[To appear in R. Stainton, ed., *Contemporary Debates in Cognitive Science* (Blackwell)]

Can Cognition be Factorised into Internal and External Components?

Timothy Williamson

0. Platitudinously, cognitive science is the science of cognition. Cognition is usually defined as something like the process of acquiring, retaining and applying knowledge. To a first approximation, therefore, cognitive science is the science of knowing. Knowing is a relation between the knower and the known. Typically, although not always, what is known involves the environment external to the knower. Thus knowing typically involves a relation between the agent and the external environment. It is not internal to the agent, for the internal may be the same whether or not it is related to the external in a way that constitutes knowing. Cognition enables agents to achieve their goals by adjusting their actions appropriately to the environment. Such adjustment requires what is internal to the agent to be in some sense in line with what is external; that matching depends on both internal and external sides. Thus if cognitive science were restricted to what is internal to the agent, it would lose sight of its primary object of study.

Although cognition depends on both the internal and the external, one can try to analyse it into internal and external factors. Call a state S *narrow* if and only if whether an agent is in S at a time *t* depends only on the total internal qualitative state of S at *t*, so that if one agent in one possible situation is internally an exact duplicate of another agent

in another possible situation, then the first agent is in S in the first situation if and only if the second agent is in S in the second situation.[1] Call S *broad* if and only if it is not narrow. A state is *environmental* if and only if it depends only on the total state of the environment external to the given agent (as it were, the total internal state of the external environment) at the relevant time. The factorising strategy attempts to analyse cognitive states into combinations of narrow states and environmental states (likewise for cognitive processes). If the strategy succeeds, it may be tempting to regard the narrow states that emerge from that analysis as the only strictly mental or psychological states, at least for the purposes of cognitive science. If the strategy fails, it will be tempting to regard some broad cognitive states as themselves strictly mental for the purposes of cognitive science. *Internalism* is here the view that (for the purposes of cognitive science) all mental states are narrow. *Externalism* is the negation of internalism. Of course, what matters is not so much how we use the word 'mental' as whether the factorising strategy succeeds at all.

For present purposes let us not worry about exactly what counts as internal: (internal to the brain or merely to the body, and 'internal' in what sense?). A radical view is that there is no natural boundary between the internal and the external: they somehow flow into each other. That would be bad news for internalism rather than externalism, since only the former attributes a central significance to the internal-external boundary with respect to the mental. Thus it is legitimate to assume a natural internal-external boundary for the sake of an argument *against* internalism.[2] But given that plausible assumption, internalism may seem an obviously attractive or even compelling view. Isn't the aim of cognitive science precisely to identify the internal mechanisms that are the

contribution of mind, and to understand how their interaction with the external environment produces knowledge?

The aim of this chapter is to sketch an argument that the internalist programme of factorisation rests on a myopic view of what it would take to understand the role of cognition in guiding action. The cognitive states that explain action at an appropriate level of generality are not merely broad; they are broad in a way that makes their factorisation into narrow and environmental constituents impossible in principle.

The internalism-externalism debate has ramified in many directions over recent decades. It would be hopeless to attempt here to survey all the arguments that have been advanced in favour of internalism. Instead, one central line of externalist thought will be developed. That should give the reader a sufficient clue as to where the internalist arguments go wrong.[3]

Section 1 considers two main ways in which the intentional states of folk psychology as delineated by natural language predicates are broad. Internalists may respond by denying that those folk psychological states are strictly mental in the sense relevant to cognitive science and attempting to factorise them into narrow and environmental components. Section 2 sketches some abstract structural considerations, to explain why the factorising strategy is problematic in principle. Section 3 develops that argument in more detail by reference to some examples. Section 4 suggests a more general moral about cognition.

1. Knowing is central to the subject matter of cognitive science; it is also a central source of broad intentional states. Consider, for instance, the state of knowing that one is holding

a glass of water, the state such that, necessarily, one is in it if and only if one knows that one is holding a glass of water. Two agents in different possible situations may be exact internal duplicates even though one of them knows that she is holding a glass of water while the other merely has a false belief that she is holding a glass of water (actually it is gin). One cannot know something false, although one can believe falsely that one knows it. Thus the state of knowing that one is holding a glass of water is broad.

Knowing is a *factive* attitude, in the sense that, necessarily, one has it only to truths. Other factive attitudes include seeing and remembering (in the usual senses of those terms). Unless the glass contains water, you cannot see that it contains water; if you are misperceiving, you only think that you can see that it contains water. Unless you drank water yesterday, you cannot remember that you drank water yesterday; if you are misremembering, you only think that you remember that you drank water yesterday. By contrast, the attitude of believing is non-factive; false belief is possible, indeed actual and widespread. Arguably, knowing is the most general factive attitude, the one that subsumes all others: any factive attitude is a form of knowing.

In unfavourable circumstances, for instance when they are being hoaxed, agents falsely believe that P, and believe that they know that P, without being in a position to know that they do not know that P. Thus agents are not always in a position to know which knowledge states they are in; such states lack first-person transparency. If first-person transparency were a necessary condition for being a mental state, knowledge states would not be mental. But why impose any such condition on mentality? There is no obvious inconsistency in the hypothesis that someone has dark desires without being in a position to know that he has them: for example, the opportunity to put them into effect

may not arise. It is a commonplace of cognitive science that agents' beliefs as to their mental states are sometimes badly mistaken. That is not to deny that we often know without observation whether we are in a given mental state: but then we often know without observation whether we are in a given knowledge state. The partial transparency of knowing is consistent with its being a mental state.

Do non-factive attitudes such as believing yield narrow mental states? Whatever the glass contains, one can *believe* that it contains water. Nevertheless, famous externalist arguments show that the content of intentional states can yield a broad mental state even when the attitude to that content is non-factive. Indeed, the current debate between internalism and externalism in the philosophy of mind originally arose with respect to the contents of the attitudes rather than the attitudes themselves.

We can adapt Hilary Putnam's classic example to argue that even the state of believing that one is holding a glass of water is broad.[4] Imagine a planet Twin-Earth exactly like Earth except that the liquid observed in rivers, seas and so on is not H20 but XYZ, which has the same easily observable properties as H20 but an utterly different chemical composition. Thus XYZ is not water. However, the thought experiment is set in 1750, before the chemical composition of water was known. Oscar on Earth and Twin-Oscar on Twin-Earth are internal duplicates of each other; they are in exactly the same narrow states. Oscar is holding a glass of H20; Twin-Oscar is holding a glass of XYZ. Since H20 is water and XYZ is not:

(1) Oscar is holding a glass of water.

(2) Twin-Oscar is not holding a glass of water.

They both express a belief by saying ‘I am holding a glass of water’. Since Oscar uses the word ‘water’ to refer to water, we can report his belief thus:

(3) Oscar believes that he [Oscar] is holding a glass of water.

Suppose (for a reductio ad absurdum) that believing that one is holding a glass of water is a narrow state. By (3), Oscar is in that state. By hypothesis, Oscar and Twin-Oscar are in exactly the same narrow states. Therefore Twin-Oscar is also in that state:

(4) Twin-Oscar believes that he [Twin-Oscar] is holding a glass of water.

Now consider elementary schemas governing true and false belief:

(T) If S believes that P, S believes truly that P if and only if P.

(F) If S believes that P, S believes falsely that P if and only if not P.

Here is an instance of (T):

(5) If Oscar believes that he [Oscar] is holding a glass of water, Oscar believes truly that he [Oscar] is holding a glass of water if and only if he [Oscar] is holding a glass of water.

Here is an instance of (F):

(6) If Twin-Oscar believes that he [Twin-Oscar] is holding a glass of water, Twin-Oscar believes falsely that he [Twin-Oscar] is holding a glass of water if and only if he [Twin-Oscar] is not holding a glass of water.

From (1), (3) and (5) we easily conclude:

(7) Oscar believes truly that he [Oscar] is holding a glass of water.

By the same token, from (2), (4) and (6) we conclude:

(8) Twin-Oscar believes falsely that he [Twin-Oscar] is holding a glass of water.

But (7) and (8) imply an implausible asymmetry between Oscar and Twin-Oscar. There is no more reason to impute error to Twin-Oscar than to Oscar. They manage equally well in their home environments. Since (1), (2), (3), (5) and (6) are clearly true, the culprit must be (4). Since (4) follows from (3) and the assumption that believing that one is

holding a glass of water is a narrow state, we conclude that believing that one is holding a glass of water is a broad state, not a narrow one.

The Putnam-inspired argument generalizes from believing to all sorts of other propositional attitudes, such as wondering, desiring and intending. It also generalizes beyond natural kind concepts like *water* to a huge array of contents that turn out to depend on the agent's natural or social environment, or simply on the identity of the agent. For example, the qualitatively identical twins Tweedledum and Tweedledee both speak truly when they simultaneously utter the words 'Only I believe that *I* won'; each correctly self-attributes a belief the other lacks. Similarly with perceptual demonstratives: both twins speak truly when they point at each other and simultaneously utter the words 'Only he believes that *he* won'. Intentional content as attributed in natural languages depends on reference, which in turn depends on causal and other relations of the agent that are left undetermined by all internal qualities. The broadness of reference extends further to folk psychological relational states such as seeing Vienna, thinking about Vienna and loving or hating Vienna. If intentionality (aboutness) is the mark of the mental, and our thoughts are typically about the external environment, then our mental states are typically broad.

Thus both the contents of folk psychological intentional states and the attitudes to those contents make the states broad. Of course, that does not prevent internalists from postulating a core of narrow mental states for the purposes of cognitive science and trying to identify them by more theoretical means. But it is not obvious that such a core exists. Granted, when human agents are in an intentional states they are also in some underlying narrow physical state or other, but it does not follow that the latter are intentional in any

useful sense. We can clarify the issue by considering the role of cognitive states in the causal explanation of action.

2. Suppose that we are trying to explain an agent's action in terms of the state of the world at a prior time *t*. A first internalist hypothesis is that all we really need consider is the agent's internal state at *t*. If causation is local, that will determine what she does (insofar as quantum mechanics allows it to be determined in advance at all); other aspects of the state of the world make a difference only to whether what she does gets her what she wants.

That internalist hypothesis raises a question about the individuation of actions. Suppose that the agent's action was *drinking from this stream here*. That action was not determined purely by her total internal state. She could have been in that state even if she had seen a numerically distinct but qualitatively identical stream, or been the victim of an illusion with no stream nearby. In the former case the same internal state would not have led to her drinking from *this* stream; in the latter it would not have led to her drinking from any stream at all. Many actions are individuated broadly and are consequently not determined by preceding narrow mental states (Evans 1982: 203). Of course, the internalist will now insist on individuating the action to be explained in narrow terms, for instance as *making as if to drink from a stream of this qualitative appearance*. There is a danger of a stand-off here, with the internalist accusing the externalist of begging the question by individuating the action to be explained broadly and the externalist in turn accusing the internalist of begging the question by individuating the action to be explained narrowly. The externalist can point out that the content of the agent's intention

typically corresponds to the broad individuation of the action in terms of the external objects that it involves; if we are trying to explain an intentional action, don't we need to explain the action as specified in the intention? The internalist may reply that the factorising strategy must be applied to the agent's intention too. Alternatively, the internalist explanation may be restricted to a class of basic actions more primitive than either *drinking from this stream here* or *making as if to drink from a stream of this qualitative appearance*.

Even if actions are narrowly individuated, other problems arise for the attempt to explain them on the basis of the agent's prior internal state. Action is typically not instantaneous, and not merely because there is no such moment as the one immediately preceding the moment of action if time is dense. It takes time even to bend one's head until one's lips touch the water and more time to drink. That is enough time to spot piranha fish in the water and withdraw before completing the action. Thus the internal state of the agent at *t* does not even determine whether she will go through the sequence of internal states corresponding to drinking in some brief interval shortly after *t*. Again, the internalist may respond by attempting to analyse extended actions such as bending and drinking into sequences of more basic actions, each to be explained in terms of an 'immediately' preceding internal state of the agent.

To restrict ourselves to concatenating explanations of basic actions would drastically curtail our explanatory ambitions. Consider a tourist in a strange town whose camera and passport have just been stolen. We may be able to explain on general grounds why he will sooner or later get to the local police station, whose location he does not yet know, without assuming anything specific about the layout of the town, his current

position in it, or his beliefs about those things. By contrast, if we concatenate explanations of the basic actions by which, step by step, he actually gets to the local police station in terms of 'immediately' preceding internal states, we shall need to invoke a mass of detailed extra assumptions about the sequence of his mental states (for example, his perceptions of the streets around him and of the responses to his questions about the way to the nearest police station) far beyond anything that the more general explanation need assume. Thus by moving down to the level of basic actions we lose significant generalisations that are available at the higher level. If the layout of the town, his position in it or the responses to his questions had been different, he would have gone through a different sequence of basic actions, but he would still have got to the police station in the end. To understand cognition at an appropriate level of generality, we often need to understand non-basic actions themselves, not merely the sequences of basic actions that realize them on particular occasions. Since the performance of non-basic actions depends on the environment as well as the agent, they are not to be explained solely in terms of the agent's internal states.

Appeals to the locality of causation do not motivate the idea that only narrow states of the agent at *t* are causally relevant to effects that are not complete until some time after *t*, for environmental states at *t* are also causally relevant to those effects. Rather, the issue is whether the causally relevant states at *t* can be factorised into narrow states and environmental states.

We should therefore assume that when we are trying to explain the agent's action in terms of the state of the world at *t*, both the state of the agent at *t* and the state of the external environment at *t* are potentially relevant. Nevertheless, it might be argued, the

factorisation strategy is bound to succeed, for the *total* (maximally specific) narrow state of the agent at *t* and the *total* environmental state (state of the external environment) at *t* together determine the total state of the whole world at *t* (no difference in the latter without a difference in one of the former), which in turn determines the action (insofar as it is determined at *t* at all). Let us grant the determination claim, although it is not completely uncontentious ─ in principle there might be global states of the whole world not determined by the combination of states local to the agent and states local to the external environment. Even so, when we explain an action we do not assume a particular maximally specific state of the world at *t*. It is not just that we cannot know exactly what the state of the world was at *t*. Even if we did know, it would be methodologically undesirable to build all that knowledge into our explanation of the action, because the maximal specificity of the assumption implies the minimal generality of the explanation. It would not apply to any of those alternative states of the world that would have led to the same action. A significant generalisation would again have been missed. Compare Putnam's example (1978: 42):

> A peg (1 inch square) goes through a 1 inch square hole and not through a 1 inch round hole. Explanation: (?) The peg consists of such-and-such elementary particles in such-and-such a lattice arrangement. By computing all the trajectories we can get applying forces to the peg (subject to the constraint that the forces must not be so great as to distort the peg or the holes) in the fashion of the famous Laplacian super-mind, we determine that some trajectory takes the peg through the square hole, and no

trajectories take it through the round hole. (Covering laws: the laws of physics.)

In our cases of interest, what matters is whether an explanation that assumes an unspecific state of the world at *t* (very roughly, the one necessary and sufficient for the action to be taken) can be analysed into an explanation that assumes a narrow state of the agent at *t* and an environmental state at *t*, not both of them maximally specific.[5]

The answer depends on the nature of the assumed unspecific state of the world. To see this, consider a very simple model with just three possible maximally specific narrow states of the agent, I1, I2 and I3, and three possible maximally specific environmental states, E1, E2 and E3. Any possible narrow state is compossible with any possible environmental state, so there are just nine possible maximally specific states of the whole world, each of the form I*i*&E*j*. For example, the world is in the unspecific disjunctive state (I1&E1) ∨ (I1&E2) ∨ (I2&E1) ∨ (I2&E2) when and only when it is in one of the specific states I1&E1, I1&E2, I2&E1 and I2&E2. Thus that unspecific state is equivalent to the state (I1 ∨ I2)& (E1 ∨ E2), which the world is in when and only when the agent is in the unspecific narrow state I1 ∨ I2 and the external environment is in the unspecific state E1 ∨ E2. Contrast that with the unspecific state of the world (I1&E1) ∨ (I2&E2) ∨ (I3&E3). The latter is not equivalent to any conjunction of the form I&E, where I is a narrow state and E is an environmental state. For suppose that such an equivalence holds. Then I is a disjunction of some subset of I1, I2 and I3. If I1 is not in the subset then I excludes I1, so I&E excludes I1, which it cannot because I1&E1 is possible and by hypothesis implies I&E. Thus I1 is in the disjunction, so I1 implies I. By a parallel argument, E2 implies E. But then I1&E2 implies I&E, which it cannot, for by

hypothesis I&E excludes I1&E2. Hence (I1&E1) ∨ (I2&E2) ∨ (I3&E3) is not equivalent to the conjunction of a narrow state and an environmental state. Thus whether assuming an unspecific state of the world is equivalent to assuming a narrow state and an environmental state depends on whether that state of the world is like (I1&E1) ∨ (I1&E2) ∨ (I2&E1) ∨ (I2&E2) or like (I1&E1) ∨ (I2&E2) ∨ (I3&E3).

Call a state *composite* if and only if it is equivalent to the conjunction of a narrow state and an environmental state. All narrow states trivially count as composite, because they are equivalent to the conjunction of themselves and the maximally unspecific environmental state that always obtains; but not all composite states are narrow. For example, I1&E1 is composite but not narrow. Call a state *prime* if and only if it is not composite. Trivially, all prime states are broad; but not all broad states are prime. I1&E1 is broad but not prime. In the toy model, there are 512 states altogether, of which 8 are narrow and 504 broad, 50 composite and 462 prime.[6]

We already have a very general reason to expect that many cognitive states will be prime. For we noted that cognition is supposed to help agents achieve their goals by interacting more successfully with the environment, which typically requires their narrow states to be in some sense in line with environmental states. That idea is vague, but it implies at least this much: in the relevant respect, a specific narrow state I1 may be in line with a specific environmental state E1 but not with a specific alternative environmental state E2, while a specific alternative narrow state I2 is in line with E2 but not with E1. Now consider the cognitive state C that the agent is in when and only when her specific internal state is in line with the specific environmental state in the given way. Thus the conjunctive states I1&E1 and I2&E2 imply C, while the conjunctive states I1&E2 and

I2&E1 exclude C. But then an argument just like that two paragraphs back shows that C must be prime. On purely structural grounds, the state of being in some total narrow state or other that matches the total environmental state in some given way cannot be equivalent to the conjunction of a narrow state and an environmental state, however unspecific. We will see later how to apply this abstract point to particular cases.

Could an internalist object that although a state such as (I1&E1) ∨ (I2&E2) ∨ (I3&E3) is prime, it has still been analysed into a disjunction of conjunctions of narrow states and environmental states, so the factorising strategy remains viable? That objection trivialises the factorising strategy in several ways:

First, the relevant disjunctions will in practice have infinitely many disjuncts, or at least vastly more than we can list. It is unlikely that such analyses will be available to cognitive science in usable form.

Second, the possibility of such analyses was not derived from any distinctive feature of the internal-external boundary, but merely from the quite general idea that for any boundary whatsoever, the total global state of the world is fixed by the conjunction of the total local state on one side of the boundary and the total local state on the other side. That holds even for utterly *ad hoc*, arbitrary, unnatural boundaries. Thus the mere possibility of such analyses is uninformative. That a cognitive state is equivalent to a disjunction of conjunctions of states tailored to any one of those gerrymandered or irrelevant boundaries shows nothing interesting about its underlying nature.

Third, and most important, if a prime state C is equivalent to a disjunction of conjunctions of narrow states and environmental states, it does not follow that an explanation that postulates C can be reduced to an explanation that postulates a narrow

state and an environmental state. That would follow only on the further assumption that C is equivalent to the conjunction of those states, in which case C is composite, contrary to hypothesis. Of course, on any particular occasion on which C obtains, so too does some specific conjunction I*i*&E*j* that implies C (but not vice versa), where I*i* is narrow and E*j* environmental. But it would be a mistake to conclude that the 'real explanation' invokes I*i*&E*j* rather than C. For the explanation that invokes I*i*&E*j* typically involves a drastic loss of generality. It applies only to one specific type of case, whereas the explanation that invoked C covered a greater variety of cases. As already emphasized, significant generality is a crucial virtue in scientific explanations. The moral is that when an explanation assumes a given prime state, there need be no pair of a narrow state and an environmental state by assuming which one can do the same explanatory work.

An explanation that invokes a composite state may be able to achieve some generality on one side of the internal-external boundary by sacrificing generality on the other side. For example, fix the total environmental state E*i*. Let I be the disjunction of all total narrow states that, conjoined with E*i*, yield the outcome that we are trying to explain. Thus I may have some generality, since a range of different internal states in that environment may yield the outcome (some features of the tourist's brain make no difference to whether he eventually gets to the police station). Since I too is narrow, the state I&E*i* is composite, and yields the outcome to be explained. But even this explanation involves an undesirable loss of generality: most obviously with respect to the environmental state, but also with respect to the agent's internal state, since I excludes internal states that would yield the outcome only in combination with environmental states other than E*i* (the tourist's intention to wait for the bus to pass before crossing the

road would not have been needed for getting to the police station if the road had been empty). Similarly, fix the total internal state I*i*. Let E be the disjunction of all total environmental states that, conjoined with I*i*, yield the outcome that we are trying to explain. Thus E may have some generality, since I*i* may yield the outcome in a range of different environmental states. The composite state I*i*&E also yields the outcome to be explained. But this explanation too involves an undesirable loss of generality: most obviously with respect to the internal state, but also with respect to the environmental state, since E excludes environmental states in which only another internal state would yield the relevant outcome (consider the tourist crossing a road).

Whichever way we turn, we attain the requisite level of generality in explaining action only if our explanations cite prime states of agents.

3. To argue that a cognitive state is composite, we need only find a narrow state and an environmental state of which it is the conjunction. But how can we argue that a cognitive state is prime? We cannot go through all possible conjunctions of narrow states and environmental states and argue case by case that it is not equivalent to any of them. A more useful criterion is this:

PRIME A state S is prime if and only if some narrow state I and some environmental state E are both separately compatible with S but I&E is incompatible with S.[7]

Suppose that we can find possible combinations of narrow states I1 and I2 and environmental states E1 and E2 such that I1&E1 and I2&E2 imply the cognitive state C while I1&E2 excludes C. Then I1 and E2 are compatible with C while I1&E2 is incompatible with C. Consequently, by PRIME, C is prime. Furthermore, PRIME enables one to argue that for each prime state there is a pair like I1 and E2.

For a simple example, let C be the unspecific state of *knowing which direction home is in* (in egocentric space). Consider two possible scenarios. In scenario 1, you know which direction home is in by knowing that home is straight in front of you. Thus home *is* straight in front of you, and you believe that it is. Your total narrow state is I1; the total environmental state is E1. In scenario 2, you know which direction home is in by knowing that home is straight behind you. Thus home *is* straight behind you, and you believe that it is. Your total narrow state is I2; the total environmental state is E2. The narrow state I1 is compatible with C because you are simultaneously in both in possible scenario 1. The environmental state E2 is compatible with C because you are simultaneously in both in possible scenario 2. But the conjunctive state I1&E2 is incompatible with C, because if I1&E2 obtains you do not know which direction home is in; rather, you believe that home is straight in front of you while in fact it is straight behind you; you have a mere false belief as to which direction home is in. Consequently, by PRIME, C is prime. The state of knowing which direction home is in is not equivalent to the conjunction of a narrow state and an environmental state.

For another example, let C* be the state of *consciously thinking about Rover* (not necessarily under the verbal mode of presentation 'Rover'). Again, consider two possible scenarios. In scenario 1*, you can see two very similar dogs, Rover on your right and

Mover on your left. You are consciously thinking about Rover, visually presented as 'that dog on my right' and in no other way; you are not consciously thinking about Mover at all. Your total narrow state is I1*; the total environmental state is E1*. In scenario 2*, you can see Rover on your left and Mover on your right. You are consciously thinking about Rover, visually presented as 'that dog on my left' and in no other way; you are not consciously thinking about Mover at all. Your total narrow state is I2*; the total environmental state is E2*. The narrow state I1* is compatible with C* because you are simultaneously in both in possible scenario 1*. The environmental state E2* is compatible with C* because you are simultaneously in both in possible scenario 2*. But the conjunctive state I1*&E2* is incompatible with C*, because if I1*&E2* obtains you are not consciously thinking about Rover; rather, you are consciously thinking about Mover, visually presented as 'that dog on my right'. Consequently, by PRIME, C* is prime. The state of consciously thinking about Rover is not equivalent to the conjunction of a narrow state and an environmental state.

It is not difficult to multiply such examples. In general, the folk psychological intentional states individuated by natural language predicates tend to be prime. But do those states matter for purposes of cognitive science?

Much depends on whether we are interested in short term or long term effects. The short term effects of merely believing that home is straight in front of you may be the same as the short term effects of knowing that home is straight in front of you. For example, if you want to go home, you take a step forward. Thus what the internalist may conceive as the narrow state of believing that home is straight in front of you may be more relevant in the short term than the broad state of knowing that home is in front of

you. But the long term effects are likely to be quite different. If you believe falsely that home is straight in front of you, you will not get there by going forward. Nor are the long term differences confined to your broad states when you get home; even your internal states are likely to be different if you do not get home (it may be a matter of life and death).

Of course, internalists can appeal to the composite state (as they may see it) of believing truly that home is straight in front of you, the supposedly narrow conjunct being the state of believing that home is straight in front of you and the environmental conjunct the state of home's being straight in front of you. But even if you believe truly that home is straight in front of you without knowing that it is, for example because that true belief is based on a false belief about where you are, and home is many miles further on through the forest than you think, then you are liable subsequently to discover the falsity of the belief on which your true belief is based and abandon the latter.[8] Thus the long term effects even of believing truly without knowing are not generally the same as the long term effects of knowing. If we are interested in whether the agent actually gets home, the state of knowing which direction home is in has a significance that cannot be subsumed under the significance of having a belief as to which direction home is in, or even of having a true such belief. Of course, if the agent maintains a true belief as to which direction home is in all the way home, that trivially suffices to get home, but appealing to the maintenance of that true belief throughout the journey fails to address the original challenge, which was to understand the long term effects of the agent's cognitive state at a fixed time *t*. Whether the agent's true beliefs at *t* are likely to be maintained

after *t* is part of what we are trying to understand in terms of the agent's cognitive state at *t*; merely assuming that they are maintained does not help us answer that question.

The externalist can produce a more general explanation by invoking the state of knowing which direction home is in, as opposed to the more specific state of knowing that home is straight ahead. It would not help the internalist to match that generality by invoking the state of having a true belief as to which direction home is in, as opposed to the more specific state of believing truly that home is straight ahead. For the state of having a true belief as to which direction home is in is itself prime, as one can show by using scenarios 1 and 2 above. Invoking that prime state does not yield an explanation that reduces to one that just invokes a narrow state and an environmental state. As usual, the explanatory work is done by the prime state itself. Moreover, we have already seen that the substitution of true belief for knowledge involves explanatory loss.

Similar considerations apply to the state of consciously thinking about Rover. The short term effects of consciously thinking 'that dog on my right' may be the same whether doing so constitutes consciously thinking about Rover, about Mover or about nothing at all (in case of hallucination). For example, you may move to the right. That may differ from what you do if you realize the state of consciously thinking about Rover (in scenario 2*) by thinking 'that dog on my left'. Thus what internalists conceive as the narrow state of consciously thinking with that visual mode of presentation may be more relevant to short term effects than is the broad state of consciously thinking about Rover. But the long term effects of consciously thinking with that visual mode of presentation may depend critically on whether doing so constitutes thinking about Rover. In one case, the effect may be that you buy Rover; in the other, that you buy Mover, a dog of similar

visual appearance but totally different personality, or discover that you were hallucinating. From then onwards, further developments are liable to diverge increasingly, even with respect to your narrow states. Similarly, the long term effects of consciously thinking about Rover may be the same whether you think of him as 'that dog on my right' or 'that dog on my left'. You may be more likely to buy him either way.

Obviously, the long term effects under discussion are highly sensitive to the agent's other mental states, such as desires. Nevertheless, a pattern emerges. The more short term the effects we consider, the greater the explanatory relevance of narrow states. As we consider longer term effects, narrow states tend to lose that explanatory advantage and the intentional states of folk psychology come into their own. Those states are typically not just broad but prime. We have already seen that if we ignore long term effects we fail to capture significant general patterns in cognition. Thus prime cognitive states are no mere curiosity of folk psychology: they are central to the understanding of long term cognitive effects.

When we investigate cognitive effects that depend on continuous feedback from a complex environment, we cannot expect to lay down strict laws. We can hope to identify probabilistic tendencies, perhaps comparable to those of evolutionary biology at species level. For example, we might want to investigate the general cognitive effects of literacy (as a means of public communication and not just of private note-taking). In doing so, we seek generalisations that hold across different languages and scripts. But literacy is itself a prime cognitive state, for it involves a kind of match between the individual's dispositions to produce and respond to written marks and those prevalent in the appropriate social environment. The narrow states that go with mastery of written

communication in English-speaking countries today did not go with mastery of written communication in Babylon three thousand years ago, and vice versa. Knowing how to read is not a narrow or even composite state; it is prime and broad.[9] Perhaps recent developments in cognitive science in the study of embodied, situated and distributed cognition, particularly cognition that relies on continuous feedback loops into the external environment, can be interpreted as investigations of prime cognitive states.[10]

4. We have seen that the two main sources of broadness in the intentional states of folk psychology ─ factive attitudes and environmentally determined contents ─ are also sources of primeness, in ways that make those states especially fit to explain long term effects at an appropriate level of generality.[11] In particular, such states as knowing, seeing, remembering and referring play key roles in those explanations. The appeal to such states in understanding cognition raises a general question about the nature of the theoretical enterprise. For words like 'know', 'see', 'remember' and 'refer' are *success terms*. They describe what happens when cognition goes well. Even if we replace the ordinary language terms by more theoretical ones for the purposes of cognitive science, the argument of previous sections suggests that some of those more theoretical terms will need to have a relevantly similar character. To give success terms a central role in our theorising about cognition is to understand it in relation to its successes. That is not to ignore the failures; rather, we understand them *as* failures, deviations from success. For example, to a first approximation, we can treat merely believing that P as merely being in a state with the content that P that the agent cannot distinguish from knowing that P, having it merely visually appear to one that P as merely being in a state with the content

that P that one cannot distinguish from seeing that P, and misremembering as merely being in a state with the content that P that one cannot distinguish from remembering that P.[12] Again, we might understand cases of reference failure as cases merely indistinguishable by the agent from cases of reference. That is to employ a sort of teleological strategy for understanding cognition. The internalist follows the opposite strategy, starting from states that are neutral between success and failure, and then trying to distinguish the two classes by adding environmental conditions.

The externalist's point is emphatically not to deny that the successes and the failures have something in common. Indeed, the failures were all described as merely indistinguishable by the agent from successes; since everything is indistinguishable from itself, it follows that both successes and failures are indistinguishable by the agent from successes. The point is rather that the failures differ internally amongst themselves, and that what unifies them into a theoretically useful category with the successes is only their relation to those successes. For example, many different total internal states are compatible with its perceptually appearing to an agent that there is food ahead; what they have in common, on this view, is their relation to cases in which the agent perceives that there is food ahead ('perceive' is factive).

To use a traditional analogy, consider the relation between real and counterfeit money. Uncontroversially, a counterfeit banknote can in principle be an exact internal duplicate of a real banknote. The internalist strategy corresponds to taking as theoretically fundamental not the category of (real) money but the category that contains both real money and all internal duplicates of it.[13] One would then have to circumscribe the real money by further constraints (presumably concerning origin and economic role).

But that strategy seems quite perverse, for being real money cannot usefully be analysed as having a certain intrinsic property and in addition satisfying some further constraints: those further constraints do all the theoretical work. Indeed, the property of being money is prime, in the sense that by a criterion like PRIME it is not the conjunction of a purely intrinsic property and a purely extrinsic one, for being money is compatible with being gold, and it is compatible with being in a social environment in which only silver counts as money, but it is incompatible with the conjunction of those two properties. It is no use complaining that real money and its internal duplicates have the same causal powers. For purposes of economic theory, the category of real money is primary, the category of counterfeit money must be understood as parasitic on it, and the category of all internal duplicates of real money is of no interest whatsoever.[14]

Of course, the analogy is not decisive. But it does show that the factorising strategy is not always compelling or even attractive. Each application of it must be argued on its individual merits. The argument of this paper has been that, for the study of cognition, the factorising strategy is often inappropriate. Sometimes, we need to use concepts like *knowledge* and *reference* (or *money*) in our explanations, and we cannot replace them without loss of valuable generality by conjunctions of purely internal and purely external constituents.

Notes

1 On some alternative definitions, a state S is ‘narrow’ if and only if S has no existential implications outside the subject of S (compare Putnam’s influential definition of ‘methodological solipsism’ as ‘the assumption that no psychological state presupposes the existence of any individual other than the subject to whom that state is ascribed’ in his 1975: 220). Even after various ambiguities are cleared up, such definitions are awkward because they deprive the class of ‘narrow’ states of closure properties required by the conception of those states as forming a self-enclosed domain. In particular, they permit the conjunction of two ‘narrow’ states not to be ‘narrow’. For let N1 and N2 be two incompatible ‘narrow’ states, and B a non-‘narrow’ state. Then the disjunctive states $N1 \vee B$ and $N2 \vee B$ are also ‘narrow’ by such definitions, since N1 implies any existential implication of $N1 \vee B$ and N2 implies any existential implication of $N2 \vee B$. But $(N1 \vee B)\&(N2 \vee B)$ is not ‘narrow’, for it is equivalent to B. By contrast, definitions such as that in the text on which the narrow is whatever supervenes on or is determined by the internal automatically make conjunctions of narrow states narrow (although the notion of the internal may be problematic for some dualists).

2 In the same spirit, the text treats both narrow states and environmental states as synchronic. That is a significant over-simplification. Whether one is in a given folk psychological intentional state is typically sensitive to causal origins. Even though reference cannot be defined in causal terms, reference to particulars and kinds in the environment is still normally carried by causal connections to those particulars and kinds

through memory and perception. Similarly, even though knowledge cannot be defined in causal terms, the difference between knowing and merely believing is often partly constituted by the presence or absence of an appropriate causal connection. In applying the internal-external distinction, we must therefore decide how to classify the agent's past internal history. Since the rationale for drawing the distinction depends on the assumed locality of causation, which is both spatial and temporal, the natural ruling is that the past history of both the agent and the environment counts as external for the purposes of distinguishing broad from narrow. This makes the conception of the internal and the external as independent dimensions harder to maintain, since the external includes all the causal antecedents of the internal. Fortunately, that complication does not undermine the conclusions in the text.

3 Williamson 2000 develops the argument of this chapter in greater detail with reference to epistemology, and responds to some internalist challenges.

4 Compare Putnam 1975, where the argument (formulated rather differently) is directed only against internalism about linguistic meaning, and takes internalism about psychological states for granted. Burge 1979 made the natural generalization to psychological states, which Putnam later accepted. See Pessin 1996 for more on the debate.

5 The gloss 'necessary and sufficient for the action to be taken' is indeed very rough, for we do not expect an explanation of an outcome to generalise to cases in which

the same outcome occurred for completely different reasons. Satisfying explanations have a certain unity and naturalness; they do not rope together essentially disparate cases. Nevertheless, subject to this vague constraint, the point stands that generality is an explanatory virtue. For more on the importance of generality in causally relevant properties and the trade-off between generality and naturalness see Yablo 1992, 1997 and 200?.

6 More generally, if there are m possible maximally specific narrow states and n possible maximally specific environmental states, and the possible maximally specific states of the world correspond to all pairs of the former and the latter, then there are 2^m narrow states, 2^n environmental states and 2^{mn} states of the word altogether. For simplicity, the calculation includes both the universal state (which always obtains) and the null state (which never obtains); they are the only states that are both narrow and environmental. The null state is composite; every non-null composite state corresponds to a unique combination of a non-null narrow state and a non-null environmental state, so there are $1 + (2^m-1)(2^n-1)$ composite states altogether. In the toy model in the text, $m = n = 3$.

7 To establish PRIME, we must assume the principle of *Free Recombination*, according to which any possible narrow state is compatible with any possible environmental state. Although this principle is not beyond question, it is a natural one for the internalist to assume, since it reflects the internalist analysis of the state of the world into two logically independent dimensions, the internal state of the agent and the external

state of the environment. Thus it is not unfair to the internalist to assume Free Recombination. Moreover, it is independently plausible that Free Recombination holds at least to a first approximation. We can now argue for PRIME as follows. First, suppose (for reductio) that some narrow state I and some environmental state E are both separately compatible with S but I&E is not, yet S is not prime. Thus for some narrow state I* and environmental state E*, S is equivalent to I*&E*. Hence I is compatible with I*&E*, so I&I* is a possible narrow state. Similarly, since E is compatible with I*&E*, E&E* is a possible environmental state. Therefore, by Free Recombination, I&I* is compatible with E&E*, so I&E is compatible with I*&E*, and so with S, contrary to hypothesis. This establishes the sufficiency of the condition in PRIME for primeness. For its necessity, suppose that for every narrow state I and environmental E, if I and E are separately compatible with S then so too is I&E. Consider all conjunctions of the form I&E that imply S, where I and E are possible maximally specific narrow and environmental states respectively. Let I* be the (infinite) disjunction of all the first conjuncts of such conjunctions and E* the (infinite) disjunction of all the second conjuncts. Thus I* is narrow and E* environmental. Moreover, S implies I*&E*, for if S obtains then so does some conjunction I&E that implies S, where I and E are possible maximally specific narrow and environmental states respectively; but then I implies I* and E implies E*, so I&E implies I*&E*, so I*&E* also obtains. Conversely, I*&E* implies S, for if I*&E* obtains then for some possible maximally specific narrow states I and I** and environmental states E and E**, I&E also obtains and both I&E** and I**&E entail S; both I&E** and I**&E are possible by Free Recombination, so I and E are both compatible with S; therefore, by hypothesis, I&E is compatible with S; since I

and E are maximally specific, I&E entails S, so S obtains. Thus S is equivalent to I*&E* and so is not prime.

8 Gettier 1963 has classic examples of true beliefs that fail to constitute knowledge because they are essentially based on false premises.

9 See Stanley and Williamson 2001 for a general argument that knowing how is a species of knowing that. If so, the discussion of propositional knowledge applies to knowledge how as a topic for cognitive science.

10 See Clark 1997, Gigerenzer 1999 and Hurley 1998. For example, 'smart' as used in Gigerenzer's title presumably refers to a prime state, one that depends on the appropriateness of the agent's simple heuristics to the nature of the environment.

11 Factiveness and reference need not be independent sources of broadness, for knowledge-based constraints may play a constitutive role in the determination of reference (Williamson 2004).

12 This idea is arguably the core of the so-called Disjunctive Theory of Perception. For a recent discussion see Martin 2004. Other papers in the same volume are also relevant.

13 The internalist category contains much beyond real and counterfeit money: for instance, if some distant society uses internal duplicates of my paperclips as money, then my paperclips themselves fall into the category, although they are neither real nor counterfeit money.

14 It is not even safe to assume that counterfeit money that was an exact internal duplicate of real money would be undetectable; it might be detected by its extrinsic properties, such as location.

References

Burge, T. 1979. 'Individualism and the mental'. *Midwest Studies in Philosophy*, 4: 73-121.

Clark, A. 1997. *Being There: Putting Brain, Body, and World Together Again*. Cambridge, Mass.: MIT Press.

Evans, G. 1982. *The Varieties of Reference*. Oxford: Clarendon Press.

Gettier, E. 1963. 'Is justified true belief knowledge?'. *Analysis*, 23: 121-3.

Gigerenzer, G., Todd, P., and the ABC Research Group. 1999. *Simple Heuristics that Make Us Smart*. Oxford: Oxford University Press.

Hurley, S. 1998. *Consciousness in Action*. Cambridge, Mass.: Harvard University Press.

Martin, M.G.F. 2004. 'The limits of self-awareness'. *Philosophical Studies*, 120: 37-89.

McDowell, J. 1977. 'On the sense and reference of a proper name'. *Mind*, 86: 159-85.

Pessin, A., Goldberg, S., and Putnam, H. 1996. *The Twin Earth Chronicles: Twenty Years of Reflection on Hilary Putnam's "The Meaning of 'Meaning'"*. Armonk, NY: M.E. Sharpe.

Putnam, H. 1975. 'The meaning of "meaning"', in *Mind, Language and Reality*, Cambridge: Cambridge University Press.

Putnam, H. 1978. *Meaning and the Moral Sciences*. London: Routledge & Kegan Paul.

Stanley, J., and Williamson, T. 2001. 'Knowing how'. *Journal of Philosophy*, 98: 411-44.

Williamson, T. 2000. *Knowledge and its Limits*. Oxford: Oxford University Press.

Williamson, T. 2004. 'Philosophical "intuitions" and scepticism about judgement'. *Dialectica*, 58: 000-000.

Yablo, S. 1992. ‘Mental causation’. *Philosophical Review*, 101: 245-80.

Yablo, S. 1997. ‘Wide causation’. *Philosophical Perspectives*, 11: 251-81.

Yablo, S. 200?. ‘Prime causation’. *Philosophy and Phenomenological Research*, forthcoming.