You Can't Share Concepts

Cressida Gaukroger

DRAFT. DO NOT CITE WITHOUT PERMISSION

November 2017

Abstract

This paper begins by accepting the premise that if concepts can be shared, they must be individuated relationally. However, concept sharing is only an important feature of a theory of concepts if the sharing itself can explain behaviours such as linguistic communication. There is a widely-believed thesis that the relational properties of mental states do not have causal powers over behaviour. As any sharing must be in virtue of the relational properties of concepts, any account of concept sharing cannot, in fact, causally explain any behaviour, including linguistic communication. There is, therefore, no value in a theory of concepts being able to account for concept sharing. This undermines one of the strongest arguments in favour of externalism about concepts.

Introduction: What is Meant by Concept Sharing?

One of the most fundamental convictions held about concepts, independent of which theory of concepts one endorses, is that they are the kinds of things that can

be shared. As Gottlob Frege expressed it: '[O]ne can hardly deny that mankind has a common store of thoughts which is transmitted from one generation to another.' (Frege, 1948: 212) According to Jerry Fodor concepts '[a]re the sorts of things that lots of people can, and do, share.' (Fodor, 1998: 28) Jesse Prinz similarly includes concept sharing in his account of requirements on theories of concepts: 'Concepts must be capable of being shared by different individuals and by one individual at different times.' (Prinz, 2004: 14) These authors, and many, if not most, others agree: *Concepts are shared*. If true, then theories of concepts that cannot or do not account for how concepts are shared, are at best deficient, and, more likely, just false. It is appropriate at this stage, then, to ask: 'Why must concepts be shared?'.

The view that concepts are shared is intuitively appealing. The idea that having different experiences, living at different times, coming from different cultures, or any one of a range of possible variations between persons, would result in two people not sharing a concept such as 'food' or 'warmth' seems immediately counter-intuitive. But, *mere* intuitive appeal is poor grounds for rejecting theories that have other, more demonstrable advantages (see §1). Luckily, there is another answer to the question of why concepts must be shared, and this is that concept sharing explains a range of phenomena: coordination, behavioural success, linguistic communication (Frege (1948); Putnam (1970); Fodor (1998)) and the success of psychological explanations (Rey (1985); Fodor (1998)). The claim is that *none of these things would be possible* without shared concepts.

Take, for example, the case of language. In so far as the meanings of words can be understood by many people, argues Putnam (1970), they are constituted by shared concepts. What it is then to hold a concept, as with what it is to understand the meaning of a word, cannot be unique to its holder, but must be both simple and

general enough that its sharability can be explained. If concepts are not shared, what could explain the success of linguistic communication?

Similarly, the success of psychological explanations also appears to rely on concept sharing. Georges Rey (1985: 228) argues that concepts can be invoked to explain both interpersonal and intrapersonal psychological explanations. We believe there are important things in common in cases where, for example, two different individuals order champagne at a restaurant. These include a shared desire for champagne and a shared belief that ordering it is the way of fulfilling that desire, etc. Such an understanding of the psychology of others seems, once again, to require our being able to share concepts which we then attribute to them in order to explain their intentional states. If we did not share concepts then it is unclear how we could ever understand other people's mental states. Yet we have good reason to believe that we are often very successful in representing the mental states of others – we can use past explanations to predict future behaviour, and we can confirm whether our explanations are correct just by asking the relevant individuals to report on their own intentional states. Indeed, the very fact that these behaviours exist in the first place – behaviours that appear to require a shared categorisation of 'restaurant', 'champagne', 'money' etc., as well as coordination of somewhat complex behaviours between customers and waiters, for example – appears to further confirm the belief that most people, most of the time, share concepts.

We can understand this argument in favour of concept sharing as including the following premises:

1) Mental Sharing Premise: Concepts are mental things, which are common to many minds.

OR

1)* Public Accessibility Premise: Concepts are non-mental things, which are accessible to many minds.

2) Shared Behaviour Premise: Concepts explain shared behaviour.

3) Sharing Explains Behaviour Premise: Concept *sharing* explains shared behaviour.

In this paper I will take 'concept sharing' to cover both the Mental Sharing Premise and the Public Accessibility Premise. The dominant view of concepts understands them to be mental representations. If this is the case, then what it means for two people share a representation is that they have distinct token representations, that are of the same type. My elephant concept may be uniquely represented by my mind/brain, but it is a concept I have in common with many others because it was caused by contact with elephants, or because I am a member of a linguistic community where the term 'elephants' refers to elephants, or because it functions to detect elephants, etc. There are many theories regarding how concepts should be individuated, but what they have in common is that they agree that the correct taxonomy of concepts will identify concept types that are multiply realised across different individuals. If you believe concepts are mental representations, you are likely to accept the Mental Sharing Premise.

In contrast, an alternative view of concepts understands them as non-mental ab-

stract objects. In accounting for how concepts play a role in explaining the success of language, Frege (1948), for example, was concerned with how to reconcile the following: to be able to communicate linguistically, communicators need to have some kind of shared understanding or knowledge of the words they use. However, we cannot know everything about the referents of our words and often have very different mental representations associated with them from the representations held by others. For Frege, the units of which there must be shared knowledge to enable linguistic communication cannot be the referents of our words, because it is possible for one person to associate different concepts with the same referent. Similarly, these shared units cannot be our 'conception' of a thing – i.e. that which includes internal images, memories and sense impressions – as these features are vague, will vary between individuals, and will change over time. Conceptions, Frege argued, cannot explain the publicity of language – in fact they threaten to undermine it. Frege, therefore, introduced the idea of a 'mode of presentation'. As abstract objects, modes of presentation (herein, MOPs) do not vary from person to person (although people can have varying conceptions of the same MOP), but they are also not synonymous with reference, since more than one MOP could apply to the same referent.

If one takes concepts to be abstract objects, then what it means for concepts to be shared is that they are *publicly accessible*, meaning that they can be represented by multiple individuals, even if these representations themselves vary from person to person. While the Mental Sharing Premise can be understood as stating that concepts are shared in the way that a book is shared if multiple people own copies of *the same* book, the Public Accessibility Premise can be understood as stating that concepts are shared in the way that a book is shared if multiple people have

access to, or co-own, one and the same book.

If concepts did not explain behaviour, then the success of language and psychological explanations could not be explained in terms of concept sharing. This paper will challenge premises 1), 1)* and 3), but it will not challenge the Shared Behaviour Premise. If concepts do anything, they explain behaviour. And particular complex behaviours, 'Shared Behaviour' (herein SB), are some of the best candidates for the kinds of behaviour which would not be possible without concepts. The fact that humans are capable of complex linguistic behaviour; the fact that people perform the same detailed actions with what appear to be the same intentions; the fact that humans organise objects into the same categories for what appear to be the same reasons – all of these are possible because of concepts.

Being able to explain concept sharing may be an asset or even a non-negotiable requirement of a theory of concepts, but this does not mean that *any* type of concept sharing will do. Recall that, if we take concepts to be mental representations, then an account of concept-sharing will be a taxonomical account – one that classifies concepts such that people who have non-identical mental representations still have mental representations of the same type. However, there are an infinite number of ways we *could* classify concepts that would account for concept sharing in this sense. Most of them will be no better than not accounting for concept sharing at all. An account, for example, that said that individuals share a book concept if and only if they are a member of a book club, would satisfy the Mental Sharing Premise. Similarly an account that says that an individual has a duck concept if they have a concept that was caused by experiences of ducks prior to 2067, or a concept that was caused by experiences of rabbits after 2067, would also fulfil this requirement. Such a theory would, once again, be able to account for concept sharing, but would

nonetheless be a *terrible* theory of concept individuation. The fact that such theories satisfy the Mental Sharing Premise, does not make them any better than if they hadn't satisfied this premise, because they are unable to satisfy the Sharing Explains Behaviour Premise.

The Sharing Explains Behaviour Premise is crucial for the case in favour of concepts being shared. It can be understood as saying that either 1) or 1)* explains 2). This premise captures the belief that concepts explain SB *in virtue of their being shared* and, therefore, that any account of concepts that cannot account for how concepts are shared *cannot explain SB*. It clear that it is not the case that *just any* account of concept sharing could play a role in explaining SB. For the ability to account for concept sharing to be a *virtue* of a theory of concepts, the theory must first explain what the explanatory significance of its account of concept sharing is. In other words, it must explain *how* its account of concept sharing explains SB. It is the Sharing Explains Behaviour Premise that answers our question of *why* concepts must be shared.

As vital as the Sharing Explains Behaviour Premise is, however, it is rarely explicitly stated. Instead it is *assumed* that if one has an account of concepts according to which they are shared or publicly accessible, then the account automatically explains language and intentional explanations of behaviour. However, concepts being shared is not the same as concepts explaining the success of language. A step is missing. An account of concepts must not only be able to explain SB, but, if the Sharing Explains Behaviour Premise is correct, it must be able to answer the following question: What is it about the fact that (or the way in which) concepts are shared, that explains SB?

Rather than assuming that being able to account for concept sharing is a virtue of

a theory of concepts, I will take it to be a virtue only if a theory that accounts for concept sharing also satisfies the Sharing Explains Behaviour Premise. However, I will argue that none of the accounts that satisfy the Mental Sharing Premise or the Public Accessibility Premise can satisfy the Sharing Explains Behaviour Premise. In fact, concept sharing *cannot* explain SB. I conclude that being able to account for concept sharing is not a necessary feature of a theory of concepts after all.

1. Concepts as Internally-Individuated Mental Representations (i.e. Concepts as Conceptions)

Not all theories of concepts can account for concept sharing. There are theories of concepts that have many virtues, but are often dismissed because (among other things) they satisfy neither the Mental Sharing Premise nor the Public Accessibility Premise.

Assuming for the present that concepts are psychologically real entities or states, such entities will be comprised of certain internal or intrinsic psychological components. We can think of these as Fregean (1948) 'conceptions'. What these internal components are may vary depending on your theory of concepts, but at the very least they are whatever is in the head, whether that is mental images, basic 'switches' in the mind that indicate or detect features of the external world, mental feature-lists, exemplars, theories, or any other number of internally-defined mental components. As every person has different psychological states in virtue of their dispositions and experiences, no two people have concepts that share qualitatively identical psycho-

logical properties.

For theories that take concepts to be mental representations of some kind or another, what it means to account for concept sharing is to provide a taxonomy which individuates concepts such that different individuals possess concepts of the same type. This is what is required to satisfy the MENTAL SHARING PREMISE. There are two ways in which one could approach the question of concept individuation if concepts are mental representations. The first way of individuating concepts – of providing a theory that tells us what it takes for any two concept tokens to be the same concept type – does so on the basis of their internal characteristics alone. To say, for example, that where two concepts share the same internal features, they are the same concept, but where their internal features differ, they are different concepts. This way of individuating concepts can be understood as being 'internalist' as it honours one of the fundamental internalist litmus tests, namely that it entails that the 'Oscar-doppelgängers' of Putnam's (1975) "Twin Earth" thought experiment will have the same WATER concept.

The second way of individuating concepts can, in contrast, be understood as 'externalist', though here it may be easier to think of it as 'relational' as, insofar as we accept that concepts are psychological entities, it individuates such entities based on one or more of their relational properties.[1] These relational properties can be determined, for example, by the concept possessor's social and/or linguistic community (e.g. Putnam (1975); Burge (1979)), their teleological role (e.g. Millikan (1984)), or

[1]Segal (2000) defines the distinction between internalism and externalism about mental content in terms of intrinsic and relational properties. According to Segal, externalism is the theory that cognitive content is a relational property, while internalism is the denial that mental content supervenes on anything but intrinsic properties. Note that by 'relational' properties Segal and I both mean relational between mind and (external) world. Relations between an individual's concepts are 'internal features' for the purposes of this distinction.

their history of interaction with natural kinds (e.g. Kripke (1981); Putnam (1975)). A common argument in favour of having theories of concepts that individuates them externally, is that this the *only way* that a theory can satisfy the Mental Sharing Premise.

To understand this challenge to 'internalist' theories of concepts, consider imagism – the theory, roughly held by Locke and Hume, that postulates concepts as internal images, fainter copies of our perceptual experiences. Imagism can be understood for our purposes as being an 'internalist' theory of concepts. This theory is charged with failing to satisfy the Mental Sharing Premise because, insofar as a person's concepts are made up of combined images that have arisen from that person's perceptual experiences, they are going to differ from the concepts of other people in virtue of the fact that no two people will have indistinguishable perceptual experiences. If what it takes for two people to share a concept is for them to have *the same* (i.e. identical) mental images, then concept sharing will never happen.

For a more contemporary example, take prototype theory. The best known version of prototype theory explains concepts in terms of stored lists of the prototypical features of categories (Rosch (1973), (1975)). The features that make up one person's concept dog, for example, may include having a tail, barking, having a snout etc. According to prototype theory, the important properties of concepts – those properties that explain the role concepts play in guiding behaviour, and the role they play in thought (i.e. their relationship to other concepts and non-conceptual thoughts) – are their internal properties, namely the structure and contents of these feature lists. If we want to understand concepts, which includes understanding what makes two concepts the same or different, we need to look no further than how they are represented (as opposed to looking at what they refer to, or what the corresponding

word means in their linguistic community).

Prototype theory has a number of virtues. Most importantly it is able to capture and explain certain kinds categorising behaviour. As it originated in psychology, much of this behaviour is at a fine level of detail, performed by test subjects. For example, prototype theory predicts that if an individual has only (or mostly) been exposed to small, European birds, then a bird resembling a robin is likely to be the prototype she uses to distinguish her concept 'bird' (Barsalou (1987)). This is measured by timing how long it takes subjects to classify something as a bird – individuals are quicker at identifying and classifying birds that are closer to their prototype. This theory predicts that prototypes will vary depending on the birds one has been exposed to, but even with variation in prototypes there is usually enough similarity that there is a very strong overlap between what different individuals classify as a 'bird', even if their classification times vary.

Prototype theory is not without its problems, and there are several other competing theories which can be found in psychology, for example, which argue from experimental data that concepts are fundamentally different in their nature.[2] However, a primary criticism from philosophy has focused on the inability for this theory to satisfy the Mental Sharing Premise.

According to prototype theory, possessing a complete feature list is sufficient

[2]Further, Edouard Machery (2009) highlights the great deal of evidence that supports all the three main theories of concepts found in psychology: prototype theory (which states that concepts are representations of the most typical features of category members), exemplar theory (which states that concepts are a set of singular representations of particulars) and theory theory (which states that concepts are theories which store causal or functional information about category members). The fact that there is evidence to support each of these theories, argues Machery, indicates that none of these are the 'correct' theory of concepts. Rather, we have many heterogeneous types of concepts which may co-refer, but have very little in common. Machery's conclusion is that there is no correct or unified theory of concepts in psychology, so we should give up using the term 'concept' in this field altogether and instead recognise that different theories may not be in competition, but rather serve different ends.

for concept possession, although no one feature may be necessary for something to count as a member of the concept category. One person's DOG concept feature list may lack 'barks' as a feature, and another person's DOG concept feature list may lack 'has a tail' as a feature yet it could contain enough of the features of dogs that it could still be understood as being a DOG concept. If feature lists needed to be identical across people to count as being concepts of the same type, then it would be very rare for people to share concepts because feature lists are rarely identical across people.

There is evidence to suggest that the features considered to be prototypical of a particular category will differ depending on what tokens of that category-type individuals have been exposed to. In fact, it is even more complicated than this. According to prototype theory, our feature lists are not like a list on a page. Rather, the features that make them up will differ in importance or weight. Both 'has a snout' and 'is brown' might be on my DOG concept feature list, but 'has a snout' is a far more important feature than 'is brown'. It is, for example, more likely to come to mind when I think of dogs, and I will use it more when identifying dogs. This is referred to as 'graded structure'. The graded structure of concepts governs how likely you are to identify something as a typical instance of a category member. Consider, for example, the fact that for pretty much any category people will 'rank' items as being more or less typical members of that category (Rosch (1973, 1999)). Penguins are considered less typical birds than sparrows, for example. Indeed, this graded structure is thought to be at work even in obscure or novel concepts, for example, having an anvil dropped on your head is considered a less typical way of being killed by the mafia than being shot in an Italian restaurant (Barsalou (1987)). For two prototypes to be the same based solely on their internal properties would

mean that they were not only identical feature lists as regards their contents, but also that these contents are identically-weighted. However, it is unlikely that any two people's concepts will ever be the same in this way. As was shown by Barsalou (1987), people's graded categories change with context; there is a lot of variation between individuals (something that is not shown when results are only given as averages across groups); and the way people structure and apply their concepts does not even remain the same in one individual over time.

Insofar as it individuates concepts internally, prototype theory is unable to give an account of how normal people, under normal circumstances could have *the same* concepts. However, this is not something that it tries to do. Advocates of this theory reject the claim that a theory of concepts needs to satisfy the Mental Sharing Premise. Instead prototype theorists argue that it is sufficient to account for *similarity* between concepts. Even if feature lists are not shared, they can resemble one another. Rosch and Mervis (1975) describe the relationship between concepts in terms of the Wittgensteinian idea of 'family resemblance'. In other words there is nothing that we could define as being the real features of a particular concept, but rather we can identify a family of concepts which share some features with different members of other concepts in this family, even if there is no defining feature common to them all.

The approach of talking in terms of similarity as an alternative to concept identity has faced serious challenges. Jerry Fodor and Ernest Lepore (Fodor and Lepore (1992); and Fodor (1998)) reject the idea that there can be a meaningful account of concept similarity. They argue that for two things to be similar they must share some identical features. But, they continue, if identity is necessary for similarity, and one has a theory that cannot account for identity, they cannot simply avoid this

problem by claiming that all they are aiming for is similarity. They use prototype theory to illustrate this point: It is easy to claim that two concepts are sufficiently similar, they say, if they have enough overlap of feature-lists, but what is it for a feature on one list to overlap with a feature on a different list? If my concept DOG and your concept DOG both include 'has a snout' as an item on our feature-list then we are still going to have to explain what it means for both of us to mean the same thing by 'has a snout'. Such a feature minimally requires the concepts POSSESSION and SNOUT, but this assumes that these concepts are identical between two people who have similar concepts that combine such features. Fodor and Lepore argue that, while different theories of concepts may construct their accounts of concept similarity in different ways, they all face the same problem: "[A]ll the theories of content that offer a robust construal of conceptual similarity do so by presupposing a correspondingly robust notion of concept identity." (Fodor (1998): 34)[3]

The problem prototype theory faces, that it entails concepts are idiosyncratic, is equally true of other internally-based theories of concept individuation. Consider theories that individuate concepts on the basis of inter-conceptual relationships. If what it is to be a particular concept relies on other concepts (or areas of mental information) which themselves rely on yet other concepts, then we need to explain the whole network to say what a single concept is. As Fodor (1995b: 6) points out: "If what you're thinking depends on *all* of what you believe, then nobody ever thinks the same thing twice, and no intentional laws ever get satisfied more than once..."

Without factoring in some relational properties it seems that no account of concept individuation can classify concepts as types as opposed to individuating them

[3]There have been attempts to answer Fodor and Lepore, see Churchland (1993); Laakso and Cottrell (2000); and Prinz (2004).

so narrowly that only token-identical concepts can be said to be the same concept. Consider the book analogy: What it is to share concepts for any theory that takes concepts to be mental states or representations is like different people having different copies of the same book. However, if what makes two books count as 'the same' is that they share all the same words, have the same cover, and are in the same condition, then if everyone owns similar books that have different covers, contain a few different words, or range in condition, they cannot be said to have different copies of the same book. Any model that individuated concepts *solely* in terms of their psychological properties cannot explain how two individuals whose concepts differed internally, *even slightly*, could have the same concept. Indeed, if we can no longer talk in terms of two concept tokens being members of the same type then we cannot even give an account how one person can retain the same concept over time in the face of changing beliefs and experiences. Theories of concepts found in psychology – prototype theory, exemplar theory, and theory theory, for example – all face this problem. This is also a problem for internalist theories of mental content.[4]

We can understand, therefore, the significance of the belief that theories of concepts must either satisfy the Mental Sharing Premise or the Public Accessibility Premise. If not being able to account for concept sharing is a reason to reject a theory of concepts, then this may lead us to reject several theories that have many

[4]Joseph Mendola rejects this critique of internalism as he denies that internal properties are as private and idiosyncratic as we assume they are (Mendola (2008): 227). This follows from his unusual definition of internalism, which individuates the contents of thoughts as being the same if they lead to the same behaviour (defined a-contextually such that behaviour is determined exclusively by bodily movements) (Mendola (2008): 229; for some problems with this account see Ebbs (2013); Richard (2013)). However, it is not as if behaviour can avoid problems of vagueness that would undermine attempts at robust individuation, particularly if we cannot individuate it in terms of the relationship between a body and an environment. If behaviour is defined merely as the position of limbs, for example, what does it mean for limbs to have the 'same position' either in two people or in one person over time?

other virtues.[5] However, remember that these two premises are, on their own, not particularly important. Rather, what makes them important is their conjunction with Sharing Explains Behaviour Premise. The mere fact that internalist theories cannot account for concept sharing shouldn't be so bad, if it turns out that, counter to the way it is normally presented, concept sharing does not in fact explain sharing behaviour.

2. Concepts as Abstract Objects

Conceptions differ between people – this is the problem. As stated in the introduction, one way of accounting for how concepts are shared is to identify concepts with conceptions, but to individuate them according to their relational properties (as opposed to the features of these conceptions themselves). An alternative is to deny that conceptions are concepts.

Frege (1948) demonstrated that reference cannot be the only thing meant by a proper name, or rather that the extension of a proper name could not constitute its meaning. Frege went on to establish the notion of 'modes of presentation', which could play the role of the meaning of proper names. Rejecting a psychologistic understanding of MOPs, Frege instead identified MOPs as something objective. According to Frege, the meaning of a sentence is a proposition – an abstract object with truth bearing properties – so sentence meanings exist independently of the mind. Propositions are what one grasps as one comes to understand the meaning of a sentence. When two people have the same belief – such as the belief that 'elephants

[5] These virtues include being able to account for the significance of cognitive content; explaining concept acquisition; and explaining the role concepts play in categorisation (Prinz (2004)).

are large' – what makes it the case that they have the same belief? According to the Fregean tradition, 'elephants are large' is a proposition to which many people can all bear a belief-relation (it can be the object of multiple beliefs). Applied to concepts, one can understand propositions as being made up of more basic units – similar to sentences that are made up of words. As the proposition itself is abstract and mind-independent, this entails that the units that make it up will similarly be abstract objects. It is these basic units that make up Fregean propositions that are understood by some to be concepts.[6]

Georges Rey (1985; 1999), defends the view that concepts are abstract objects and rejects the idea that concept ontology requires a psychological component. According to Rey, concepts are like numbers – entities that can be grasped mentally, and used in rational thought, but which are neither mental, nor natural artefacts. This theory understands concepts in a Platonistic or Fregean way (Rey (1999): 296). Rey refers to his particular account of concepts as the 'Hypothesis of External Definitions': "the correct definition of a concept is provided by the optimal account of it, which need not be known by the concept's competent users." (Rey, 1999: 293)

Rey's theory of concepts, sometimes known as definitionism, appears to easily satisfy the Public Accessibility Premise. For example, the fact that being composed of H_2O is necessary and sufficient for a substance to be water is part of the WATER concept on this account, independent of whether anyone knows this. What it takes for a individual to 'possess' a WATER concept is for that person to have the appropriate relationship with the definition of water – a relationship that may well stand short of their actually (or perhaps fully) knowing this definition. For two peo-

[6]Peacocke (1992) and Zalta (2001) hold such a belief about the ontological nature of concepts. See Margolis and Laurence (2007) for a more detailed account of the view of concepts as abstract objects.

ple to 'possess' the same concept, they do not need to represent the same definition in the same way, but rather to both be referring to a category that is individuated by its definition conditions, even if they are not entirely aware of them. There is, therefore, no problem posed by the fact that there is variation in individual conceptions according to Rey.

It appears then that defining concepts in entirely non-psychological terms avoids the problem of people with different psychological lives still being able to share concepts. However, this is not the case. Arguing that there must be an externally located conceptual component that accounts for how psychologically or subjectively different individuals can possess the same concept does not avoid the problem that arises when we consider how these external components are represented internally. After all, the problem for which concepts as external entities were posited to answer is how we get from differing internal mental content to public communication. It seemed as if the problem was that internal ways of representing different referents would always differ. Rey and Frege claim that we do not internally represent the referent of a concept, but the concept itself. How does this solve the problem? These new external entities – abstract objects – must themselves be represented internally. The varying nature of internal representations, if it is a problem for theories that do not involve MOPs, cannot be solved via the introduction of MOPs.

Our having two different conceptions of the same abstract object may satisfy the Public Accessibility Premise, but it cannot explain SB any more than our having two different conceptions of the same concrete object can. This problem with the failure of abstract object theories of concepts to satisfy the Sharing Explains Behaviour Premise is mirrored in criticisms others have made of such theories. For example, Fodor (1998) makes the following observation: Frege brings in the idea of

MOPs, as reference alone is not enough to differentiate concepts. However, Frege was concerned to make MOPs external – otherwise they would fail to account for their public accessibility; for, if they were internal then they would differ from person to person. If concepts are conceptions, thinks Frege, then this would mean that there are, potentially, an infinite number of concepts relating to the one referent and, therefore, we cannot explain communication. But, Fodor points out, there are many ways that modes of presentation can entertained:

> [I]f MOPs *aren't* mental, what kind of thing *could* they be such that *necessarily* for each MOP there is only one way in which a mind can entertain it? (And/or: what kind of mental state could entertaining a MOP be such that *necessarily* there is only one way to entertain each MOP?) (Fodor (1998): 20)

Consider Rey's definitionism. If concepts are externally located definitions, a question arises over how these definitions are manifested mentally (or recognised, or learned). Even if we were to agree that only certain definitions need to be mastered for us to share the same concept, we must nevertheless be provided with an account of what it takes to master a definition. If mastering a definition requires forming a particular mental representation, then we are left needing to explain how these mental representations come to be identical.

Remember from §1 that concepts in the Fregean sense were brought in to solve the problem that conceptions, in virtue of the fact that they vary from person to person, could not explain shared behaviour. It was not just that an explanation in terms of abstract objects was given for SB. Rather it was believed that such an explanation was *needed* to explain SB precisely because explanations that looked to conceptions

were fundamentally unable to account for such behaviour. Defining concepts in entirely non-psychological terms was meant to avoid the problem of explaining how people with different psychological lives were still able to share concepts. However, unless we are puppets unconsciously controlled by the realm of abstract objects, the presence of such abstract objects means nothing for explaining human behaviour if they are not represented some way or another in actual human minds. If the question is how we get from differing internal mental content to public communication, the answer cannot come from bypassing a discussion of the role of conceptions in causing public communication.

3. Conceptions as Externally-Individuated Mental Representations

Abstract object theories of concepts are unable to satisfy the Sharing Explains Behaviour Premise, without which the Public Accessibility Premise is worthless. The problem that the abstract object theories of concepts face is that they claim to be able to explain SB, but fail to account for the mental mechanisms that actually cause SB. It is unclear what explanatory work abstract objects could do when it comes to accounts of SB. An alternative to the abstract object approach that has been more widely adopted is to argue that, while concepts are mental representations, the variation between them is irrelevant because their content is determined externally. We can understand this line of argument by using the analogy of words. A word, say 'canal', can be represented in many ways that vary widely – it can be written or spoken, pronunciation and writing style can differ etc. However, these

variations are irrelevant because what is significant about the word 'canal' – what explains how it is used, how it relates to other words, how it influences thoughts and behaviour – is what it refers to. The content of the word is not determined by any of the intrinsic features of its token presentations, but by a relational property, namely the reference relation that it bears to canals. Similarly to words, concepts pick out things in the world, and so if they can be individuated in terms of these relational properties, we can have an explanation of SB which is not undermined by the fact that there is variation in the intrinsic properties of individual conceptual representations. Or, at least, so the argument goes.

This approach to concept sharing is common to pretty much any 'externalist' theory of mental content which argues that mental content supervenes on the relational properties of mental states. For an example of someone who is very explicit about how this works in the case of concepts, consider Fodor. In "Concepts: A Potboiler", Fodor (1995a) distinguishes between the two different types of properties of concepts – internal properties (which he calls 'causal' properties) and relational properties (which he calls 'semantic' properties). Consider, for example, a SPOON concept. How my SPOON concept appears to me; what I associate with it; what prototypes I employ to identify spoons; even what patterns of neurons firing realise my concept – all of these are the internal properties of my concept. Consistent with the arguments presented in §2, Fodor argues that these properties must be irrelevant to concept sharing. According to Fodor, the SPOON concept held by different people will be realised by symbols with different syntactic features, or that play different roles in relation to their other concepts – this concept will differ in its internal properties from person to person. However, just because people differ internally, does not mean that they cannot share a SPOON concept. If individuals do share concepts,

it must be in virtue of the relational properties of their concepts.

Fodor (1995a) argues that I share my SPOON concept with anyone who has the right relationship with spoons: anyone who, like me, has *some kind of symbol* that is reliably activated when and only when spoons are present. For Fodor (1995a, see also Fodor and Pylyshyn (2014)), when we individuate concepts – when we say what it is for my SPOON concept to be a *spoon* concept, or when we identify what it is that accounts for psychologically different people being able to share concepts – this must be done on the basis of relational properties.

The problem with this view lies in the following: The relational properties of mental states do not have causal power over behaviour. (Herein the RELATIONAL INERTNESS THESIS.)

The RELATIONAL INERTNESS THESIS is a widely-believed thesis, even amongst those who believe that mental states must be individuated relationally.[7] A person's context may well have an effect on how they behave, but only insofar as it is mediated by intrinsic mental properties such as states of their brain. You needn't even believe that the mental supervenes on the neural (though many people do) to accept that the world cannot bypass the brain in causing a person to act.

A change purely in the relational properties of mental states, but no difference in intrinsic (or as we are calling them here, 'internal') properties of mental states, will see no difference in the causal powers of those mental states. This thesis is, in fact, so widely accepted that it is assumed in many of the scenarios used to illustrate the case *for* relationally individuating concepts (i.e. for externalism about concepts). Oscar and Twin Oscar continue to have very similar interactions with the watery

[7] Stich (1978); Burge (1982/1996); McGinn (1982); Fodor (1987); Jacob (1992), for example, all argue that, regardless of one's theory of content, causation is local.

stuff around their respective selves; the subject in Burge's (1979) arthritis example goes to the doctor and complains of the pain in his thigh in all possible worlds Burge imagines for him; and, despite the fact that he argues for all content being broad and, therefore, a complete change in broad content faced by Donald Davidson's (1987) swampman, both the original Davidson and his swampman replica exhibit the exact same behaviour.

The Relational Inertness Thesis has received a range of different replies arguing that the relational properties of mental states actually do have causal powers. Such replies mostly focus on the question of whether people with internally identical but relationally different mental states, such as Oscar and Twin Oscar, actually behave the same on their respective planets, or whether they behave differently.[8] In the simplest version of this objection it is pointed out that the doppelgängers are, after all, going to be drinking different substances, swimming in different substances, washing in different substances, and so on. It is argued that while their physical actions might remain the same, their *behaviour* differs between Twin Worlds.

However, this line of arguing misses the point. When Oscar and Twin Oscar behave differently (if you believe that they do), is this behaviour not caused by the internal properties of their mental states? The difference in their behaviour is explained by the difference in their environments, and the difference in their respective relations to their environments, *not* the difference in the *relational properties* of their respective mental states. It is only if the *relational properties* of mental states are doing the work then we can feel confident rejecting the thesis that the relational properties of mental states don't have causal power. Were the two Oscars 'switched'

[8] See Peacocke (1981), Evans (1982) and Hornsby (1986). See Fodor (1987); Jacob (1992); Gaukroger (2017) for replies.

so that Oscar was now on Twin Earth but still, presumably for a while at least, retained his 'water' concept, Oscar would behave exactly as Twin Oscar did on Twin Earth in spite of their different concepts – Oscar would drink XYZ just as Twin Oscar did, would categorise XYZ under the label 'water', would be able to successfully communicate about the watery stuff around him. This suggests that, in this case the difference in behaviour would not be explained by the difference in the relational properties of their respective concepts, but rather that the difference in their behaviour *and the difference in the relational properties of their mental states* are jointly explained by a third factor, the differences in their contexts. Accepting the role of context in determining behaviour does not require believing that the relational properties of mental states have causal power.[9]

To reject the Relational Inertness Thesis in a way that was consistent with satisfying the Sharing Explains Behaviour Premise, one would need to present an account where it is not just *any* relational properties of mental states that cause public behaviour, but specifically those properties that form the basis of a theory of concept individuation which can account for concept sharing. It is easy to understand what form an explanation of SB could take if given in internal terms as it is possible for the explanation to be entirely mechanistic. In contrast, individuating concepts relationally is an exercise in taxonomy or classification, not in identifying alternative causal mechanisms. Introducing an alternative classificatory system is not going to be enough to explain SB if the internal mechanism does not exist to

[9] I will use 'explaining behaviour' below to mean 'causally explaining behaviour'. While the relational properties of a concept could be used in an explanation of behaviour as a way of picking out a particular concept, its use would be *non-explanatory.* Consider someone in the past who, believing whales are fish (Sainsbury (2014)), utters: "The boat is damaged because it was hit by a fish." The classification as 'fish' helps the listener identify the subject of the description. This classification does not contribute to a causal explanation of what happened to the boat. See §4.

support it. Classification cannot explain what a failure in the mechanism is unable to explain.[10]

Some, for example, Loar (1988, 2003); Block (1986); Chalmers (2002); Prinz (2004), argue that there is a significant role for narrow content (content that supervenes on the internal properties of mental states), but it should be combined with broad content (that which supervenes on the relational properties of mental states). Such two-factor theories, however, are at no advantage when it comes to satisfying the Sharing Explains Behaviour Premise. Two-factor theories have no resources in explaining SB that go beyond those of internalist and externalist one-factor theories. If they individuate concepts on the basis of their intrinsic properties they will face the same criticism as regards accounting for concept sharing that those solely-internal theories do, namely they fail to satisfy the Mental Sharing Premise. If, alternatively, they individuate concepts relationally their theories may inherit other virtues, which I will discuss briefly below, but, for the reasons outlined above, they will be still fail to show how *concept sharing* explains SB. The tools you need to account for concept sharing just aren't the kinds of tools that can explain SB.

[10] There are additional challenges for externalists who believe their theories of concepts can explain behaviour such as communication. Pollock (2015) argues that social externalism is inconsistent with the standard account of communicative success. Mendola (2008) argues that vagueness or indeterminacy of reference causes problems for theories of semantic content based on reference. This Quinean point can extend to non-referential theories of concept individuation – it is not enough to say that DOG concepts are shared because they are caused by dogs, one needs to give an account of why it was something about the category *dog* that was doing the causal work as opposed to particular breeds of dog, a set of the specific dogs an individual has come into contact with, or animals that have dog-like properties etc.

4. Roles for and Alternatives to Concept Sharing

I have argued here that those theories of concepts that can account for the either the Mental Sharing Premise or Public Accessibility Premise are unable to account for the Sharing Explains Behaviour Premise. This is a problem if you take SB as evidence for there being concept sharing which, I have argued, is precisely what advocates of concept sharing do. The argument that an account of concepts must explain concept sharing to be able to account for SB is akin to an argument that concept sharing is a *necessary* requirement for a theory of concepts. I have shown that this argument does not work. However, this does not mean that there is no virtue in being able to account for concept sharing.

Being able to individuate concepts such that we can generalise across them is extremely *useful* – making generalisations on the basis of individual instances is the basis of science, it is the basis of folk psychology, it is what allows us to meaningfully talk of concepts in the first place. However, the fact that it is useful to be able to generalise across concepts, is not a reason to reject internally-based theories of concepts which do not have written into them an account of how to broadly individuate concepts. Such theories do not fail *as theories of concepts* even if they do not provide the grounds for developing a robust theory of concept individuation. In fact, remaining neutral on the question of how best to individuate concepts may be an advantage for such theories, as different models of concept individuation may be appropriate under different circumstances.[11]

[11] Burge (1986) argues that psychology itself type-distinguishes mental states on the basis of relational properties, and this extends to folk psychology; after all, we use intentional language to describe behaviour. However, the fact that broad language is used in psychology, does not entail that psychologists are (or should be) committing themselves to the position that the relational properties of concepts determine their content.

It is helpful to see the parallels between the value of having accounts of concept individuation, and the way that we understand and use statistical probability. Statistical probability is a function over a population – its value depends on the interests of the user for whom it is a tool. For an individual to have a particular probability of, say, developing Alzheimer's disease, means that she is a member of a population x where n-number of members of x have historically developed Alzheimer's. The probability of developing a disease for any one person is not an instantiated trait. Any one individual's likelihood of developing the disease is merely a function over a particular group of which she is a member. Her actually developing the disease, if she does, has a local explanation. Statistics can be very useful for making predictions, but they are not causes within themselves. Similarly, talking in general terms about behaviour can help us make group predictions, while we still acknowledge that individual mental states vary, depending on internal properties. If we individuate concepts via their relational properties in order to generalise across concept-possessors, we can still acknowledge that in any individual case the cause of behaviour will require an explanation in terms of internal properties.

To continue the analogy, any one individual will have a range of probabilities of developing a certain disease in virtue of the fact that they are a member of a variety of population groups. Jemima may, for example, have a particular probability of getting diabetes based on her family history, say, a 3% likelihood. However, when she is seen as a British citizen, she may have a 7% probability of getting diabetes. As she lives in Nauru, however, the country with the highest rate of type-2 diabetes, her probability in virtue of being a resident there rises to 30%. All these things are true of Jemima – she simultaneously has a 3%, 7% and 30% probability of getting diabetes. Probability in this case is a tool that allows us to do many of the things that

an individuation model should allow us to do – aid in explanations and predictions. However, its success in doing so should not be taken to reflect an ontological feature of the subject, in this case of Jemima.

This is essentially what we do in the case of concepts when it comes to groups and generalisations. To say that everyone who has a concept caused by seeing ducks has a DUCK concept allows us to make a number of predictions about their behaviour under likely and hypothetical conditions (for example, if I were to ask each one of them if they knew what a duck was, or to ask them to describe a duck). Talking generally in terms of behaviour and prediction does most of what we want in making conceptual generalisations – we have defined our population, and made (largely) testable predictions about them *as a population*. The variation in behaviour doesn't matter too much so long as it is in line with what we are using the generalisation for and the population that we have defined.

It is important to note that there are cases of useful generalisations across groups that do not pick out the features of *any* one member of that group. The idea of the 'learning curve', for example – a curve which measures the rate of progress in acquiring new knowledge or skills – has proved to be a useful tool, particularly when making predictions about the rate and structure of learning for a population. However, no one actually learns in a curve. Rather, graphs that show the rate of learning in single subjects will consist in steps and spikes – the smooth 'curve' comes from averaging over trails and/or participants. No one individual might have concepts in the way we describe them when individuating them in accordance with relational properties, and still the relational individuation could be useful so long as we were talking in terms of populations rather than individuals.

It is possible to argue, then, that it is not the role for a theory of concepts to pro-

vide a robust account of concept individuation. Therefore, being able to account for concept sharing is not a requirement (or even a desiderata) of a theory of concepts – it should not be used as the basis for rejecting a theory of concepts or preferring another.

And what about explaining SB? Well now we know that SB cannot be causally explained by concepts in virtue of their relational properties. This means that if we want conceptual explanations of SB, we must look to their intrinsic properties – which is what psychologists have been trying to do. Note that this does not mean that we cannot incorporate accounts of concept individuation to allow us to generalise across findings, but rather that we should understand the ability to generalise as a tool available to, as opposed to a feature of these accounts of concepts. It is not itself doing the explanatory work, but is rather something that aids the explanation by putting it in terms that we are able to understand and apply elsewhere.

Throughout the history of the concept sharing debate we have had two pieces of data that have seemed irreconcilable: internal mental states vary between people, and people usually behave as if they don't. But the fact that these are two points of *data* means that they are reconciled – they are both true of the world and we know this because we have observed them to be true. If we want to explain sharing behaviour in conceptual terms, we have to go back to their internal properties (the very things that vary) and understand that, counter to common beliefs, all the tools we need for giving a conceptual explanation of this behaviour actually lie there.

Conclusion

Understanding that being able to account for the Mental Sharing Premise or Public Accessibility Premise is not the same as (indeed is at odds with) accounting for the Sharing Explains Behaviour Premise is both complicated and important. It is complicated because in the past there has been equivocation between 'being shared' / 'being publicly accessible' and 'explaining SB'. It is easy to show that these don't *always* mean the same thing, but it is only when you realise that they *never* mean the same thing that you can begin to see the problem. It is important because this equivocation has lead to the rejection of certain theories of concepts on the premise that they can't explain concept sharing and the conclusion that they can't explain SB. It is time to reconsider such internalist theories and re-evaluate our committment relational theories of concepts which, it turns out, are no better for being able to account for concept sharing.

References

Barsalou LW (1987) The instability of graded structure. In: Neisser U (ed) Concepts and Conceptual Development: Ecological and Intellectual Factors in Categorization, Cambridge University Press, Cambridge, pp 101–140.

Block N (1986) Advertisement for a semantics for psychology. Midwest Studies in Philosophy 10(1):615–678.

Burge T (1979) Individualism and the mental. Midwest Studies in Philosophy 4(1):73–121.

Burge T (1982/1996) Other bodies. In: Pessin A, Goldberg S (eds) The Twin Earth Chronicles: Twenty Years of Reflection on Hilary Putnam's "The Meaning of 'Meaning'", M. E. Sharpe, New York, pp 142–160.

Burge T (1986) Individualism and psychology. The Philosophical Review 95(1):3–45.

Chalmers DJ (ed) (2002) Philosophy of Mind: Classical and Contemporary Readings. Oxford University Press, Oxford

Churchland P (1993) State-space semantics and meaning holism. Philosophy and Phenomenological Research 53(3):667–672.

Davidson D (1987) Knowing one's own mind. Proceedings and Addresses of the American Philosophical Association 60(3):441–458.

Ebbs G (2013) Mendola's internalism. Analytic Philosophy 54(2):248–257.

Evans G (1982) The Varieties of Reference. Clarendon Press, Oxford

Fodor J, Lepore E (1992) Holism: A Shopper's Guide. Blackwell, Cambridge, Mass.

Fodor JA (1987) Psychosemantics: The Problem of Meaning in the Philosophy of Mind. MIT Press, Cambridge, Mass

Fodor JA (1995a) Concepts: A potboiler. Philosophical Issues 6:1–24.

Fodor JA (1995b) The Elm and the Expert. MIT Press, Cambridge, Mass.

Fodor JA (1998) Concepts: Where Cognitive Science Went Wrong. Oxford University Press, Oxford

Fodor JA, Pylyshyn ZW (2014) Minds without Meanings: An Essay on the Content of Concepts. MIT Press, Cambridge, Mass.

Frege G (1948) Sense and reference. The Philosophical Review 57(3):209–230.

Gaukroger C (2017) Why broad content can't influence behaviour. Synthese 194(8):3005–3020.

Hornsby J (1986) Physicalist thinking and conceptions of behaviour. In: Pettit, McDowell (eds) Subject, Thought and Context, Clarendon Press, pp 95–116.

Jacob P (1992) Externalism and mental causation. Proceedings of the Aristotelian Society 92:203–219.

Kripke S (1981) Naming and Necessity. Blackwell, Maldon, Mass.

Laakso A, Cottrell GW (2000) Content and cluster analysis: Assessing representational similarity in neural systems. Philosophical Psychology 13(1):47–76.

Loar B (1988) Social content and psychological content. In: Grimm R, Merrill D (eds) Thought and Content, University of Arizona Press, pp 99–110.

Loar B (2003) Phenomenal intentionality as the basis for mental content. In: Reflections and Replies: Essays on the Philosophy of Tyler Burge, MIT Press, Cambridge, Mass.

Machery E (2009) Doing Without Concepts. Oxford University Press, Oxford

Margolis E, Laurence S (2007) The ontology of concepts – abstract objects or mental representations? Noûs 41(4):561–593.

McGinn C (1982) The structure of content. In: Woodfield A (ed) Thought and Object: Essays on Intentionality, Clarendon Press, Oxford, pp 207–258.

Mendola J (2008) Anti-Externalism. Oxford University Press, Oxford

Millikan RG (1984) Language, Thought and Other Biological Categories. MIT Press, Cambridge, Mass.

Peacocke C (1981) Demonstrative thought and psychological explanation. Synthese 49(2):187–217.

Peacocke C (1992) A Study of Concepts. MIT Press, Cambridge, Mass.

Pollock J (2015) Social externalism and the problem of communication. Philosophical Studies 172(12):3229–3251

Prinz J (2004) Furnishing the Mind: Concepts and their Perceptual Basis. MIT Press, Cambridge

Putnam H (1970) Is semantics possible? Metaphilosophy 1(4):187–201.

Putnam H (1975) The meaning of meaning. Minnesota Studies in the Philosophy of Science 7:131–193.

Rey G (1985) Concepts and conceptions. Cognition 19:297–303.

Rey G (1999) Concepts and stereotypes. In: Margolis E, Laurence S (eds) Concepts: Core Readings, MIT Press, Cambridge, Mass., pp 279–300.

Richard M (2013) Content inside out. Analytic Philosophy 54(2):258–267.

Rosch E (1973) Natural categories. Cognitive Psychology 4:328–350.

Rosch E (1999) Principles of categorization. pp 189–206.

Rosch E, Mervis C (1975) Family resemblances: Studies in the internal structure of categories. Cognitive Psychology 7:573–605.

Sainsbury M (2014) Fishy business. Analysis 74(1):3–5.

Segal GMA (2000) A Slim Book About Narrow Content. MIT Press, Cambridge, Mass.

Stich SP (1978) Autonomous psychology and the belief-desire thesis. The Monist 61(4):573–591.

Zalta EN (2001) Fregean senses, modes of presentation and concepts. Philosophical Perspectives 15:335–359.