We have seen that folk psychological explanations appear to float free of any appeal to the underlying cognitive states and processes that sustain our capacities for perception, action and language. And yet a person's thoughts, wants and wishes are not entirely independent of the perceptual, cognitive and affective states they are in. So there is scope for both for a re-examination of the way philosophers characterize folk psychology and of what people really do appeal to or depend on in making ascription of mental states to others. Additionally, we need not contemplate seriously the eliminativist option in trying to reconcile the lived experience of our inner lives with the findings of neurobiology and neuroscience, for these disciplines too must make sense of the experience, thoughts and reflections of subjects at the personal level. They may cast light on why our experience has the form and character it is, and what happens when the underlying mechanisms break down, but it cannot dispense with the level at which fixes on the mental states it is interested in explaining. Thus a non-reductive cognitive neuroscience and a non-reduced but rich and detailed folk psychology must eventually be aligned.

7 Internalism and Externalism in Mind

Sarah Sawyer

Internalism and Externalism: The Basics

The individuation conditions of psychological properties is the topic of this chapter.[1] There are two opposing views: *internalism* and *externalism*. According to the former – also known as *individualism* – psychological properties are individualistically individuated, which is to say that their instantiation by an individual depends entirely on the individual's intrinsic physical make-up. According to the latter – also known as *anti-individualism* – psychological properties are anti-individualistically individuated, which is to say that their instantiation by an individual depends not only on the individual's intrinsic physical make-up, but in addition on objective relations she bears to objective properties in her environment. If a psychological property is individualistic, then its associated content is said to be *narrow*; if it is anti-individualistic, then its associated content is said to be *broad*.

Internalism, then, is the view that psychological properties supervene locally on physical properties: no two individuals could differ psychologically without differing in some intrinsic physical respect. Externalism rejects this local supervenience thesis, maintaining in contrast that individuals could be exactly alike with respect to their intrinsic physical properties and yet differ psychologically – if, for instance, they were related to relevantly different environments. Both internalism and externalism are consistent with global psycho-physical supervenience, the claim that no two worlds could differ psychologically without differing physically. Local supervenience entails global supervenience (since worlds can be construed as individuals), but not vice versa. The local supervenience thesis is, therefore, the stronger claim, and the question of its truth lies at the heart of the internalism/externalism debate.[2]

This paper provides an overview of the prevailing issues concerning the debate. In the second section I distinguish various kinds of externalism and outline some considerations in their favour. In the third section I discuss various forms of internalism. In the fourth section I deal with metaphysical considerations concerning naturalism and mental causation that have motivated internalism

and been thought to tell against externalism. In the fifth section I deal with epistemological considerations concerning the direct, non-empirical, authoritative nature of self-knowledge that have been thought to tell against externalism. I then conclude briefly in the sixth section.

Kinds of Externalism

Different considerations are thought to be relevant to the individuation conditions of different kinds of psychological property. Consequently, one might embrace externalism for certain kinds of psychological property but internalism for others. In this section I introduce a number of considerations that favour externalism and catalogue various resulting kinds of externalism. The kinds of externalism fall into two broad camps: externalism about concepts expressed by predicative terms, which I will call predicative externalism, and externalism about concepts expressed by singular terms, which I will call singular externalism.

Predicative externalism

The most widely recognised consideration in favour of externalism emerges from reflection on counterfactual scenarios in which a subject's intrinsic physical make-up is hypothesized to remain constant while the broader physical environment in which she is embedded is hypothesized to differ.[3] Such an environmental difference, it is urged, would be responsible for a difference in the subject's psychological states precisely because non-intentional causal relations to objective properties in one's environment partly determine what one can represent in thought. For example, a subject S, related in the right kind of non-intentional way to silver, might have various thoughts involving the concept *silver*, such as that silver jewellery is cheaper than gold jewellery. She may be unable to distinguish (either practically or theoretically) various other actual or possible metals from silver and may well acknowledge this. Nevertheless, she possesses the concept *silver* because she is related in the right kind of way to silver, and hence can think various things about silver by means of that concept. Now suppose S had lived in different circumstances, circumstances in which there was no silver for her to be related to either directly (via perception) or indirectly (via other people). In such a situation S would be unable to think about silver as such because there would be nothing to ground her possession of the concept *silver*. How could she have acquired the concept? Suppose instead that she had been related to one of the actual or possible metals that she is unable to distinguish from silver. Call this metal 'twilver'. In such circumstances,

S would have been related to twilver in just the same way as she is actually related to silver, and hence it is plausibly the concept *twilver* that S would have acquired. Consequently, where S thinks that silver jewellery is cheaper than gold jewellery, counterfactual S thinks instead that twilver jewellery is cheaper than gold jewellery. The difference in representational content between the belief S has and the belief S would have lies in the difference between the objective properties to which she is related (silver) and would be related (twilver) respectively. If these considerations are persuasive, then what determines the representational content of a subject's beliefs goes beyond her intrinsic physical make-up and her discriminative capacities (which are hypothesized to be identical in the actual and the counterfactual scenarios alike) and depends in addition on the objective properties to which she is related.[4]

This kind of thought experiment is taken by many to establish externalism specifically with respect to natural kind concepts: concepts that 'carve nature at its joints' and feature in the true final set of scientific theories: concepts such as (perhaps) *quark, electron, hydrogen, water, heart, tiger, planet*.[5] However, reflection on two further kinds of counterfactual scenario favours a more general externalism. The first draws upon the possibility of incomplete linguistic understanding[6]; the second draws upon the possibility of non-standard theory[7]. I outline each in turn.

First suppose that a subject S has a wide range of ordinary beliefs attributable by means of the term 'game': she believes that some games are more fun than others, that chess is a game, that children like party games, and so on. However, she believes in addition (and mistakenly) that games must involve at least two people, a point she would readily accept correction on if her mistake were pointed out to her. Next consider a counterfactual scenario in which her intrinsic physical make-up is hypothesized to remain constant while her linguistic community is hypothesized to differ. In the counterfactual scenario the term 'game' is defined and standardly used to apply to games that involve at least two people. Since 'game' and 'game involving at least two people' mean different things, the word-form 'game' in the counterfactual scenario has a different meaning than it does in the actual situation. In like fashion, the concept expressed by the word-form differs in the actual and the counterfactual situations. In the actual situation the word-form 'game' expresses the concept *game* and includes in its extension games such as solitaire and patience. In the counterfactual situation, in contrast, the word-form 'game' expresses a different concept that does not include in its extension either solitaire or patience. Consequently, S may in fact believe that pass the parcel is a game, but had she been a member of the counterfactual linguistic community she would have possessed a distinct concept, believing instead that pass the parcel is a 'shgame', say. Once again, S's intrinsic physical make-up and classificatory capacities are identical in the actual and the counterfactual situations,

and the difference in representational content between the belief she has and the belief she would have lies beyond her intrinsic physical properties, this time anchored by the classificatory practices of the wider linguistic community of which she is and takes herself to be a part.

Behind this counterfactual scenario lies a certain understanding of linguistic meaning according to which the conventional linguistic meaning of a term (roughly its dictionary definition) is a complex abstraction from communal rather than individual use. Linguistic meaning is determined by actual and possible agreement among the most competent users, where the most competent users are those to whom others do and would defer if a question about an individual's use were to arise. On this view, understanding the meaning of a word is not an all-or-nothing thing, but rather comes in degrees. And it is the possibility of understanding a word incompletely that allows for the difference in linguistic meaning in the actual and the counterfactual situations to be consistent with there being no difference in intrinsic physical make-up between actual and counterfactual S. The difference in linguistic meaning is then taken to imply a difference in concept expressed.

The final consideration that favours a general externalism trades on the fact that even a subject with a full understanding of the linguistic meaning of a term can doubt whether the dictionary definition that reflects that meaning correctly characterizes the things referred to by that term. Thus suppose a subject S has a full understanding of the term 'sofa' and yet comes to wonder whether sofas are really religious artefacts and not pieces of furniture made for sitting. Her proposed theory about sofas is false, but this need not compromise either her full understanding of the term 'sofa' or her ability to think with the concept *sofa*; rather, it reflects a strange view about the nature of sofas thought of as such. Now hypothesize a counterfactual situation in which S's false theory is standard and true of a different yet superficially indistinguishable class of entities (call them 'safos').[8] The linguistic meaning of the term 'sofa' in the actual situation differs from the linguistic meaning of the term 'sofa' in the counterfactual situation even though the entities referred to are superficially indistinguishable. This is because the actual linguistic community and the counterfactual linguistic community have agreed upon different characterizations of the relevant entities. Moreover, the concept expressed by the term differs in the two situations because the entities referred to differ: in the actual situation they are sofas (pieces of furniture made for sitting), whereas in the counterfactual situation they are safos (religiδus artefacts). Consequently, while actual S believes that sofas are religious artefacts, counterfactual S believes that *safos* are religious artefacts.

Behind this counterfactual scenario is a certain understanding of the difference between the linguistic meaning of a term and the concept expressed by that term. The linguistic meaning of a term goes beyond individual use and is

grounded in communal use, as mentioned above. Communal use may well change over time, and hence the linguistic meaning of a term may well change over time. (Dictionaries are plausibly updated in part to reflect such changes in linguistic meaning.) But the concept expressed by a term may well remain unaltered even while the linguistic meaning of that term changes. This will happen, for instance, when entities of a given kind are identified through perception and then characterized. The concept will be anchored to the entities through perception, whereas the linguistic meaning will reflect received views about the entities, and this characterization may well need updating as investigation proceeds and even while the concept remains unchanged. It is the fact that we can be mistaken in our characterizations of the things we perceive that allows for non-standard theory to be entertained, and this in turn grounds a general form of externalism. In the sofa/safo case, of course, S's theory would prove false under empirical tests and hence would not lead to a change in linguistic meaning. Cases where proposed theories are adopted, however, would lead to corresponding changes in linguistic meaning. This is what allows us to make sense of genuine theoretical disagreement about a class of entities thought about by means of the same concept, and grounds constancy of reference through theory change.

Thus far I have distinguished two broad kinds of externalism: natural kind externalism, based on noting subjects' relations to natural kinds; and social externalism, based on the possibility of incomplete linguistic understanding and the possibility of theoretical doubt. Both are kinds of what I have called 'predicative externalism' since they concern concepts expressed by predicative terms. Natural kind externalism has gained more support than social externalism, but so long as we take seriously, as we must, the thought that our concepts concern a world about which we can be in error, there is reason to adopt a general predicative externalism.[9]

Singular externalism

According to singular externalism, the representational content of a subject's thoughts about particulars (singular thoughts) is individuated partly by the particulars those thoughts concern. This is directly analogous to predicative externalism according to which the representational content of a subject's thoughts about properties is individuated partly by the properties those thoughts concern.

There are two main kinds of singular externalism: externalism about thoughts expressed by sentences containing demonstratives; and externalism about thoughts expressed by sentences containing proper names. To take a demonstrative example first, suppose that actual S is looking at a particular apple, A1,

while counterfactual S is looking at a different apple, A2. Suppose further that S and counterfactual S utter the sentence 'That is nutritious'. It is clear that S's utterance (and thought) concerns A1, whereas counterfactual S's utterance (and thought) concerns A2. This is so even if S's intrinsic physical make-up is identical in the two situations. Moreover, S's utterance (and thought) is true if and only if A1 is nutritious, while counterfactual S's utterance (and thought) is true if and only if A2 is nutritious. Crucially, according to singular externalism this difference in truth conditions is due to a difference in representational content.

Parallel remarks hold for externalism concerning thoughts expressed by sentences containing proper names. Thus if S utters the sentence 'Danny is interesting', referring to Danny Alpha, with whom she is acquainted, and counterfactual S utters 'Danny is interesting', referring to Danny Beta, with whom she is acquainted, their utterances and thoughts have different truth conditions, and this is consistent with S's intrinsic physical make-up being the same in both the actual and the counterfactual situations. Again, according to singular externalism this difference in truth conditions is due to a difference in representational content.

Singular externalism is upheld by a number of people in a number of different ways. According to Gareth Evans and John McDowell, all thoughts are composed of Fregean senses, but singular thoughts contain de re senses which exist only if there is an object to which they refer. Evans and McDowell advocate this kind of singular externalism for all thoughts about particulars, whether the particulars are thought about by means of demonstratives or by means of proper names.[10] According to direct reference theorists, in contrast, the thought expressed by a sentence containing a proper name contains not a de re sense of the object named but the very object itself.[11] Here again, the existence of the thought depends upon the existence of the object thought about. This view is typically not extended to demonstrative thought, although in principle it could be. A variant of the direct reference theory that accommodates Fregean insights (about different ways of thinking about an object) without countenancing de re senses holds that the thought expressed by a sentence containing a proper name contains the object named together with a mode of presentation of that object, but the implication is the same: the existence of the thought depends upon the existence of the object thought about. This view could also in principle be extended to the demonstrative case. What makes all these views forms of singular externalism is the common claim that the *content* of a singular thought is object-dependent.

Considerations that bear on singular externalism thus far parallel considerations that bear on predicative externalism, as noted at the outset. But the question of the individuation conditions of singular thoughts introduces the

possibility of a distinction between a singular thought and its representational content; and this distinction has no analogue in the predicative case. The distinction opens up the possibility of accepting that the truth conditions of singular thoughts are object-dependent while denying that this is in virtue of a difference in representational content. What results is a theory according to which the representational content of a singular thought is preserved across intrinsic physical duplicates but can be thought of (and hence true or false of) different individuals on different occasions. To take the demonstrative example above, on this view, S and counterfactual S both have a thought the representational content of which is given by the open sentence 'is nutritious'. Actual S thinks this of A1, whereas counterfactual S thinks this of A2. The difference in truth conditions between the thoughts is on this view due to a difference in contextual application rather than representational content. To accept such a distinction between a thought and its content is to embrace a kind of two-factor theory of singular thought according to which the object thought about is a constituent of the thought but is not referred to by a conceptual constituent of the thought. On such a view the object contributes to the truth conditions of a thought concerning it but does not affect its representational content. Hence the view is a form of singular internalism.[12]

This view has been popular in the demonstrative case, but has gained little support in the proper name case due to the dominance of direct reference theories. However, if one were to accept singular internalism for the demonstrative case and in addition think of singular uses of proper names as involving a demonstrative element, then one would naturally be led to embrace singular internalism for the proper name case too. To take the second example above, a singular use of a name such as 'Danny' is to be understood as involving a demonstrative element and hence as semantically equivalent to 'That Danny', which can be used to refer to different Dannys on different occasions. On this view, S and counterfactual S both have a thought the representational content of which is given by an open sentence something like 'is a Danny and is interesting'. Actual S thinks this of Danny Alpha, whereas counterfactual S thinks this of Danny Beta. The difference in truth conditions between the thoughts is again due to a difference in contextual application rather than representational content.[13]

Kinds of Internalism

I have already discussed singular internalism above to contrast and clarify singular externalism. Consequently I will confine my discussion in this section to versions of predicative internalism, of which there are four primary forms.

Two kinds of thorough-going internalism

The most straightforward way of being a predicative internalist is to reject outright the interpretation of the counterfactual scenarios taken above to ground externalism. An alternative, internalist interpretation would maintain instead that psychological properties are necessarily preserved across intrinsic physical duplicates precisely because they are and must be grounded in the discriminative capacities and transparent epistemic outlook of the individual.[14] An individual's psychological make-up cannot outstrip what that individual can do and how things seem to her, as it were. S and counterfactual S in each of the scenarios have the same capacities to discriminate and classify things, and (in some sense) have the same views about the things they encounter: there is nothing that allows them to distinguish the actual from the counterfactual situation in each case. Consequently, according to thorough-going internalism, there can be no psychological difference between them.

One way of upholding the view is to think of the relevant concepts as *descriptive*, encapsulating the subject's beliefs (or theories) about the things referred to. For instance, the concept both S and counterfactual S express by the term 'silver' might be *shiny metal often used to make jewellery and that needs to be polished to be kept clean and* . . . The concept they express by the term 'game' might be *kind of activity undertaken for enjoyment, involving at least two people, involving rules in accordance with which you can win or lose*; and the concept they express by the term 'sofa' might be *religious artefacts that look as if they may be sat upon but* . . . Note that if the original concepts (here thought of as descriptive) are to be individualistically individuated, then the concepts used in the descriptions must of course *themselves* be individualistically individuated. But this kind of descriptivism (which many will view as independently problematic) is not essential to the view. One could instead treat the concepts minimally.[15] On this view, the concept both S and counterfactual S express by the term 'silver' is a concept that has in its extension silver, twilver and everything else that S cannot distinguish from them (as it does on the descriptive view). But in order to express the concept we would need to introduce a new term such as 'shmilver'. Similarly, the concept they express by the term 'game' is the concept *shgame*, which has in its extension games that involve at least two people; and the concept they express by the term 'sofa' is the concept *safo*, which has in its extension religious artefacts that look like sofas. The subject's discriminative capacities and epistemic outlook here serve to individuate her concepts and hence determine the extensions of those concepts but are not taken up as descriptive elements of the concepts themselves. On both the descriptive and the non-descriptive versions of through-going internalism new terms need to be introduced into our language in order to express with accuracy the concepts had by individuals whose beliefs differ from the norm (or, more generally, from our own),

as illustrated by the use of new terms in the examples just given. And on both views, S possesses the same concepts as her counterfactual self in virtue of having the same discriminative capacities and epistemic outlook on the world, but she has different concepts from those in her linguistic community. This stands in marked contrast to predicative externalism, according to which S has different concepts from her counterfactual self but shares many concepts with those in her linguistic community despite varying degrees of understanding and competence which result in a wide variety of discriminative capacities and epistemic outlooks across individuals within that community.

Two kinds of two-factor internalism

The third and fourth kinds of predicative internalism are more complicated. They acknowledge that the counterfactual scenarios outlined establish that S and counterfactual S have different thoughts in some sense, but aim nonetheless to retain a sense of content which is preserved across intrinsic physical duplicates, in order to respect the internalist conception of sameness of epistemic outlook. Both therefore maintain that a thought has a narrow *and* a broad content and are thus kinds of two-factor theory.

According to the first of these, the internal component of a thought (its narrow content) is a function that determines its external component (its broad, truth-conditional content) given a context (an environment).[16] Thus when S and counterfactual S utter the sentence 'Silver jewellery is cheaper than gold jewellery', the narrow content of their thoughts is the same, but the broad content of their thoughts differs simply in virtue of their location in different environments: S's thought concerns silver, and is true if and only if silver jewellery is cheaper than gold jewellery; whereas counterfactual S's thought concerns twilver, and is true if and only if twilver jewellery is cheaper than gold jewellery.

There are similarities between this two-factor theory of predicative thought and the two-factor theory of singular thought discussed in the Singular Externalism section above. On both views the only form of truth-conditional content is broad. And yet on both views the thoughts of intrinsic physical duplicates share a kind of content even though they have different truth conditions. However, the similarity does not extend beyond the superficial level, and the differences are important. According to the two-factor theory of singular thought, singular thoughts have contents that are intrinsically representational independent of context, and can be applied to (or thought of) different individuals in different circumstances. The two factors involved in a singular thought are first, a content ('is nutritious', say), and second, (potentially) an individual of whom the content is thought (a particular apple, for instance).

The content of a singular thought is not itself divided into a narrow and a broad component. According to the two-factor theory of predicative thought now under consideration, in contrast, the content of a predicative thought is itself divided into a narrow component and a broad component. Crucially, the narrow component is not representational: only the broad component is. The narrow component is a function and can be understood only in terms of its inputs and outputs: that is, only in terms of the broad, truth-conditional content it produces once the individual is situated in a particular environment. This puts pressure on the idea that the narrow component of a predicative thought is properly conceived as a form of content at all.

The second kind of two-factor theory of predicative thought is also attracted both by the externalist interpretation of the counterfactual scenarios and by the internalist conception of sameness of epistemic outlook. However, it aims to draw a distinction between broad and narrow content consistent with all content being representational in some sense. On this view, the broad content of a subject's thought is determined, in line with externalist considerations, in part by relations she bears to objective properties in her environment. The narrow content of a thought, on the other hand, is individuated by the epistemic possibilities it allows and excludes.[17] The underlying thought here is that intrinsic physical duplicates are in the same epistemic position in the sense that they cannot distinguish between the relevant actual and counterfactual situations and that the narrow content of a thought encapsulates this fact. For example, suppose that S and counterfactual S both utter the sentence 'Silver jewellery is cheaper than gold jewellery'. The thoughts they thereby express have different broad, truth-conditional contents: one concerns silver whereas the other concerns twilver. However, the thoughts are taken to have the same narrow content because a purely qualitative description of a situation in which silver jewellery is cheaper than gold jewellery is identical to a purely qualitative description of a situation in which twilver jewellery is cheaper than gold jewellery. Given the epistemic position of S and counterfactual S, both situations 'verify' their thoughts and hence they share a narrow content. The success of the position clearly depends on the possibility of describing situations in purely qualitative terms – terms not subject to externalist considerations. As such, the position depends upon the truth of a restricted rather than a general form of externalism. If all terms were subject to externalist considerations then there would be no terms available to feature in the qualitative descriptions required to ground this notion of narrow content. Moreover, there is a question about whether it makes sense to think of a thought as having two forms of representational content where only one of these is truth conditional.

I have discussed and argued against all four forms of internalism elsewhere and will not repeat the arguments here.[18] Instead I now turn to some of the primary metaphysical and epistemological considerations that surround the internalism/externalism debate.

Metaphysical Considerations

According to externalism, psychological properties do not supervene locally on a subject's intrinsic physical make-up. This throws up two related metaphysical concerns: first, how to retain a naturalistic theory of the mind; and second, how to make sense of mental causation. The two concerns are intimately connected and have provided much of the motivation for internalism.

Since the late 1950s and early 1960s, the question of how psychological properties relate to 'lower level' properties – and ultimately to properties of interest to the physical sciences – has dominated discussions in philosophy of mind. The requirement that they must be related in some intimate and significant way has been regarded as crucial to an account of the mind that is scientifically and naturalistically respectable. Type-physicalism, according to which psychological properties are identical to physical properties, clearly satisfies the requirement. However, the postulated identity of psychological properties with physical properties comes under pressure from arguments to the effect that psychological properties are multiply realizable – that individuals in different physical states could nonetheless be in the same mental state. Such arguments have motivated forms of token-physicalism, according to which each token mental state of an individual is identical to or realized by some physical state of that individual, even if the psychological property of which it is an instance is not identical to the physical property of which it is an instance. But the dispute between type-physicalism (of various kinds, including behaviourism) and token-physicalism (of various kinds, including functionalism) takes place within a common theoretical framework: physicalism.

Externalism, in contrast, rules out all forms of physicalism.[19] The minimal claim of physicalism is that every token psychological state of an individual is either identical to or realized by a token physical state of that individual.[20] More specifically, physicalism is defined by its commitment to local psychophysical supervenience. This makes clear why it is inconsistent with externalism. As such, externalism has been thought to sever the psychological from the physical, and hence to rule out a naturalistic theory of the mind. However, although externalism is inconsistent with physicalism, it is consistent with materialism – the view that every entity is composed of physical matter. This is a weaker doctrine than physicalism, but is strong enough to secure a naturalistically respectable theory of the mind. Since externalism is consistent with materialism, it is consistent with naturalism.

There is, however, a related worry about anti-individualistically individuated properties. In particular, it has been thought that only individualistic properties can be causal properties: that causal powers must be intrinsic. If this is right, then externalism is committed to the claim that psychological properties are not causal properties, which undermines the intuitive and commonly held idea that our beliefs and desires *cause* our actions.[21] The worry here is that although externalism is consistent with naturalism, its understanding of psychological properties renders them insignificant because psychological properties thus conceived would have no causal powers and hence make no difference in the world. Consequently, even if externalism is naturalistically respectable, it does not yield an account of the mind that is scientifically respectable.[22]

However, the assumption that causal properties must be intrinsic is misguided. Indeed, scientific practice demonstrates that many sciences study patterns of causation involving entities in their normal environment, and the properties to which they appeal in causal explanations are individuated in a way that presupposes such relations between entities and their environment.[23] Thus 'astronomy studies the motions of the planets; geology studies land masses on the surface of the Earth; physiology studies hearts or optic fibres in the environment of a larger organism; psychology studies activity involving intentional states in an environment about which those states carry information; the social sciences study patterns of activity among persons' (Burge, 1989, p. 317). Because such properties are individuated with reference to a normal environment, the properties are anti-individualistic; and yet such properties are also individuated with reference to their causal powers, and hence there is no question that they are causal properties.

The view that emerges is a view according to which psychological properties are causal and yet fail to supervene on lower level properties. It follows that individuals who are classified as of the same kind from the perspective of one science may be classified as of different kinds from the perspective of another. Thus for example two individuals may be exactly similar from the perspective of neuroscience but significantly different from the perspective of psychology because they instantiate the same neurophysiological properties but different psychological properties. This is the case for S and counterfactual S in each of the scenarios described in the Predicative Externalism section. This allows us to identify an error in the internalist's reasoning. Internalists often point out that S and counterfactual S would exhibit the same behaviour non-intentionally described (they would follow exactly similar trajectories through space, exhibit the same speech patterns, classify things in the same way, and so on), which of course is true, but they go on to conclude that S and counterfactual S instantiate the same psychological properties. However, the similarity in behaviour is to be explained by the fact that S and counterfactual S instantiate the same neurophysiological properties and is consistent with their instantiating

different psychological properties. The former may well be individualistically individuated even though the latter are not.

Externalism is inconsistent with physicalism. However, it is consistent both with a naturalistic theory of the mind and with the claim that psychological states are causally efficacious.[24]

Epistemological Considerations

Central to the internalism/externalism debate in the philosophy of mind has been the question of whether externalism is consistent with the intuitive claim that a subject knows what she is thinking in an epistemically privileged way. There are two primary areas of concern. The first is 'the achievement problem' and the second is 'the consequence problem'. I deal with each in turn.[25]

The achievement problem

According to externalism, what concepts we possess and hence what thoughts we can think depends on contingent, empirical relations we bear to objective properties in our environment. The question then arises, how can we know our thoughts in a direct, non-empirical, authoritative manner when those thoughts depend on our relations to the environment? Imagine that S is periodically switched from the actual situation (in which she is related to silver) to the counterfactual situation (in which she is related to twilver). Suppose further that after each switch she stays long enough to acquire the concept appropriate to the new environment. Under such a hypothesis S will at certain points in time think that silver jewellery is cheaper than gold jewellery, and at other points in time think that twilver jewellery is cheaper than gold jewellery. And yet there would be no break in the continuity of S's life because there would be (by hypothesis) no discernible difference between the two environments. The changes in her environment would pass undetected, and so, crucially, would the changes in her thoughts. On this basis it is argued that S does not know what she thinks in a direct, non-empirical, authoritative manner. Rather, S requires empirical knowledge of her environmental relations in order to know what she thinks. There are various ways one might take the argument: one might conclude that the mere possibility of such switches undermines the direct, non-empirical, authoritative nature of self-knowledge; or one might conclude that the close epistemic possibility of such switches undermines the direct, non-empirical, authoritative nature of self-knowledge; or one might conclude that only actual switching undermines the direct, non-empirical,

authoritative nature of self-knowledge.[26] However it is taken, externalists have been broadly uniform in their response, which has two main strands.

First, it is pointed out that the concepts available at the second-order level of thought (concepts employed to think about one's first-order thoughts) are determined (in part) by relations to the very same set of environmental conditions that determine the concepts available at the first order level of thought (concepts employed to think about the world). As such, S could not be in error about her thoughts simply in virtue of the dependence of those thoughts on her environmental relations. S could not, as it were, think she was thinking that silver jewellery is cheaper than gold jewellery but really be thinking that twilver jewellery is cheaper than gold jewellery. This kind of error would involve using the concept *twilver* at the first-order level while simultaneously using the concept *silver* at the second-order level. Rather, the same concept (whether it be *silver* or *twilver*) would be employed at all levels of thought. Consequently, the kind of threat envisaged is ill-conceived.[27]

Second, it is pointed out that the (partly environmental) conditions that individuate a thought are presupposed in the thinking of that thought but need not themselves be known in order for it to be known that that is the thought one is thinking. Perceptual knowledge presupposes that certain background conditions obtain (that lighting conditions are reasonable, that one is not hallucinating, and so on), but such background conditions need not be established by the subject before she can be said to know by looking that there is, for instance, an apple on the table in front of her. Similarly, it may be that particular instances of self-knowledge presuppose relations to objective properties in one's environment, but a subject need not know that such relations obtain in order to know what she thinks.[28] Indeed, if one had to know the individuating conditions of a thought in order to know one was thinking it, then neither internalism nor externalism would be consistent with the direct, non-empirical, authoritative nature of self-knowledge: we do not have such direct, non-empirical, authoritative knowledge about our environmental relations (as would be required if externalism were true); but we do not have such direct, non-empirical, authoritative knowledge about our intrinsic physical make-up either (as would be required if internalism were true). This merely shows that the demand for such knowledge of individuating conditions is irrelevant to questions about self-knowledge.[29]

But the achievement problem surfaces again in the guise of the argument from memory.[30] Suppose S thinks a second-order thought at t1: *I think silver is shiny*. Suppose she is then switched from the actual to the counterfactual environment where she remains long enough to acquire the concept *twilver*. At t2 (some point later), when reflecting on what she thought at t1, she will think with concepts relevant to the counterfactual environment and hence think, it is argued: *I thought twilver was shiny*. The content of her thought at t2 is

false, since it does not capture the content of her thought at t1. Consequently, S does not know at t2 what she was thinking at t1. This is taken to undermine self-knowledge of externally individuated past thoughts. Moreover, it is argued, self-knowledge of externally individuated current thoughts is also undermined. After all, if S does not know at t2 what she was thinking at t1, and there is no reason to think she has forgotten anything in the interim, there is reason to think she never knew at t1 what she was then thinking.

As with the initial argument, there are various ways the argument might be taken depending on whether one thinks mere possibility, close possibility or actuality the relevant epistemic factor. But here two different responses have emerged. According to the first, the argument shows that externalism does undermine the direct, non-empirical, authoritative nature of one's knowledge of one's past thoughts, but it does not show that externalism undermines the direct, non-empirical, authoritative nature of one's knowledge of one's *current* thoughts.[31] This can be made plausible, for instance, by acknowledging a new way in which one might be said to forget something (namely, by being switched between subjectively indistinguishable environments), or alternatively, by maintaining that forgetting is not the only way in which one might fail to know at t2 what one knew at t1 (since one might instead be switched between subjectively indistinguishable environments). The general moral here is that although one may need to rely on empirical considerations to the effect that the environment has remained broadly stable in order to know what one thought in the past, the non-empirical warrant for knowledge of one's current thoughts is not thereby undermined. The disruption, as it were, is confined to knowledge of past thoughts.

According to the second line of response, the argument does not show that externalism undermines the direct, non-empirical, authoritative nature of one's knowledge either of one's current or of one's past thoughts.[32] This second line of response is bolder and can be made plausible, for instance, by showing how the content of past thoughts can be preserved in memory even across undetectable switches between differing environments. This has been advocated by Burge, who distinguishes substantive event memory, which refers back to earlier events, from preservative memory, the function of which is to hold the contents of thought in place so the subject can determine logical and epistemic relations between them. Preservative memory does not refer back to earlier thinkings, but rather holds contents in place for the purposes of, for instance, critical reasoning.[33] While substantive event memory might be undermined by externalism, preservative memory will not be, precisely because preservative memories do not refer back to independent events.[34]

It is important to note that externalists have not offered a theory of self-knowledge in response to the achievement problem in either of its guises. Rather, they have tried to show how the arguments are misguided. Two things

are clear. First, there is no widely accepted theory of self-knowledge, either internalist or externalist, and work remains to be done here. But, second, the achievement problem does not bring to light any specific difficulties for the externalist. Rather, and perhaps unsurprisingly, externalist theories of self-knowledge and of memory will look rather different from internalist ones.

The consequence problem

The consequence problem emerges once an answer to the achievement problem has been assumed. The problem arises when one combines a non-empirical warrant for the claim that one is thinking a particular thought, with a non-empirical warrant for the claim that thoughts of that kind depend on the environment's being a particular way, to yield, surprisingly, a non-empirical warrant for a claim about the nature of one's environment. For example, S might reason as follows: (P1) I think silver is shiny; (P2) if I think silver is shiny then I must be related to silver; therefore (C) I am related to silver. (P1) is an instance of self-knowledge and hence taken to be non-empirically warranted; (P2) is arrived at through philosophical theorising and hence taken to be non-empirically warranted; but then it seems that (C) can be warranted non-empirically, which is generally thought to be wildly implausible. Given the implausibility of having a non-empirical warrant for claims such as (C), it is argued, externalism is inconsistent with the direct, non-empirical, authoritative nature of self-knowledge.[35]

The consequence problem has generated a vast amount of literature and three primary externalist responses have emerged. The first concerns the externalist conditional that connects thoughts with the environmental conditions they presuppose. According to this position, conditionals such as (P2) are false because they commit the externalist to a stronger thesis than is either plausible or established by the counterfactual scenarios that support it. And a true externalist conditional, which stated a genuine dependency relation between the thinking of a thought and environmental conditions necessary for it, would, according to this view, be so weak that non-empirical knowledge of its consequent would not be implausible at all. For instance, it might state that in order for S to think that silver is shiny, S would have to be related to some basic kinds of things (but not necessarily to silver).[36]

The second response abstracts away from questions about the content of the externalist conditional and focuses instead on what is wrong with S's reasoning even if her reasoning is sound. According to this strategy, arguments such as the one S reasons through are epistemically defective in roughly the same way that a question-begging argument is epistemically defective: neither a question-begging argument nor an externalist argument of this kind will be persuasive to a subject who doubts the conclusion. According to this position, the non-empirical warrant available at (P1) fails to transmit across the known conditional (P2) to provide a non-empirical warrant for (C), because, roughly, the non-empirical warrant for (P1) is only available on the assumption that the environmental conditions stated in (C) obtain.[37]

The final position argues, in contrast, that there is nothing epistemically wrong with S's reasoning. On this view, self-attributions are warranted non-empirically and without the need of a prior warrant for the claim that the environmental conditions that help to individuate the thoughts obtain. Moreover, in the absence of a doubt about whether the environmental conditions obtain, the non-empirical warrant for (P1) remains undefeated and can legitimately transmit via (P2) to (C). The response has met with incredulity, but is, I think, reasonable given two facts. First, a claim is equally open to doubt whether it is warranted empirically or non-empirically. This is important because it separates the question of whether there is anything epistemically defective with S's reasoning from the question of whether S's conclusion can be deployed in a straightforward argument against the sceptic. S's reasoning is epistemically legitimate, but does not refute external world scepticism. Second, (although I have not argued this here) externally individuated thoughts depend not only upon the existence of relations between the thinking subject and objective properties in her environment, but on the subject's knowledge of such relations. Consequently, while externalist arguments of this kind can provide a subject with non-empirical warrants for claims about her environmental relations, the subject will already have empirical warrants for such claims.[38, 39]

Conclusion

In this chapter I have tried to provide a relatively comprehensive overview of the internalism/externalism debate in the philosophy of mind and its implications. However, it should be clear where my allegiance lies. In the predicative case I find the considerations that motivate externalism persuasive, the theoretical gains of externalism significant, and the considerations against externalism misguided. On the metaphysical side, externalism is consistent with a naturalistic theory of the mind and with the claim that mental states are the causes of our actions. On the epistemological side, externalism is consistent with the direct, non-empirical, authoritative nature of self-knowledge, and in addition has the potential to ground an adequate theory of justification.[40] In the singular case, however, a two-factor theory strikes me as theoretically superior

for thoughts expressed both by sentences containing demonstratives and by sentences containing proper names. This yields a theory according to which names and demonstratives do not function in the same way as predicates, whether in language or in thought: our fundamental contact with the world is demonstrative, which itself grounds an externalist theory of the mind.

8 The Philosophies of Cognitive Science[1]

Margaret A. Boden

Cognitive Science and Pluralism

There's no such thing as the philosophy of cognitive science. Rather, there are competing philosophies of – and within – the field. That's partly because the concepts and techniques of artificial intelligence and artificial life have been changed and enriched since the 1940s. For although psychology is the thematic heart of cognitive science, its intellectual heart is AI/A-Life – or AI, for short.

Psychology (both animal and human) is the thematic heart because cognitive science studies all aspects of the mind or mind/brain, or, if you prefer, embodied experience and behaviour. It ranges from low-level vision to enculturated thought, from infantile development to adult personality, and from individual behaviour to social phenomena. It investigates not only cognition, but emotion and motivation too. So the field is badly named: outsiders are often misled, assuming that it deals only with cognition.

Cognitive science differs from other forms of psychology in using computational concepts of various kinds. Very broadly speaking, these fall into two main types: formalist/symbolic and connectionist/dynamical. Much of the philosophical interest lies in the differences between these approaches.

Some 'computational' concepts, in the broad sense intended here, denote formal computations on symbolic representations. These typify classical AI, or GOFAI, Good Old-Fashioned AI (Haugeland, 1985, p. 112). Others draw on cybernetic ideas about embodied and self-organizing systems. These include situated robotics, wherein the robots rely on direct 'reflex' responses to environmental cues; dynamical systems understood in terms of physical laws; and self-equilibrating neural networks. And all approaches sometimes include the sort of 'computation' (mutation and natural selection) that's effected by evolution.

Cognitive scientists often express their theories as computer models, because this is the best way of testing their coherence and implications. (Testing for their truth, of course, involves comparisons with the actual phenomena.) The computational concepts implemented in such models are substantive theoretical