we need to compare the roles extended dynamics would have in what-quality explanations and in quality-enabling explanations.

What-quality externalism appeals to characteristic dynamic sensorimotor patterns in explaining the qualities of experiences: of visual versus auditory experience, or specific qualities within one modality. It holds that in some cases qualities of experiences can best be explained in terms of extended dynamics in which brain, body and world all participate, while in other cases the best explanations may be in purely internal. In particular, I argued, what-quality explanation may need to appeal to extended dynamics to explain qualities with variable neural correlates. Qualities of experience adapt to follow characteristic extended patterns when their neural portions are re-implemented as a result of, say, normal developmental neural plasticity, or wearing distorting lenses such as those in Kohler's goggles (see Hurley and Noë 2003 for other examples, e.g. TVSS, or the projection of tactile inputs to visual cortex in blind persons). In such cases the quality of experience can defer to extended sensorimotor dynamics despite variable neural correlates. Since they are not counter-examples to neural supervenience, they underscore that externalist what-quality explanation is compatible with neural supervenience. What-quality externalism holds that what predicts and explains phenomenal quality in some such cases are extended dynamics, rather than the properties of a particular reimplementation of the neural portion of the dynamics, or an internal functional 'shadow' of extended dynamics. That is, when the neural portions of an extended dynamic are

reimplemented over development or in response to distorting lenses, what collects the various neural implementations together under a given quality is the extended dynamic in which they participate.

The middle ground view concedes that *what* qualities we experience can require externalist explanation, in light of variable neural correlates, but insists that vehicles of phenomenal qualities—the enabling processes that explain *how* we are able to experience given qualities-- are internal neural processes. For example, a middle ground view could concede that extended sensorimotor dynamics can explain what quality we experience, and that skills in negotiating such extended dynamics and associated expectancies of the sensory consequences of movement can enable our experiencing of qualities—while still insisting that enabling explanations in terms of skills and expectancies should be understood in terms of internal simulations of such extended dynamics. Can what-quality externalism be held apart from quality-enabling externalism in this way in this middle-ground way?

Note that this question is *not* analogous to a question about cultural extension cases, about whether what-content externalism can be held apart from a content-enabling externalism that invokes interactions with external representations. A middle ground view might be supported in cultural extension cases by the purported derivativeness of the content of external representations. But noncultural extension doesn't involve external representations; so noncultural cases remove at least this basis for occupying the middle ground. The disanalogy follows from the way external representations in cultural extension cases are already at one remove from the world represented, discussed in the last section. As a result, the relations of extended dynamics in what- versus how-explanations differ across cultural versus non-cultural extension cases.

CONTRAST:
RELATIONS OF EXTENDED DYNAMICS IN WHAT vs. HOW EXPLANATIONS FOR CULTURAL vs. NON-CULTURAL EXTENSION

| | **Externalist What-Explanations** | **Externalist How-(enabling) Explanations** | |
|---|---|---|---|
| **Cultural Extension Cases** | Direct interactions with world represented<br>→ DYNAMICS DO NOT CONVERGE IN WORLD | On-line:<br>Direct interactions with external representations, at 1 remove from world represented<br>← | Off-line:<br>Simulations of interactions with external representations, at 2 removes from world represented |
| **Non-Cultural/ Natural Extension Cases** | Direct interactions with world & its qualities<br>→ DYNAMICS CONVERGE IN WORLD | On-line:<br>Direct interactions with world & its qualities<br>← | Off-line:<br>Simulations of direct interactions with world, at 1 remove from world |

In *cultural extension* cases, externalist what-content explanations and extended content-enabling explanations do not converge in the world represented: what-content explanations typically appeal to direct interactions with the world represented, while how-explanations appeal to direct interactions not with the world but with external representations of the world (e.g. notebooks). And off-line

internal simulations of interactions with external representations—*cultural simulations*-- are at two removes from what is represented, presupposing stable external representations of the world.

By contrast, in non-cultural or *natural extension* cases, extended what- and –how explanations would indeed converge in the natural world. Neither would appeal to external representations of the world, but rather to extended sensorimotor dynamics, patterns of interaction with the natural world and its qualities. Off-line internal simulations of such dynamics—*natural simulations*-- are thus only at one remove from the natural world, and presuppose direct interactions with the world to be simulated.

This contrast predicts, e.g., that cultural simulations would show greater stability in, say, an isolation tank than would natural simulations. Moreover, we should expect externalist what-explanations and extended how- explanations to constrain one another more directly in natural cases than in cultural cases: in natural cases, it should be harder to keep enabling explanations from leaking into the world along with what-explanations, so harder to occupy the middle ground position.

*5.2. Why go radical? Explanatory relations between on-line extended dynamics and off-line simulations.* Return to the question: Haven't variable neural correlates and extended dynamics done all the externalist work they can do in arguing for the middle ground position? Why go further, to radical quality-enabling externalism? My argument concerns explanatory relations between extended on-line processes and internal off-line simulations thereof. The extended on-line processes I have in mind aren't interactions with cultural artefacts or external representations, but are more basic: direct sensorimotor couplings with a natural environment, converging with the extended dynamics that feature in externalist what-quality explanations. The corresponding off-line processes are what I called 'natural simulations': internal simulations of direct couplings with the natural world, rather than with cultural items that represent the world.

What are the explanatory relations between extended sensorimotor dynamics and simulations thereof? And how does the answer bear on whether extended dynamics as opposed to simulations thereof can provide quality-enabling explanations?

Here's a story about the explanatory relations between extended sensorimotor dynamics and simulations thereof. Consider subpersonal neural expectancies or predictive simulations of sensory feedback from movements. Such 'forward models' associated with efference copy are in effect internal feedback loops that mimic external feedback loops. Consider three contexts in which such simulations could play enabling roles. (Keep in mind that all these enabling roles are described at a subpersonal level.)

*(1) On-line simulations in comparator control systems.* Comparator control systems can compare predicted sensory feedback from movement with actual feedback during on-line environmental interactions. Such predictive simulations have two important on-line functions:

(A) Permitting smoother, faster movements directed at a certain target, by comparison with movements controlled solely by actual feedback. A thermostat can function more efficiently by predicting room temperature and turning the heat off before reaching target temperature, to avoid overshooting. Similarly, bodily control and instrumental movement can be more efficient when predictive simulations are available during on-line interactions with the environment.

(B) Distinguishing sensory events deriving from exogenous environmental events from those resulting from endogenous movements. Once correlations are established between actual and simulated feedback from movement, divergence between them can indicate an exogenous rather than endogenous source of sensory input, making a contribution to enabling sensory experience.

Note that it is the extended dynamic, including external and internal feedback loops, that provides improved control and distinguishes exogenous and endogenous events; internal simulations alone would not do this work.

*(2) Off-line simulations with monitoring of inhibition.* Once internal simulations of the results of movement are available for on-line functions, they can be exapted for off-line use also, permitting the results of inhibited movement to be simulated. Off-line processes detach predictive simulative from the environmental aspects of the on-line dynamics with which it was originally coupled. Off-line simulations can enable instrumental cognition such as imagining the likely results of your own alternative acts and assessing which is the best means to a goal, instead of relying on costly trial and error learning. Dennett's 'Popperian' animals let simulations die in their stead.[19]

For off-line simulations to do this enabling work, the information must be available that they are off-line simulations of results of inhibited possible movements, not on-line simulations of results of actual movements. Very different responses are appropriate when simulations predict results of possible as opposed to actual movements. Two capacities thus work together in this enabling explanation: capacities to simulate off-line while inhibiting actual movement, and to monitor off-line status or inhibition.

*(3) Off-line simulations without monitoring of inhibition.* However, these two capacities might dissociate: off-line simulations might occur without monitoring of their off-line status. This could explain how some illusions or hallucinations work: if on-line simulation normally makes an enabling contribution to sensory experience (1B), then when off-line simulation occurs without inhibition monitoring so is not distinguished from on-line simulation (2), it can be predicted to have effects on sensory experience. Resulting illusions would be a natural by-product of cognitive functions enabled by off-line simulation, which in turn is a by-product of functions enabled by on-line simulation.

Put the other way round, in this account the contribution of off-line simulations to enabling illusions presupposes their contribution to enabling instrumental cognition, which in turn presupposes the contribution of on-line simulations to enabling effectively controlled movement and to distinguishing endogenous from exogenous sensory events. The account is not obviously biased towards how-externalism; Grush (2003) tells a story similar to parts of this account in arguing against enabling externalism. So why do I think something like this account favors enabling externalism?

As I see it, the issue is this. Internal simulations can occur on-line, as part of an extended dynamical process (as in context 1 above) or off-line (as in contexts 2 and 3); in both cases, they can provide at least part of a quality-enabling process. When internal simulations occur off-line, they can provide internal enabling explanations of qualities of experience (as in context 3). But do internal simulations alone provide the best quality-enabling explanation *when they occur on line, embedded in an extended dynamic?* Or can an extended dynamic that includes internal simulations provide the best explanation of qualities of on-line experience? Arguably, internal simulations are necessary for the enabling of experience; if so, a creature with no predictive simulations, and only external feedback control mechanisms would lack experience with phenomenal qualities. But it remains open whether enabling explanations must be purely internal in on-line as well as off-line cases. The internalist holds that the internal simulations that explain how qualities are enabled off-line also explain how they are enabled on-line; the external portions of the extended dynamics in on-line cases are not part of the enabling process. The externalist holds that even though internal simulations explain how qualities are

[19] See Millikan's (2006) squirrel; Hurley 2005 relates predictive simulation to processes that can enable understanding of others' actions.

enabled off-line, extended dynamics that include internal simulations can explain how qualities are enabled on-line; in on-line cases, the external portions of the extended dynamic can be of the enabling process.

How should this issue be decided? Can the processes that enable a given quality of experience vary? Can they be internal for some instances of the quality, and extended for others? The internalist may argue that if the same qualities of experiences can result off-line (say, in hallucinations) as on-line, then the external parts of extended dynamics are not needed to explain how experience works, any more than to explain what it's like. But as we've seen in section 3, this type of argument about what-quality explanations doesn't work. The neural correlates of a given quality can vary across illusory and veridical cases, and across veridical cases. On this basis I argued that extended dynamics can in some cases provide what-quality explanations, which explain sameness of quality despite varying neural correlates. For example, an extended dynamic might explain the quality shared by the illusion of movement in the paralyzed eye case and the veridical perception of movement in the sideways earthquake case, despite different neural correlates. It's no objection when explaining quality type that the type-explanatory external factors are absent in illusory cases, any more than it's an objection when explaining content type. The extended dynamic in which an internal simulation normally participates can explain quality type in illusory cases, just as normal causes might explain content-type in cases of mistake, where normal causes are absent (recall Burge 1986 on cracks and shadows; ditto proper functions).

The internalist argument doesn't work for quality-enabling explanations either, for related reasons. We've seen that neural correlates can vary across veridical and illusory instances of the same quality, but still be collected under one extended what-quality explanation. So why not allow that quality-enabling processes can vary so as to be extended in on-line cases and internal in off-line cases of the same quality, but similarly be collected under the same extended what-quality explanation? Externalists claim that, in some on-line cases, what enables qualitative experience is ongoing embodied interactions with the environment, probings and samplings and movements with external feedback loops intact, not merely the internal simulative portions of those interactions (see Keijzer and Schouten, in press, on change-blindness). If so, the extended what-quality explanation and the corresponding extended quality-enabling explanation in on-line cases would converge in the natural world.

Support for this externalist claim is provided by the above account of how the off-line enabling roles of internal simulations presuppose their more fundamental roles within extended dynamics. On this account, the enabling roles of internal simulations are explanatorily derivative from, not independent of, their role in on-line dynamics. Enabling explanations in on-line cases have explanatory priority, just as what-quality explanations in veridical cases do. By contrast, the internalist view that internal simulations explain how qualities are enabled both on-line and off-line gives internal simulations explanatory independence of extended on-line dynamics. This gets the cart before the horse. If the enabling role of internal simulations in off-line cases is derivative from their role in extended dynamics, it provides no reason to hold that only internal processes can do quality-enabling work in the primary, on-line cases.

So far in this section I've argued that extended on-line dynamics are explanatorily prior to off-line simulations thereof, and that this supports the externalist view that extended dynamics can provide quality-enabling explanations in on-line cases, even though internal simulations do so in off-line cases.

*5.3. Neural plasticity and development: Avoiding the internal endpoint error.* Extended on-line dynamics provide internal simulations thereof with ongoing tuning and maintenance (see also Clark 1997, 479).

The way many illusions adapt away, yielding variation in neural correlates of given qualities, illustrates how on-line processes, with the external loops of their dynamics intact, continually set and reset the parameters of off-line simulations. Illusory experiences can themselves reflect on-going tuning by on-line dynamics—e.g. illusory after-effects of adaptation when goggles are removed.

The internalist may regard the tuning and maintenance of internal simulations by extended dynamics as 'merely causal, not constitutive': processes of acquisition, over development or learning, of a mature capacity for the internal processes that do the real quality-enabling work. However, if we avoid the 'causal-constitutive error' error of assuming that only internalist explanations can be constitutive, we shouldn't assume extended tuning and maintenance processes can not be part of the sought-for explanation of how experience works, as well as of what it is like.

In particular, the distinction between acquisition and mature capacity should be treated with empirical caution in this context. Nervous systems, especially human ones, are by nature more plastic than we've tended to suppose. Neural correlates don't vary only in response to distorting goggles and pathologies such as congenital blindness. Over normal childhood and adolescence, the overall shape of the neural correlates of many types of experience changes dramatically, from relatively diffuse and bilateral to more efficiently localized, while the capacity for the relevant experiences is sustained (Huttenlocher 2002). This makes good evolutionary sense, allowing environmental interactions to influence the efficient specification of neural functions. But we shouldn't assume that it is only *after* on-line processes have finally fixed the parameters of internal simulations that given types of experience can be enabled: this is the *internal endpoint error* (a close relation of the 'causal/constitutive error' error). Rather, many neural processes are continually open to and re-parameterized by on-line interactions with the environment, as body and brain grow, and into adulthood. Quality-enabling externalism holds that extended on-line dynamics needn't be just a way of acquiring a mature capacity for an internal endpoint, but can enable and sustain qualities of experience across normal developmental variation in neural correlates.

Quality-enabling externalism may not be as radical as it first seemed. The view isn't that external factors by themselves enable experience, or that internal factors by themselves cannot enable experience. Rather, it's that *purely* internal processes are not the *only* way experience can be enabled. In on-line cases, what the world is like can be part of what enables us to experience what it is like. Evolution has no reason of principle to respect the skin in enabling experience, no reason not to enable experience by exploiting both interactions with the world and internal processes. It may be a mystery why evolution should enable experience at all--but that point is a double-edged sword, as the magical membrane problem reveals. If we really have no idea how experience is enabled, why be so sure the explanation must be internal? Perhaps inner-outer interactions are part of the needed gap-antidote.

## 6. Concluding summary.

*Taxonomy.* I've distinguished 'what' externalism, about the content or quality of mental states, from 'how' externalism, about the processes that enable mental states with given contents or qualities. A two-by-two taxonomy of varieties of externalism results: what-content externalism, what-quality externalism, content-enabling externalism, and quality-enabling externalism. The 'what'/'how' distinction doesn't align cleanly with a constitutive/causal or an internal/external distinction. Many intuitions resist moves from 'what' to 'how' externalism or from content to quality externalism, and are most resistant to that most exotic form of externalism, about the processes that enable phenomenal qualities.

| SUMMARY | | |
|---|---|---|
| | *content* | *quality* |
| *what* | •Supervenience vs. explanation<br>•Supervenience thought experiments and control<br>• Separability and Unpluggability | •Magical membrane problem<br>•Variable neural correlates<br>•Illusions & brains in vats<br>•'Causal/constitutive error' error |
| *how* | •Cultural extension: external representations, raise issues about underived content<br>•Cf. Natural extension: more basic issues | •Natural extension: extended dynamics can enable on-line while internal simulations thereof enable off-line<br>•Explanatory priority of on-interactions to simulations |

Two general principles have animated my discussion under these headings: First, externalism should be understood in both its 'what' and 'how' varieties as making explanatory rather than metaphysical claims. Second, veridical and on-line cases are explanatorily prior to cases involving illusions or hallucinations and to off-line cases.

*What-content externalism* is usually supported by externalist intuitions in supervenience thought experiments (STEs), which postulate twins who are internal duplicates but embedded in different environments. STEs are controlled thought experiments that seek to separate out the explanatory roles of internal and external factors; they presuppose explanatory separability, which requires that internal factors be unpluggable from external factors. Since the truth of an internal supervenience claim does not require unpluggability, internal supervenience is necessary but not sufficient for the possibility of an STE. Internalist explanation requires explanatory separability and unpluggability; if the relevant STE is not possible, internal supervenience provides no support for internalist explanation. Supervenience claims should aim to draw boundaries that are neither too wide, including explanatorily redundant factors, nor too narrow, cutting between explanatorily nonseparable factors. Supervenience boundaries should be open to revision in a process of reflective equilibrium between intuitive evidence and theorizing.

*What-quality externalism.* Intuitions about content in STEs provide authoritative evidence, but intuitions about quality don't. While STEs for phenomenal qualities typically yield strong internalist intuitions, these co-exist with meta-intuitions to the effect that qualities are ultimately autonomous, that internalist intuitions could just turn out to be wrong. Such autonomy meta-intuitions express the explanatory gap separating neural processes and internal functions from phenomenal qualities. The combination of strong internalist intuitions with autonomy meta-intuitions presents a puzzle about the explanatory significance of the internal/external boundary, the magical membrane problem: if we have so little understanding of how phenomenal qualities could possibly be explained, why are we so confident that *if* they can be, the explanation must be internalist?

Two internalist responses to the magical membrane problem were considered, both of which attempt to support internalist intuitions in STEs: one on the basis of illusions, the other on the basis of

brains in vats. In reply I argued that neural correlates of a given quality can vary, across normal development and perceptual adaptation, as well as between illusory and veridical experiences. Variable neural correlates make trouble for both responses: what explains why they are collected by the same quality type? In some cases, plausible answers can be provided by extended dynamics in which internal and external factors are not explanatorily separable but which admit of varying neural implementations. If so, what the world is like can be part of what explains what experience is like.

Externalist what-quality explanations need not provide counterexamples to supervenience claims: without unpluggability and separability, there's no violation of internal supervenience. Failures of internal supervenience are not the touchstone of externalism; rather, externalism competes case by case with internalism to provide the better explanation. There's no shortcut to internalism via the claim that only internalist explanations are constitutive; we should avoid the 'causal/constitutive error' error, of assuming a causal/constitutive distinction that coincides with an external/internal distinction.

*Content-enabling externalism.* Extended conceptions of the processes that enable cognition often appeal to cultural examples, involving interactions with external representations, such as Otto's notebook or the accountant's pencil and paper, plus a principle of parity. One internalist objection to cultural arguments for extended cognition claims that cognitive processes must have underived content, which internal representations have and external representations lack. Now the underived content criterion itself raises difficulties, and has rivals, such as an availability off-line criterion motivated by dynamical cognitive science. But these issues about relations between the contents of internal and external representations distract attention from a more basic underlying issue, concerning relations between on-line interactions with the natural world—not external representations of it—and off-line simulations of such interactions.

*Quality-enabling externalism.* What-quality externalism may seem radical enough already. Why not stop at a middle ground that combines what-quality externalism with quality-enabling internalism, instead of going all the way to the latter? Quality-enabling externalism claims that, in some on-line cases, what enables qualitative experience is ongoing embodied interactions with the environment, probings and samplings and movements with external feedback loops intact, not merely the internal simulative portions of those interactions. By contrast, internalism claims that the internal simulations that explain how qualities are enabled off-line also explain how they are enabled on-line. If the same qualities of experiences can result off-line (say, in hallucinations) as on-line, it may be argued, then the external parts of extended dynamics are not needed to explain how experience works, any more than to explain what it's like. However, I've argued that neural correlates can vary across veridical and illusory instances of the same quality, yet still be collected under one extended what-quality explanation. If so, why not allow that the processes that enable a given quality of experience can be internal in off-line cases and extended in on-line cases? The explanatory priority of extended on-line interactions with the natural world to internal simulations thereof supports the externalist view that extended dynamics can provide quality-enabling explanations in on-line cases, even though internal simulations do so in off-line cases. What the world we are interacting with is like can be part of what enables us to experience what it is like.

References

Adams, Fred, and Aizawa, Ken, 2001. The bounds of cognition. *Philosophical Psychology* 14(1):???.

Block, Ned, 2005. *Journal of Philosophy.*

Brooks, Rodney, 1999. *Cambrian Intelligence.* Cambridge, MA: MIT Press.

Brooks, Rodney, 1991. Intelligence without representation. *Artificial intelligence* 47:139-159.

Burge, Tyler, 1986. Cartesian error and the objectivity of perception. In *Subject, Thought, and Context*, eds. Phillip Pettit and John McDowell, 117-136. Oxford: Clarendon Press.

Clark, Andy, 1997. The dynamical challenge. *Cognitive Science* 21(4):461-481.

Clark, Andy, 2005. Intrinsic content, active memory and the extended mind. *Analysis* 65(1):1-11.

Clark, Andy, in progress. Material symbols and the extended mind.

Clark, Andy, and Chalmers, David, 1998. The extended mind. *Analysis* 58(1):7-19.

Clark, Andy, and Grush, Rick, 1999. Towards a cognitive robotics. *Adaptive Behavior* 7(1):5-16.

Clark, Andy, forthcoming. Pressing the flesh: Exploring a tension in the study of the embodied, embedded mind.

Clark, Andy, forthcoming. Word, niche, and super-niche: How language makes minds matter more. *Theoria* 20:54:2005 p. 255-268. Special issue on Language and Thought: Empirical and Conceptual Viewpoints, eds. J. Acero and F. Rodriguez.

Dawkins, Richard, 1982. *The Extended Phenotype.* Oxford: Oxford University Press.

De Gaynesford, Maximilian, 2004. *John McDowell.* Polity Press.

Dennett, Daniel C., 1991. *Consciousness Explained.* Boston: Little Brown.

Dretske, Fred, 1996. Phenomenal externalism: If meanings ain't in the head, where are qualia? *Philosophical Issues*, 7, Enrique Villanueva, ed. Ridgeview Publishing Company; Atascadero, CA. Pp. 143-158.

Edelman, 2005. Mostly harmless.

Greenberg, Mark, forthcoming. A new map of theories of mental content: Locating normative theories. *Philosophical Issues.*

Grush, Rick, 2003. In defense of some 'Cartesian' assumptions concerning the brain and its operation. *Biology and Philosophy* 18:53-93.

Harman, Gilbert, 1990. The intrinsic quality of experience. In *Philosophical Perspectives*, vol. 4, *Action Theory and Philosophy of Mind*, ed. James Tomberlin, 31-52. Atascadero, Calif.: Ridgeview.

Hurley, Susan, 1998a. *Consciousness in Action*. Cambridge, MA: Harvard University Press.

Hurley, Susan, 1998b. Vehicles, contents, conceptual structure, and externalism. *Analysis* 58(1):1-6.

Hurley, Susan, 2003. Action, the unity of consciousness, and vehicle externalism. In *The Unity of Consciousness: Binding, Integration, and Dissociation*, Axel Cleeremans, ed., Oxford: Oxford University Press, pp. 78-91.

Hurley, Susan, 2005. "The shared circuits model: how control, mirroring and simulation can enable imitation and mind reading", at: http://www.interdisciplines.org/mirror/papers/5.

Hurley, Susan, and Noë, Alva, 2003a. Neural plasticity and consciousness. *Biology and Philosophy 18*:131-168.

Hurley, Susan, and Noë, Alva, 2003b. Reply to Block. *Trends in Cognitive Science* 7(8):342.

Hurley, Susan, and Noë, Alva, in press. Can hunter-gatherers hear color? *In Common Minds: Essays in Honor of Philip Pettit*. Eds. G. Brennan, R. Goodin, F. Jackson, and M. Smith. (Oxford: Oxford University Press).

Huttenlocher, Peter R., 2002. *Neural Plasticity: The Effects of Environment on the Development of the Cerebral Cortex*. Cambridge, Mass: Harvard University Press.

Jacob, Pierre, 2002. Review of *The Body in Mind*, by Mark Rowlands. *Mind and Language* 17(3):325-331.

Johnston, Mark, 2003. The obscure object of hallucination. *Philosophical Studies* ??:113-183.

Keijzer, Fred, 2001. *Representation and Behavior*. Cambridge, MA: MIT Press.

Keijzer, Fred, and Schouten, Maurice, in press. Embedded cognition and mental causation: Setting empirical bounds on metaphysics. *Synthese*.

Kim, Jaegwon, 1993. *Supervenience and Mind*. Cambridge, UK: Cambridge University Press.

Kohler, I. (1951). "Über Aufbau und Wandlungen der Wahrnehmungswelt." *Österreichische Akademie der Wissenschaften. Sitzungsberichte, philosophish-historische Klasse, 227*, 1-118.

Kohler, I. (1964) *The formation and transformation of the perceptual world*. Published as a monograph in *Psychological Issues* vol.3 (monograph 12). New York International University Press. [This is a translation of Kohler 1951.]

Kupers, Ron, and Ptito, Maurice (2004). "Seeing" through the tongue: Cross-modal plasticity in the congenitally blind. *International Congress Series* (Frontiers in Human Brain Topology. Proc. ISBET 2004, the 15th World Congress of the International Society of Brain Electromagnetic Topography), Vol. 1270, pp. 79-84, August 2004.

Maudlin, Tim, 1989. "Computation and Consciousness", *Journal of Philosophy* 86, pp. 407-432

Menary, Richard, forthcoming.

McDowell, John, 1994. The content of perceptual experience. *Philosophical Quarterly* 44(175): 190-205.

Noë, Alva, 2004. *Action in Perception.* Cambridge, MA: MIT Press.

Noë, Alva, Pessoa, L, and Thomspon, E., 2000. Beyond the grand illusion: What change blindness really teaches us about vision. *Visual Cognition* 7:93-106.

O'Regan, J. Kevin, and Noë, Alva, 2001a. A sensorimotor approach to vision and visual consciousness. *Behavioral and Brain Sciences* 24(5):939-973.

O'Regan, J. Kevin, and Noë, Alva, 2001b. What it is like to see: A sensorimotor theory of perceptual experience. *Synthese* 29:79-103.

Rowlands, Mark, 1999. *The Body in Mind.* Cambridge: Cambridge University Press.

Rowlands, Mark, 2003. *Externalism.* Chesham, Buckinghamshire: Acumen Press.

Thelan, Esther, and Smith, Linda B., 1994. *A Dynamical Systems Approach to the Development of Cognition and Action.* Cambridge, Mass.: MIT Press.

van Gelder, T.. 1999a. Revisiting the Dynamical Hypothesis. Preprint No. 2/99, University of Melbourne, Department of Philosophy.

van Gelder, T. J. 1999b. Dynamic approaches to cognition. In R. Wilson & F. Keil ed., *The MIT Encyclopedia of Cognitive Sciences*. Cambridge MA: MIT Press, 243-6.

van Gelder, Timothy J., 1995. What might cognition be, if not computation? *Journal of Philosophy* 91:345-381.

van Gelder, Timothy J., 1998. The dynamical hypothesis in cognitive science. *Behavioral and Brain Sciences* 21:1-14.

Wheeler, Michael, 1997. Cognition's coming home: The reunion of life and mind. In *Proceedings of the 4th European Conference on Artificial Life*, eds. P. Husbands and I. Harvey. Cambridge, MA: MIT Press.

Wheeler, Michael, 2001. Two threats to representation. *Synthese* 129: 211-231.

Wheeler, M., 2003. Do genes code for traits? In A. Rojszczak, J. Cachro and G. Kurczewski (eds.), *Philosophical Dimensions of Logic and Science: Selected Contributed Papers from the 11th International Congress of Logic, Methodology, and Philosophy of Science. Synthese Library* vol. 320, 151-164. Kluwer.

Wheeler, Michael, forthcoming. How to do things with (and without) representations. In *The Extended Mind*, R. Menary, ed. Ashgate.

Wheeler, M., and Clark, A., 1999. Genic Representation: Reconciling Content and Causal Complexity. *The British Journal for the Philosophy of Science*, 50 (1):103-135

Wilson, Robert A., 2004. *Boundaries of the Mind.* Cambridge: Cambridge University Press.