Selfless Desires

Daniel Nolan 2005
Forthcoming in *Philosophy and Phenomenological Research*.

Abstract

David Lewis's unified theory of the contents of *de se* and *de dicto* attitudes faces a problem. Whether or not it is adequate for representing beliefs, it misrepresents the content of many of our desires, which rank possible outcomes in which the agent with the desire does not exist. These desires are shown to play a role in the rational explanation of action, and recognising them is important in our understanding of ourselves.

In Lewis 1979, Lewis offered a unified way of representing the contents of beliefs and desires. Assignments of propositions, conceived of as e.g. sets of possible worlds, have the problem that they do not straightforwardly represent *de se* beliefs and desires – beliefs about how *I* am, or desires about what happens to *me*. If I am told that two amnesiacs, D and E, have woken up in hospital in the perceptually identical Ward 1 and Ward 2, and I am told (and I can infer from observation) that I am one of those amnesiacs, I can believe full well that D is in Ward 1 and E is in Ward 2, and be unaware of whether *I* am in Ward 1 or Ward 2. If I hear that E will likely have a smooth recovery, but D has hours of painful surgery ahead of him, I might *desire* that I not be D. The belief that I am D is not the same belief as the belief that D is D (or that E is E) – I already have those beliefs, and those beliefs play a very different role in my cognitive economy (since they are trivial, or near enough, they probably do not have much of a role). The desire that I not be D is not the same as the necessarily frustrated desire that D is not D, nor the trivially satisfied one that E is not D. In some sense D and E might share the desire to not be D (suppose they are told the same things about the two amnesiacs, each are told that they are one of the two, and each are told about D's impending painful surgery). D's desire is not satisfied, and E's is. Lucky E, we might think! (Though it may be hard to see how E could have been unlucky in his desire to not be D.)

De se characterisations of beliefs and desires can give us useful information about a subject that descriptions that are not at all *de se* can lack. Some may take this to be a matter of a difference in the pragmatics of attitude reports, rather than as a difference in

mental states. Perhaps D's desire that *he* not be D is, after all, the desire that D not be D, though D may not realise that it is that desire. D's desire, one might think, has the same content as that of a deviant logician ill-wisher of D's, who hopes that D is not D – it would just not be apt to lump this mad desire in with the understandable desire D has, on the grounds that it would be misleading. (Presumably this view says that a desire ascription conveys a lot more information than the content of the desire).

Attitude reports are messy, it is true, and on most views they can be used to convey information other than supplying the content of the relevant state. But whatever we say about this, there seems to be a distinctive sort of attitude (or perhaps distinctive way of having an attitude) that goes along with *de se* characterisations. The state that I could describe as the state of my believing that *I* will undergo painful surgery is a different psychological state, with a different motivational profile, from the state described as the case of believing D will undergo surgery (even though if I realise that I am D, the first state might be rapidly inferred from the second). The state of having a perceptual belief that a ball is headed towards *me* (as we would say) may play a different sort of role from the perceptual belief that a ball is headed towards D (as we might describe it). The state that Mad Jim is in when he believes that he is Napoleon is in some sense the same state that Bonaparte is in when he believed that *he* was Napoleon – and that state they have in common is different from the state we would describe as the belief that Mad Jim is Napoleon (which perhaps neither has, and certainly Bonaparte does not have), or the state of believing that Napoleon is Napoleon (which Mad Jim and Bonaparte may share – but so do you and I!) Whatever you think the best way to refer to them is, one should recognize that these *de se* states are psychologically importantly different from non-*de-se* states. [1]

Many other "indexical" beliefs and desires can be understood as having a *de se* component. Some of these are obvious – a belief about home is a belief about *my* home, for example, and a desire to be famous is a desire that *I* be famous. Others are only

---

[1] As well as arguments in Lewis's 1979, see also papers by John Perry (1977), Hector-Neri Castaneda (1968), and Peter Geach (1957).

slightly less obvious. The perceptual belief that there is a tiger *here* is a belief that there is a tiger *near me*. Depending on one's view of believers and desirers, beliefs and desires *de nunc* may also be *de se*. An insomniac wondering what time it is may not be wondering about any non-indexical fact about the world – she may know ahead of time that she will be lying in the dark throughout the early hours of the morning, but only be wondering which time is at the same time as her thought. If believers and desirers are primarily time-slices of people, then a belief about "what time it is" can be construed as a belief about when *I* exist, (where "I" am the current believing timeslice). A desire that the train come soon is a desire that there is not very much time between the slice that exists when the train arrive and *me* (if "me" is the current desiring slice). If this reduction of "indexical" beliefs about time does not work, then self and time will need to be taken into account in mental content to adequately represent beliefs and desires. Let me suppose for present purposes that Lewis is right, and representations of "now" can be analysed in terms of representations of a time-slice self.

The content of *de se* beliefs and desires can be represented by a set of *centred worlds* – each centred world has a privileged individual, so that each <world, individual> pair represents a way things could be and which of the individuals at that world that *I* am represented to be. (Or that the representor is meant to be, more generally). If each individual is only found in one world (as with Lewis's counterpart theory), then instead of a pair of a world and an individual in it, we can just use the individual itself. The content of a *de se* belief or desire can then be given as a set of possible individuals – my doxastic alternatives, or the set of individual positions that I want to be in, insofar as the desire concerned goes. So the content of a desire that I be red haired is given by the set of red haired possible individuals, the belief that I am human has as its content the set of possible humans, and so on.

Armed with these contents, suited to be the contents of de se attitudes, Lewis offers a surprising and powerful reduction of attitude contents. Lewis points out that *de dicto* contents can often be represented with the contents he motivated by the case of de se attitudes. Take, for example, the proposition that some dogs have tails. That is often

represented in the possible-worlds tradition as the set of worlds where some dogs have tails – but it could also be represented by the set of possible individuals that exist in worlds where some dogs have tails. For me to believe that some dogs have tails, in effect, is to believe that I am such that some dogs have tails – that I am among those who exist in worlds where some dogs have tails. Lewis suggests that we can treat all belief and desire contents as being sets of possible individuals. To believe a proposition is to take oneself to be in one of the worlds where the proposition is true, to desire that a proposition obtain is to want to be one of the people such that the proposition is true in their world. All beliefs and desires *de dicto* are a certain sort of belief or desire *de se*. If, like Lewis, you are prepared to take properties to be sets of possible objects, then you can interpret every belief as a "self-ascription of a property" – to believe that it is raining is to self-ascribe the property of being such that it is raining. To believe one is rich is to self-ascribe the property of being rich, and so on. Lewis also offers a way of subsuming beliefs and desires that are characterised *de re* into his system (Lewis 1979 p 156), with the result that all beliefs and desires are, in the end, beliefs and desires *de se*.

(Strictly speaking, mental content is represented for Lewis by measure functions over possible individuals – credence functions over possible individuals for the "belief" side, and value functions over possible individuals for the "desire" side. (See e.g. Lewis 1981) This gives a finer grained picture than the all-or-nothing model discussed so far – despite this, I propose to largely ignore this complication for purposes of this paper, except insofar as I sometimes illustrate my claims about desires by pointing out an agent's preference for one outcome *over* another that is generated by a desire. This ranking of options is easier to make sense of in a system that assigns values to different worlds, but the complication of talking about credences and values is one that is largely unnecessary for the purposes of this paper.)[2]

The metaphysical details of how the reduction is to be carried out will vary depending on the metaphysics of worlds employed, but the idea that, in effect, *de dicto* content can be

[2] There is one interesting, though slightly technical, issue that does arise when we complicate things in this way – see footnote 6 for a discussion.

treated as a special case of *de se* content offers the prospect of a more unified treatment of mental content. Despite this, I will argue that Lewis's theory does not deal adequately with an important class of propositional attitudes. I do not suppose that this is entirely fatal – one may be able to reconstrue these attitudes, or rule them somehow out of court. I think it will be preferable to reject Lewis's unification rather than bite any of the bullets needed to save the theory, but I will be satisfied with demonstrating that Lewis's proposal faces previously unappreciated difficulties.

Lewis's discussion is primarily about beliefs, or self-ascriptions of properties, or locating oneself among a class of doxastic alternatives. The reduction of beliefs de dicto to beliefs de se is comparatively plausible – perhaps there is a difference between believing that snow is white and that I inhabit a world where snow is white (or that I am such that snow is white, if the reference to a world is suspicious), but even if there is a difference, at any rate whenever I have a true belief that snow is white it will also be true that I am such that snow is white, and that I inhabit a world where snow is white. So at least if we think the essence of the content of a *de dicto* belief has to do with its truth-condition, we should think the two ways of specifying the content are very closely connected in the way that matters.

Many desires seem to be represented adequately either way. My desire that I eat icecream is adequately represented by my desire that I am an icecream eater. My desire that it be warm is adequately represented as the desire that I am such that it is warm, or that I inhabit a world that is warm (around here, or warm everywhere if my desire is sufficiently cosmic in scope). But some desires do not fit this mould very well.

1. I sometimes wish I'd never been born at all

- Queen, Bohemian Rhapsody

Consider the desire that some have that they not exist – the desire that someone might express by wishing that they had never been born, for example. (Presumably this is not

the wish that they had come into existence by some other means – it is not the wish that they had been brought by the stork instead!) Now, this may not be a desire we approve of, or perhaps even that some think ideally rational people should have. But it is certainly a desire that we pre-theoretically think some people have – it is a commonplace of folk psychology that there are such desires, not an artificial philosophical construction.

This desire should be distinguished from the desire to not exist any more – desiring to not be found anywhere in the future can easily be represented in Lewis's system by a set of individuals that do not exist past a certain time (or counterparts of that time). But the desire to not exist at all cannot be represented by a set of individuals such that they are not found in their worlds, for there are no such possible individuals. Perhaps it is to be represented by a set of impossible individuals, such that they both do and do not exist in their worlds? I doubt Lewis would find this option very appealing, but in any case we might doubt that it does justice to the desire. It seems like the condition associated with the desire is possibly satisfied – there are plenty of worlds where neither I nor a counterpart exist – but a condition satisfied only by impossible individuals is an impossible condition, which would suggest the desire is for something impossible.[3]

The desire to not exist is unusual in one respect – it is a desire that is never satisfied when it exists. For the only way to have such a desire is to exist, and once one exists the desire to never-exist is frustrated. The desire to not exist is not unique in this respect – the desire to never have mental states, the desire to have no desires (ever), and some others are like this. These are odd desires, though, and Lewis might balk at admitting them. To have a desire, for Lewis, is to have a state that plays a certain causal role in one's psychology – and what sort of disposition to behave would be rationalised by this desire?

---

[3] It would not do to represent this desire with the set of individuals existing in worlds where I am not, or alternatively where there is no alethic counterpart of me. Given Lewis's interpretive strategy, that would be the desire that I be different from the ways I can in fact possibly be, but that is different from the desire to not exist. (For example the first, but not the second, would be entailed by the desire to be the number two, or a poached egg, unless there are some very permissive alethic counterpart relations according to which those are genuine possibilities for me.)

It is not clear what action is suited to satisfying it, for example, since necessarily it is never satisfied when it is present.[4]

There are several ways this desire can be manifest (as you would expect, since we commonsensically take ourselves to sometimes detect its manifestation in ourselves or others). One obvious way is in verbal behaviour – verbal behaviour is not conclusive proof of psychology, but someone who volunteers that they want to not exist, or want to have never been born, *prima facie* has that desire. It can manifest in other ways as well – it could help explain someone's resentment of their parents, for example, or feelings of regret when they contemplate their existence, or irritation when the subject comes up. And it can make sense of behaviour in the presence of other odd (perhaps incoherent) attitudes for which we may have good independent evidence: for example, a willingness to give money to a charlatan who claims to have a time-machine that can alter the past and erase one's existence. We can have independent reason to ascribe to someone a belief that time-travellers can make the past different from what it was the "first time around" – even if this belief is incoherent, it is not so obviously so that no-one can have it. This belief, plus the desire to not exist, plus a belief about the charlatan, can make sense of someone making a deal with the charlatan to wipe themselves out of existence. Or to take a more real-life case, there is the story of the famous presentist that was glad presentism was true, since it meant that in the future they would not exist (and according to his beliefs, not only not exist then, but not exist *simpliciter*). A desire that such a metaphysical view could turn out to be true can manifest itself in various ways – in wishful thinking, for example, or in the taking of steps to avoid the risk of wishful thinking – and we can make sense of someone's belief or desire in presentism even if we think it is incoherent.[5]

---

[4] Lewis once said about a particular desire that was impossible to fulfil "I have no idea whether or not I *do* have it. It is so disconnected from any guidance of conduct that I cannot tell how it would be manifest in my thought or action whether I had it or not" (Lewis 1986 p 126). This can be construed as a challenge to anyone who attributes a desire that is impossible to fulfil – could such a desire ever count as being manifested in thought or action, and if not, how could it satisfy functionalist criteria to be a desire with that content?

[5] Whether presentism means that a desire for one's own non-existence will eventually be satisfied is of course controversial, and it is far from obvious that presentism is conceptually incoherent. Nonetheless, even someone who thinks it is incoherent, or does not imply that a desire for non-existence will be

We might also have good reason to attribute such a desire based on behaviour that justifies us in attributing some broader desire. An agent might behave as if she hates all intelligent life, or even all life that counts as having beliefs and desires – seeking to destroy it, denouncing its corruption of Nature's harmony, and not merely seeking to destroy it, but also to prevent sentient creatures from coming into existence whenever possible. If this person tells us, with all apparent sincerity, that they desire that there be as little agency and intelligence in the world as possible, we may attribute to them the preference that they had never come into existence (or at least never come into existence as an agent or an intelligent life form). We can guess, in our usual folk-psychological and unsystematic way, how a desire that there be as little mental activity as possible might display itself. Now, it is possible to have the desire that there be as little sentient life as possible without having the desire that oneself not exist – one might desire rather to exist and be a non-sentient thing, or have an inconsistent exception for oneself. But we find it intelligible, if strange, to desire that one's own self not exist as a result of this more general preference. (I do not say rational, necessarily, only intelligible).

Other kinds of self-hatred are understandable enough, and some of those might lead to a desire for one's own non-existence. But I have laboured the point enough, I think. There is possible behaviour, voluntary and involuntary, which such a desire could explain – my guess is that some people do have such a desire for never having existed, and some of their behaviour *is* best explained as being caused by this desire. It might be thought that this is a minor anomaly for Lewis's view. Admittedly, such self-abnegation (or plain abnegation, for the purists) is a problem for many stories about *de se* content, and it may appear that this problem is primarily a problem for characterising *de se* content rather than for the reduction of other mental contents to the *de se*.

It should be noticed that this is a problem, and one that many theories of *de se* content as well as Lewis's should address (representing a *de se* content as a set of centred worlds,

---

satisfied, can make sense of a (possibly slightly counterfactual case) where a cluster of beliefs and desires about this topic could lead to behaviour that manifested a desire to not exist.

for example, it is hard to know where to locate the centre if the point of the desire is that there *isn't* one). It's main interest for my purposes, though, is that it points to something more significant about desires – that not all desires are desires only about what happens in worlds where oneself is found. We imagine that someone who desires to not exist may not be indifferent about how the world turns out in their absence – the person who hates all sentient life would prefer a world where no thinkers or desirers exist, and prefer a world with a few hundred animals that barely reach sentience to a world buzzing with billions of humans.

2. Selfless Preferences

More importantly, many of our usual desires apparently rank worlds where we do not exist. I want my loved ones to flourish, and would want the world to turn out so that they do well even if I had never existed. I am not indifferent to the possible world where my parents are kidnapped, separated, and tortured to death six months before my actual conception – I much prefer the possible world where they stayed together and were happy despite discovering that they could never have children (and in which I did not come into existence by any other means). Besides such "personal" commitments, we have desires about larger matters – the affairs of our nation, the progress of inquiry, the overall level of human happiness – which are not obviously desires that have anything particularly to do with ourselves. These desires are "selfless" in the colloquial sense, and I think there is truth to be found in this – we should take these desires at face value as involving preferences over worlds where one's own self does not exist. Many of us prefer that the Axis lose the Second World War, even if that meant that we ourselves would not have come into existence. We doubt that there is anything we can do now to make a difference one way or another to the outcome of WWII – but our desires about its outcome can manifest themselves in what we choose to regret and what we choose to celebrate, what we focus our attention on, the manner in which we would be taken in by charlatan time-travellers or behave with other rather unusual belief/desire sets, and our preference about the outcome of WWII can be evidenced by a more comprehensive belief/desire set – behaviour which suggests a general hatred of totalitarian dictatorships could be

behavioural evidence for a set of beliefs and desires which, inter alia, include the desire that the Axis lose WWII. Finally, of course, there is verbal behaviour – when someone tells you, apparently sincerely and without an obvious ulterior motive, and without evidence that they are somehow relevantly self-deceived or confused, that they think it would have sucked if the Axis had won WWII, you have grounds, even by functionalist lights, to attribute to them the desire that WWII not be won by the Axis.

Lewis's *de se* account can allow, for example, that I have the desire that the Axis lose WWII (or prefer that the Axis lose to it not losing, or whatever). However, in Lewis's reduction, my desire that the Axis lose WWII is just the desire that *I* be in a world where the Axis lose WWII. Lewis's surrogate does not rank the worlds where (a counterpart of) the Axis wins WWII but where I (or a counterpart of me) never exist. A desire that terrible things not happen is replaced by a desire that I not be around when they do (where "around" here means "am in the same world as"). Lewis himself notes this

> I note an analogy. The saintly crusader, who would like to live in a world without avoidable misery, is something like the snob who would like to live in a better part of town. Each wants a locational property. The crusader wants to be in a nice part of logical space, whereas the snob wants to be in a nice part of ordinary space. I trust the analogy redounds more to the credit of the snob than to the discredit of the crusader. (Lewis 199 p 146)

If indeed, the two desires turned out to be analogous in this way, then Lewis might be right that we should be less harsh in judging the snob. But to the extent that we think the humanitarian and the snob have different sorts of desires, we should be suspicious that the analogy is correct.

We can presumably distinguish the desire that there be no avoidable suffering from the desire that *I* not be in a world with avoidable suffering ("cosmic squeamishness", we might call it). One way the two might come apart is in what behaviour they produce when one considers counterfactual situations. The desire that there be no avoidable

suffering may produce regret or distaste when considering a world full of massacres and cruelty but in which the imaginer imagines that she never came into existence – the desire that the imaginer not be stuck in such a world may not. We might have reason to attribute one rather than another based on the source or rationalisation of the desire – someone who holds that human suffering is intrinsically wrong, and to-be-destroyed, and says that they think everyone is important in their own right, is someone who we might be inclined to attribute the "selfless" desire that there be no suffering. Someone obsessed with a certain moral "purity", that hates nearby suffering more than distant suffering, and hates recent suffering more than suffering in the recent past, and who talks about suffering always in relation to themselves, and who evinces emotions when contemplating suffering like those of people who fear contamination, might be suspected to have the desire only that *she* not be in a universe of suffering, rather than the desire that there be no suffering *per se*. (An odd way to be, perhaps, but then perhaps there is something a little unusual about the desire).

Likewise, there is a difference between desiring that the Axis lose WWII and the desire that I live in a world where the Axis loses WWII. (We could perhaps paraphrase the content of the second desire as the desire that *if* I exist, the Axis loses WWII). There are cases where the second desire seems particularly appropriate to ascribe – someone who says they are glad that the Axis lost WWII primarily, or solely, because they believe that if the Axis had won they would not have been born, presumably has the second. Likewise for an American or Russian who just wants to have been on the winning team. Someone who wants the Axis to have lost primarily because of a general hatred of totalitarian dictatorship, or because they strongly value the fact that the Nazi death camps were stopped on disinterested grounds, is more likely to have the first sort of desire – they are more likely to prefer a (not too dissimilar) world in which the Axis was stopped to a (not too dissimilar) world where the Axis is not stopped, independently of whether such worlds contain them or a counterpart of them.

3. A Wrinkle

Lewis sometimes suggested that words like "beliefs" and "desires" were "bogus plurals" (Lewis 1994 p 311) – that while ascriptions like "X believes that..." were in order, there was no reason to suppose that one's mental representation of the world, or ones preferences for the world, could be divided into single belief or desire-sized chunks in any meaningful way. Perhaps our beliefs, for example, were map-like or hologram-like – there was an overall representation of the state of the world (or the state of oneself-in-the-world, more strictly speaking), but no way to divide this representation into sub-representations with any physical reality, just as a hologram or map is less able to be divided than a page of writing into, say, sub-representations each of which contains a sentence worth of information and which together exhaust the information available.

Lewis does not say so, but this can be invoked to make the reduction of belief *de dicto* into belief de se more plausible. While my belief that some architects have noses is not on the face of it a belief about me, it is plausible that an agent like us with a total belief system sufficient to represent that some architects have noses *will* represent plenty of information about the self – including that the believer inhabits a world where some lawyers have noses. So when we are classifying possible total belief systems into the "believe that some architects have noses" sort and the "not believe that some architects have noses" sort, it is not crazy to hope that this will be the same demarcation as the one drawn by the "believes oneself is such that some architects have noses"/"does not believe that oneself is such that some architects have noses" distinction. Which suggests that the two characterisations of a belief state may not have a genuine difference between them.

Could the same thing be said about desires? Might it be that a total desire system that contains "selfless" desires might also contain enough self-involving desires so that there is no difference at the level of total systems between a selfless desire and its self-involving analogue? No. While a case can be made that rationality demands that a person assign no credence to worlds where they do not have an epistemic alternative ("I do not exist" is plausibly certainly false *a priori*), it is not plausible that there is an equivalent demand that we be indifferent between worlds where we have no desire-

alternatives. (Appetitive alternatives?) Two fairly ordinary distributions of value might differ only in worlds that the desirer takes herself to not exist in. Two people might agree about which worlds they want to live in, but disagree about whether a world with an Axis victory but without them is worse than a world without an Axis victory and without them. (And not just morally worse or better, but worse or better in terms of how they want things to be).[6]

In extreme cases, all of the most preferred worlds may be ones in which the agent does not themselves exist. Consider Sadsack. Sadsack wants most of all that he not exist – worlds where he is absent top his preference ordering. Sadsack may have preferences among worlds where he does not exist – for example, he might prefer that the Axis lose WWII to worlds where they win it. To be an intelligible agent, he may even need preferences somewhere along the line about what the world is like if he does exist – perhaps he prefers to cease existing quickly, or for people to leave him alone when he does exist, or he prefers to go unconscious quickly to going dancing, or whatever. Now, it might be that some would condemn Sadsack's preference ordering as irrational – and we would certainly be inclined to think that there was something wrong with him, at least if he was a human being. But even if there is some failure of rationality here, it is not as if Sadsack is unintelligible – it is not as if we could not use this information about his preferences to make folk-psychological predictions about what he was likely to do, for example. (Sadsack is a simple and fictional example, but he is not entirely unrecognizable, and more complicated versions of Sadsack may even be actual).

---

[6] One might think that once we move to the credence-function and value-function model of beliefs and desires, we do have a reason to hold that our desires are indifferent between outcomes that we assign credence 0 to, i.e. that we are certain will not obtain. In Jeffrey's influential model of decision theory (Jeffrey 1983), a proposition that has credence 0 has undefined desirability. This result can be resisted in various ways: if we allow conditional probabilities to be defined when the proposition conditioned upon has probability 0, for example, a very minor amendment to Jeffrey's system is all that is needed. Alternatively, we could say that Jeffrey is characterising rational belief/desire systems, and a desire in something we are certain will not come about involves us in some irrationality, and so involves a failure of Jeffrey's conditions. Finally, we could endorse some greater departure from this sort of system. The necessary repairs to decision theoretic systems that bar desiring certain falsehoods are easy, and if this paper is correct, well worth carrying out.

The point about intelligibility can be used to shed doubt on the argument about beliefs as well. Even if every entirely rational agent is sure that she exists, the sort of fallible agents who get confused or fall into misleading philosophical traditions or get very ill may still intelligibly have credences – and for those agents, their credence that *p* and their credence that (they exist and *p*) might intelligibly come apart. Perhaps, or perhaps not – but the case for desires that are independent of preferences about one's existence, or which conflict with a preference for one's existence, is much stronger.

The move to entire belief and desire attributions does not save Lewis's account, and so far as I can tell, there is no other minor refinement of Lewis's core idea which avoids the problems raised. This is not to say that there are other methods of reducing *de dicto* content to *de se* content that might be tried: representing properties of an agent *in absentia* seems to be the main thing required, so that a suitable method can ensure, for instance, that the Axis losing WWII is something that could be represented as happening *to me* in a world even when I have no doxastic/appetitive counterpart existing in that world. How much violence this does to plausibility as an account of the *de se* is another matter, though perhaps one to be settled when a specific alternative is set out and defended. And of course, there are other ways of unifying the contents of *de dicto* attitudes and *de se* attitudes besides reducing the *de dicto* to the *de se*, and there is no space here even for a survey of all the options, let alone a critical evaluation of them. However, for example, if we were to suppose that *de se* attitudes reduced to the *de dicto*, for example, rather than the other way around, I expect we would avoid the problems pressed in this paper. On the other hand, we would be left with the problems which originally drove theorists to postulate distinctive de se contents in the first place. I offer no quick resolution of the problems in this area - I suspect there are no quick solutions absent a general theory of mental content and what a theory of it should do for us.

4. Why it Matters

Suppose a Lewisian retreats. Suppose that they admit that desires *de dicto* are not all equivalent to desires *de se*, and concede that this is not just a matter of styles of reporting desires, but a difference at the level of desires themselves (or total valuation functions

themselves). Nevertheless, a Lewisian might claim, this reduction works for all practical purposes. Representing *de dicto* desires by their *de se* analogues may not characterise them precisely, but it does seem to work in the central case of decision making. For when we make decisions between options, we know we exist, and so our beliefs about how the world will turn out if we do one thing rather than another will represent each of the available outcomes as outcomes according to which we exist. (The outcomes may not be ones according to which we continue to exist in the future, but they are all ones according to which we exist at some time in the world). Whatever our preferences about how the world might have gone without us, the preferences that matter for decision are the ones about worlds where we exist. *That* choice was made for us already.

Representing all of our desires with *de se* desires will be adequate for some purposes, we should concede. Insofar as Lewis's model is simpler and more straightforward, that may be a reason to employ it, even if we think that it ignores some complexity that is really there. Lewis's model of assignments of belief and desire arguably does this in other respects in any case, by not distinguishing between some necessarily equivalent but hyperintensionally distinct representations of the world. Distinguishing the belief that 15+142=179 and the belief that 15+142=187, for example, may require some additional footwork – but see Lewis 1986 p 34-36 and Stalnaker 1984 chapter 5. This does not destroy the model's usefulness for some purposes.

While I am happy to grant Lewis's model of beliefs and desires at attitudes *de se* does not cause trouble in all applications, I think it does make a difference whether we conceive of beliefs and desires as ultimately all being *de se* or not. The first point to make is that there is more to our understanding our minds and ourselves than constructing a model that predicts behaviour correctly – even if it were true that both theories made the same predictions about behaviour, there would still be an interesting question about which one was correct.[7] We engage in philosophy of mind partly to improve our ability to make

[7] Some instrumentalists might think that there is nothing of interest over and above getting correct predictions – but this opinion is by no means universal among instrumentalists about beliefs and desires. Some think that there is an interesting question about which picture of beliefs and desires is right, but the

predictions and provide retrospective illustrations of what happened in a useful model, but also to understand ourselves, how the world seems to us, and how we value aspects of it. Running together egocentric desires and non-egocentric desires does not strike us as a "don't care" issue – it would be interesting if the humanitarian turned out to be analogous to the snob, but we care about whether that is in the end right, even if we could get along with predicting behaviour on either assumption. We care (or some of us do) about understanding our minds and ourselves – and for most of us, that is more than the task of constructing a reasonably simple model that predicts choices adequately. Perhaps this point could be put in terms of respecting the phenomenology of desire – but I will refrain, for fear of producing more smoke than light.

In any case, I think which model is right can make a difference in predicting decisions and other behaviour. People's patterns of regret and satisfaction can have something to do with how they rank worlds where they do not exist – someone who regrets having come into existence is likely to behave differently when contemplating the fact that they exist or in how they feel towards their parents. A selfless cast of mind, corresponding to having strong desires that are not *de se*, may well have behavioural consequences that are different from someone who has only *de se* desires, either constitutively, in some of the manners mentioned in this paper, or as a contingent psychological generalisation. If I am right that there is a difference between selfless desires and the "not in my cosmic backyard" variants, this can manifest in many ways relevant for the "practical purpose" of predicting and explaining behaviour.

It can even manifest in the making of choices. Our decisions about what to do are not just influenced by our desires concerning the options immediately before us, but by our beliefs about other desires we might have. If I guess that I have a preference that the Axis lose WWII, I might expect to have a reaction to, for example, reading a story according to which the Axis triumphed. I might guess that my reaction would be

---

answer is that *neither* of them is right – beliefs and desires are merely useful fictions, but any such story gets the real facts about human cognition wrong. Other instrumentalists might think there is an interesting question about which is right, but that the answer is epistemically inaccessible except insofar as a difference in theory turns up in a difference in predicted behaviour.

different than I would have if I merely had the desire that I not be in an Axis-triumphant world. And this could in turn influence my decision about whether to read the story. (*How* it manifests may be a very complicated matter – I may even read the story *because* of my desire that its contents not be true, for example because it is more interesting).

That case may be a little too sophisticated, since it dealt with a difference flowing from a difference in belief about which theory was correct. But when we predict the choices of others, information about the desires they have that do not bear directly on the choice at hand can be useful. We might expect, perhaps as a contingent generalisation, that someone who wishes that they had never existed, would make choices differently from someone who desired... whatever it is that a Lewisian would attribute to someone, e.g. disposed to say "I wish I had never existed" or "I want things to be as if I had never existed". Suppose we are faced with two people, one with the desire that there be no poverty and lack of education, and the other with the desire that there be no poverty and lack of education in *their* world (the "cosmic snob"). Suppose each is given an opportunity to ensure that they never come into contact with the poor and uneducated. My guess is that it is more likely that the person with the egocentric desire to be in a world free of such people would be more likely to take it than the person with the non-egocentric version. This may be because of a hunch I have about how different desires tend to go together, as a matter of contingent psychological fact. My guess may or may not be right (and there are many other factors at play of course). But if I am right, the two attributions lead to different predictions about choice – so at least in principle, this is another way that the distinction urged in this paper could make a difference to prediction of choices.

Finally, there are odd (perhaps necessarily irrational) cases where someone believes that their options are not all options according to which they exist. When the charlatan timetraveller offers Sadsack the chance to be erased from existence, we may well predict that Sadsack is more likely to hand over money than average. When the humanitarian hallucinates that God has offered her a prize in return for accepting her erasure from all

existence, it makes a difference to what she might do in response whether the prize is universal happiness or the triumph of the Axis in WWII.

Rejecting Lewis's reduction of attitudes *de dicto* to attitudes *de se* allows us to appreciate differences that Lewis's scheme ran together. The broader lesson is that even *de se* contents should sometimes make distinctions between possibilities in which one is represented as absent, and this is important *a fortiori* for *de dicto* contents. Lewis, of course, may not have thought the game was worth the candle, since his metaphysical picture of equally real concrete possibilities suggested that all useful information *was* merely "location" information – information about where "I" am in the space of possibilities. It is hard to see what a purely *de dicto* desire would come to in a system where, in effect, what was contingently true at all was a matter of where "I" happened to be – it is hard on this framework to make sense of a desire for something contingent that is not in effect a desire about where I am. So Lewis himself may have wished to bite the bullet, and find some other explanation of the distinction I have been urging. Insofar as this distinction in attitudes is a real one, and one recognized by our ordinary understanding of the world, the rest of us may instead take ourselves to have an additional reason to reject Lewis's metaphysical picture of our situation. *That* dispute, however, can be left for another occasion.[8]

Daniel Nolan
Departments of Philosophy
University of St Andrews
Edgecliffe, The Scores
Fife, Scotland
KY169AL
United Kingdom

[8] Thanks to audiences at Western Washington University, the University of Michigan, Cambridge University, and Keele University for questions and comments.

References

Castaneda, H-N. 1968. "On the Logic of Attributions of Self-Knowledge to Others". *Journal of Philosophy* 65:439-56

Jeffrey, R. C. 1983. *The Logic of Decision*, 2nd edition. University of Chicago Press, Chicago.

Geach, P. 1957. "On Beliefs About Oneself". *Analysis* 18:23-24

Lewis, D. 1979. "Attitudes *De Dicto* and *De Se*". Reprinted in Lewis, D. 1983. *Philosophical Papers Volume I*. Oxford University Press, Oxford, pp 133-156. Page numbers are from this reprint.

Lewis, D. 1981. "Causal Decision Theory". Reprinted in Lewis, D. 1986. *Philosophical Papers Volume II*. Oxford University Press, Oxford, pp 305-336. Page numbers are from this reprint.

Lewis, D. 1986. *On the Plurality of Worlds*. Basil Blackwell, Oxford.

Lewis, D. 1994. "Reduction of Mind". Reprinted in Lewis, D. 1999. *Papers in Metaphysics and Epistemology*. Cambridge University Press, Cambridge, pp 291-324. Page numbers are from this reprint.

Perry, J. 1977. "Frege on Demonstratives". *Philosophical Review* 86: 474-97

Stalnaker, R. 1984. *Inquiry*. MIT Press, Cambridge MA.