dialectica

*dialectica* Vol. 71, N° 1 (2017), pp. 3–33
DOI: 10.1111/1746-8361.12169

# Rational Agency without Self-Knowledge: Could 'We' Replace 'I'?

Luke Roelofs[†]

Abstract

It has been claimed that we need singular self-knowledge (knowledge involving the concept 'I') to function properly as rational agents. I argue that this is not strictly true: agents in certain relations could dispense with singular self-knowledge and instead rely on plural self-knowledge (knowledge involving the concept 'we'). In defending the possibility of this kind of 'selfless agent', I thereby defend the possibility of a certain kind of 'seamless' collective agency; agency in a group of agents who have no singular self-knowledge, who do not know which member of the group they are. I discuss four specific functions for which singular self-knowledge has been thought indispensable: distinguishing intentional from unintentional actions, connecting non-indexical knowledge with action, reflecting on our own reasoning, and identifying which ultimate practical reasons we have. I argue in each case that by establishing certain relations between agents – relations I label 'motor vulnerability', 'cognitive vulnerability', 'evidential unity' and 'moral unity' – we would allow those agents to do everything a rational agent needs to do while relying only on plural, rather than singular, self-knowledge. Finally, I consider the objection that any agents who met the conditions I lay out for selfless agency would thereby cease to qualify as distinct agents, merging into a single agent without agential parts. Against this objection, I argue that we should recognise the possibility of simultaneous agency in whole and parts, and not regard either as disqualifying the other.

## 1. *The I-concept and the We-concept*

Many philosophers have claimed that the first-person singular concept – the 'I-concept' – is indispensable for rational agents.[1] For a variety of reasons, it has been thought, we could not be the kind of reflective, critical, reasoners that we are without using this concept. In this paper I dispute this claim, arguing that at least one other concept can serve the same functions. Thus there could be rational agents which are 'selfless' in the sense of not employing the first-person singular concept.

More specifically, I argue that for agents in certain relations (to be specified), the first-person *plural* concept (the 'we-concept') can serve the same functions

[†] School of Philosophy, Australian National University, Australia; Email: luke.mf.roelofs@gmail.com

[1] As well as being explicitly endorsed by some (e.g., Burge 2000; Smith 2011, 63), this requirement is implicit in standard ways of defining 'person', such as Frankfurt's definition as a being capable of desiring that *it* have different desires (Frankfurt 1971, 7), and Locke's famous definition as "a thinking intelligent being, that … can consider itself as itself, the same thinking thing, in different times and places" (Locke 1975, 335). This definition (which is frequently accepted and endorsed, see, e.g., Olson 2007, 9) does not explicitly claim that all rational agents must also be self-conscious, but that is the natural implication of thinking that it identifies the single interesting category in this vicinity.

as the singular. I do not here argue for the even more radical position defended by Parfit (1999), that there could be rational agents who used no first-person concepts at all, and did not even think in terms of subjects or agents, but only of streams of experiences and actions. Those agents would be 'selfless' in a more radical sense than those I will discuss; call my sort of agents, which rely on a first-person concept other than the I-concept, 'moderately selfless agents'.

The possibility of moderately selfless agency is closely connected to another, commonly overlooked or denied, possibility: that of 'seamless' collective agency. Many authors have argued that we should allow for a group of human beings, when suitably related to one another, to qualify not only as acting collectively, but as collectively composing an agent (see especially Rovane 1998; Gilbert 2002; Pettit and List 2011). We should thus recognise not just collective agency, exerted by many agents together, but also group agency, exerted by the group itself. However, even if we accept this, any group agent which human beings would normally compose will be different from individual agents in a crucial way: all its operations will involve the knowledge (on the part of its members, and to that extent on its own part) that it is a group, and that its members are themselves distinct agents with potentially divergent interests and beliefs. Everything it does will be done through an individual agent self-consciously deploying their own separate agency. Could there be a group agent who lacked this automatic awareness of its own basis in individual agents, a group agent which operated with an awareness only of itself? Call such a group agent 'seamless': a seamless group agent is one whose agential functioning is independent of any knowledge of its underpinnings in individual agency. Insofar as they do rely on knowledge of their own division into members, this knowledge is like our knowledge of our bodies' division into limbs; while we may need to know about our arms to carry out many tasks, we do not need to think of our arms as agents, or to address them as such.

I define seamless *collective* agency as the agency of many agents who together form a seamless group agent. One prominent objection to the possibility of seamless collective agency is that seamless group agency would necessarily cease to be collective (Rovane 1998, 2005, 2012). Even if a seamless group agent was constituted by several human beings, they would cease to be individual agents precisely insofar as they subserved the agency of this seamless group. There would just be one multi-body agent.

Against the view that seamless group agents would no longer be groups of agents, I argue that even though a collection of suitably-connected moderately selfless agents might constitute a group agent, which could be seamless in the sense defined above, this would not impugn their own existence as rational agents. Thus I simultaneously defend two possibilities – selfless agency and seamless collective agency – which, if realised, would be two sides of the same coin.

If my argument succeeds, that will be significant both for the question of what rational agency requires, and for the question of what forms collective agency can take.

But here is a third reason to find these questions interesting. Our own human agency is clearly 'seamless' – it does not involve any awareness of being based in the agency of distinct, self-aware, parts or members of us. Thus if seamless collective agency is impossible, we can be confident that we are not *ourselves* group agents. Thus we can confidently reject any *literal* interpretation of 'homuncular' accounts of our mental functioning – we can confidently declare that if some cognitive model posits subsystems *acting* in their own right, doing agentive-sounding things like 'searching for opportunities' or 'competing for access' or 'inferring a conclusion', that model *must* be merely a fanciful metaphor (for examples of more-or-less homuncular cognitive models, see Selfridge 1959; Baars 1997; Shanahan and Baars 2004; cf. De Sousa 1976; Dennett 1991). This dismissive view is expressed nicely by Lowe, who writes:

> … the self patently does not consist of a plurality of lesser 'selves' acting cooperatively, despite the picturesque 'homuncular' descriptions of mental functioning advanced by some philosophers. Such descriptions are not intelligible if taken literally. (1996, 39)

If my arguments succeed in establishing the in-principle possibility of seamless collective agency, it becomes a live option that we are all group agents, with many rational agents seamlessly integrated within us (we might perhaps call this 'homuncular realism'). If my arguments succeed, we must accept that the 'mereological fallacy' (Bennett and Hacker 2003) is not necessarily fallacious. In a similar vein, Fernyhough (1996, 2004) argues that phenomena like inner speech should be understood as literally 'dialogic', being just internalised versions of interpersonal interaction that retain their essential features despite taking place within a single mind. Recent philosophical criticism (Gregory 2017) has challenged this approach on the grounds that it ignores the difference between inter-agent and intra-agent dynamics: since inner 'dialogue' does not involve multiple agents interacting, it cannot be the same kind of activity as inter-personal dialogue. If my arguments in this paper succeed, they undermine this kind of criticism – for all we know, there really are many agents within us.

Of course, my conceptual arguments cannot in themselves establish that we *actually are* group agents, and the full defence of this possibility requires discussing several other issues – about the unity of consciousness, the privacy of experience and the structure of the brain – which I address in other work (Roelofs 2014, 2016, Forthcoming-a, Forthcoming-b). But showing the

correlative possibilities of moderately selfless agents and seamless collective agency is one step towards a compositional view of our own agency.[2]

Quite apart from the questions of principle raised by the possibility of selfless agents and seamless group agents, such agents are common in speculative fiction. Many novels (e.g. Stapledon 1930; Vinge 1992; Naam 2012) depict collective minds, existing spread among many individuals and entirely supplanting any individual sense of self. Television shows such as *Star Trek* similarly explore the possibility of shared minds, in such forms as the Borg, the Founders, and the Vulcan Mind-Meld. Examples could easily be multiplied. The possibility of selfless agency also bears on the prospect, in our own technological future, of human collectives of whatever size that use information technology and neuroscience to integrate their brain processes. Similarly, it bears on the design of artificial intelligences whose divisions and boundaries might be more fluid than those of human individuals. In all these cases it is worth asking whether these imagined cyborgs might switch from employing the I-concept to employing the we-concept, and vice versa, without any interruption in their agency.

It will be useful in the coming discussion to have a relatively concrete example available to work with. So let us suppose that two humans, Alfie and Bettie, living about a hundred years from now, have undertaken to become not just married but hyper-married, connected so intimately that they leave behind singular self-consciousness and think only in terms of 'we the pair', regarding their distinct bodies in something like the way that someone might regard their distinct limbs (example adapted from Rovane 1998, 141). Although they may always retain the knowledge that they were and are individual agents, they aim to make their agential functioning independent of thinking of themselves as such, thereby becoming selfless agents and coming to compose a seamless group agent. In the following sections we will get a clearer sense of what Alfie and Bettie would need to do, for this project to be capable of success.

## 2. *Definitions*

Let us define some of these terms more precisely. What I will call the 'I-concept' is the concept expressed by the English words 'I', 'me' and 'myself', and which is distinguished by its referring not to any objectively specified subject but simply to whichever subject produced the thought or utterance containing it. If someone

[2] It is perhaps worth emphasising that being a realist about homunculi in the sense of agential or psychological parts of a mind does not, by itself, imply that they match up in any neat fashion with parts of the brain, divided along neuroanatomical lines – the mind and brain might both be composite, even if their respective decompositions often cross-cut (cf. e.g. Bechtel 1994; Rosenberg 1994).

uses a concept to refer to something other than themselves, it is not the I-concept – I take the I-concept's referential role to be its defining feature. What I will call the 'we-concept' is that expressed by the English words 'we', 'us' and 'ourselves', which is distinguished by its referring to some (contextually-determined) group of subjects, one or more of which of which produced the thought or utterance containing it.[3]

I will assume (though none of my arguments will rely on this) that the we-concept is a key constituent of the 'we-attitudes' discussed in the literature on collective intentionality, such as 'we-intentions', the intention to do something together with others, or 'collective guilt', the feeling of guilt at something one's group has done (see, e.g., Tuomela and Miller 1988; Searle 1990; Gilbert 2000, 2002). If this is so, then selfless agents might alternatively be described as rational agents all of whose attitudes are of the we-involving sort rather than the I-involving sort: all of their intentions are we-intentions, all of their guilt feelings are collective guilt feelings, etc.

It might seem at first that deploying the we-concept automatically implies deploying the I-concept, since by thinking of a group as 'us' a subject must think of it as including themselves.[4] But although the we-concept can be defined in terms of the I-concept, this does not show that possession or use of one must involve possession or use of the other, for there are ways to acquire concepts independently of understanding a definition of them. (After all, perhaps the closest thing to a definition of 'I' is 'the thinker of this thought', but that does not mean that a subject cannot have the I-concept without also having mental-state concepts.) A subject could learn to apply the we-concept to groups they belong to on the basis of the social experience of being and acting in a group, prior to or without ever learning to apply the I-concept, even though any subject fully possessing both would be able to derive claims involving one from claims involving the other.[5]

[3] I take no stand on whether such reference to a group should be taken as singular reference to one whole, as plural reference to many members, or as somehow indeterminate between the two, or doing both at once.

[4] Similarly, some authors writing on we-attitudes hold them to be reducible to or analysable into certain sets of I-attitudes, such as an intention to do one's part together with a belief that others will do their parts – for actual, much more nuanced, analyses, see Tuomela 2005, 340–341; and Bratman 2009, 155ff).

[5] If we individuated concepts sufficiently finely, then we might have to say that the concept learnt without the definition, and the concept defined by the definition, are in fact distinct concepts. But clearly there is some sense in which we want to say that a child who uses 'I' before they learn to self-ascribe mental states is using the same concept as they will after learning to do so, though perhaps using it with less sophistication. If fine-grained individuation of concepts stops us from saying this, so much the worse for fine-grained individuation of concepts. (Cf. the distinction between concept-possession and concept-mastery, Ball 2009, 955–956; Alter 2013.)

The claim I will be arguing against may be called 'the Indispensability Claim', and a corollary of it may be called 'agentive anti-homuncularism':

> **Indispensability Claim:** Successful deployment of the I-concept is indispensable to performing some of the essential functions of rational agency.
> **Agentive Anti-Homuncularism:** If our agency is seamless, then none of the brain parts or subsystems which subserve our agency can be themselves rational agents.

The phrase 'essential functions of rational agency' means those which something cannot be a rational agent without performing; over sections 3–7 I discuss four specific functions that support rational agency *per se* and for which the I-concept is claimed to be indispensable, and a fifth sense of 'impaired' by which it is claimed that moderately selfless agents would cease to be agents, instead being absorbed into another agent. The term 'rational agency' aims to pick out the special sort of agency that normal adult humans have and which small children and animals lack, even if the latter may still be agents in the sense of acting for reasons. That is, 'rational' here expresses the pre-theoretical idea that adult humans 'have reason' and other creatures do not. This goes beyond mere 'acting for reasons', since it seems appropriate, pre-theoretically, to explain many instances of animal behaviour by citing what the animal wanted and how it took the world to be, and such an explanation could naturally be called 'identifying its reasons for acting'.[6]

The Indispensability Claim should be distinguished from two weaker claims that I will not argue against here. One is that something could not be a fully-functional rational agent and have no 'I-representations' of any sort, conceptual or pre-conceptual. Denying this claim would mean denying, for instance, the Sartrean claim that all consciousness, of any sort, involves pre-reflective consciousness of oneself. Another claim weaker than the Indispensability Claim is that no fully-functional rational agent could *lack the I-concept* entirely, being unable even to understand what it would mean: by contrast with this, the Indispensability Claim concerns not just possession, but successful application, of the concept: I will take a 'successful' application to be one which constitutes knowledge.

In the next four sections I review four cognitive functions for which the I-concept has been claimed to be indispensable and argue that for suitably related

[6] Note that I take no stand on what 'a reason' is in isolation (a belief-desire pair, a consideration which counts in favour of something, a non-natural *sui generis* entity). I take it as obvious that many animals behave in a way that allows for certain sorts of explanations to be given (of roughly the form 'it wanted X, and saw Y as a means to get X, so it did Y'), and also that normal adult humans have a form of agency that goes significantly beyond this basic acting-for-reasons. How exactly to analyse or express these two ideas, or how in general to define 'reason' and 'rationality', I leave open.

subjects, the we-concept could perform these functions just as well. These objections all claim that selfless agents would be unable to perform vital functions – things that needed doing would not get done. The fifth objection works very differently, claiming that selfless agents just would not count as agents any more: they would have become mere bodies acted through by a single agent, which would be a 'group agent' in the sense of consisting of many human organisms, but not in the sense of consisting of many agents. This line of objection arises particularly from the view defended by Rovane (1998, 8ff) that agents are individuated by their 'rational point of view', the set of considerations they are committed to harmonising. If (as the following sections will show) selfless agents would have to be ones who shared a great deal of their rational points of view with other agents, their status as distinct agents would be in question.

### *3. Could selfless agents satisfy the knowledge condition on intentional action?*

It has been claimed that for someone to count as doing X *intentionally*, they must *know* that they are doing X (Anscombe 1963, 11–12, 49–57; Setiya 2008, 2009; cf. Paul 2009). If they are unaware of doing X, then they cannot be doing it intentionally. After all, if someone is asked for the reason for some action of theirs, the reply 'I didn't know I was doing that' is enough to dispel any requirement for a reason to be given, suggesting that the action was not done for any reason and thus plausibly was not intentional. Call this the 'knowledge condition' on intentional action.

This knowledge condition seems to immediately establish that any agent must have at least some singular self-knowledge: they must know that they are the person who is doing or has done certain things. If selfless agents do not know anything about themselves individually, then it seems they could not know that it was *them* doing anything. Thus whenever a putative selfless agent seems to do something, we can show that their 'action', because it violates the knowledge condition, is no action after all, and thus could not have been their action.

But this argument is far too quick. After all, it seems natural to describe dogs as acting intentionally, and as doing things for reasons (e.g. 'because they want food and think there's food in there'), but just from this we should not conclude that they must think about themselves as the doers of those actions. Clearly there can be agency without conceptual representations of oneself, and thus without beliefs of the form 'I am doing action X'. If there is a knowledge condition on intentional action, it must allow for this kind of non-rational, unselfconscious, agency. So it is actually not a simple matter to say what sort of 'knowledge' the knowledge condition requires, or even whether the condition is rightly interpreted

as being about *knowledge* at all, as opposed to belief, justified belief, cognitive access, or something else. And the soundness of the argument we are considering, against the possibility of selfless agency, depends on what reading we choose.

For example, here is one reading of 'knowledge' that does not rule out selfless agency: for an action to be intentional (under a description), it must be represented (under that description) by the agent in a way that allows it to be appropriately connected with reasons for action (and thereby with beliefs, desires, and so on). On this reading, what is crucial is not so much the presence of self-representations, but rather causal and informational integration among representations (this is very close to what Block 1995 calls 'access-consciousness'). And there is no obvious reason why representations other than conceptual, singular, self-knowledge could not accomplish this integration.[7]

Conversely, here is how we could read the knowledge condition to substantiate the original challenge to selfless agency: for an action to be intentional (under a description), it must be represented (under that description) by the agent *as their own action*, in a certain non-conceptual fashion which is such that, while this non-conceptual self-representation may remain non-conceptual in non-rational agents, it implies conceptual self-representation in rational agents. The idea would be that while non-rational agents lack the conceptual capability to entertain conceptual self-representations, rational agents have that capability, and so in them this non-conceptual self-representation will translate automatically into a conceptual self-representation, and thus into what I have termed 'singular self-knowledge'.

So there are ways of interpreting and elaborating the knowledge condition that would substantiate the original argument against selfless agency. But now the proponent of that argument needs to show why we should prefer that particular reading, especially since it seems to make the knowledge condition a much stronger claim than other available readings. I think that when we consider the reasoning that made the knowledge condition itself plausible, we will see that this reasoning does not give us reason to adopt the strong reading of that condition which would rule out selfless agency.

What originally motivated the knowledge condition was that answering the question 'why are you (singular) doing X?' with 'I didn't know I was doing X' serves to remove the demand for reasons. Thus if we ask Alfie 'why are you (singular) doing X?', and he answers 'I didn't know that it was *I*, and not Bettie, who was doing X', we might conclude that he was not doing X for any reason. But

[7] Many cognitive psychologists attribute our feelings of agency to a 'comparator mechanism' that discriminates 'our actions' from other perceived events by comparing perceptions of them with perceptual expectations based on 'efferent copies' of motor instructions. This mechanism seems to be at least a large part of what provides 'knowledge of our own actions', and seems to be clearly non-conceptual in its basic functioning (cf. Blakemore et al., 2002; David et al. 2008; Moore and Haggard 2008; Synofzik et al. 2008; Carruthers 2012; Frith 2012).

what if Alfie follows up with '… but whichever of us was doing X, we're doing it for the sake of Y'? Then it seems that X was intentional after all, and Alfie has enough knowledge of it to identify its reasons. Conversely, if he replied to 'why are you (singular or plural) doing X?' with 'we didn't know we were doing X', this removes the demand for reasons that would otherwise be satisfied by the answer 'we are doing X for the sake of Y.' Note that I am not here presupposing anything about the status of 'we are doing X for the sake of Y', whether it reports an irreducible we-mode intentional action or is merely a shorthand for a certain interlocking pattern of individual actions. Whether or not such a reduction is available, knowing the truth of the we-statement is compatible with ignorance of what *I* specifically am doing: knowing that I am part of a group whose interlocking actions constitute a joint action still doesn't tell me what role I am playing in the group.

So it seems that we can still distinguish intentional and unintentional actions, by reference to (some sort of) 'knowledge' of them, even when dealing with selfless agents. Thus the original motivation for a knowledge condition does not force us to read it in a way that requires individual self-knowledge. To find reasons to prefer such a reading, we will need some further argument, such as an argument that agents lacking singular self-knowledge would be impaired *qua* rational agents. And it is precisely such arguments that I have been considering in this and the next four sections.

## *4. Could selfless agents connect objective knowledge with action?*

Secondly, I-knowledge might be essential to connecting non-indexical knowledge with action. This is because action seems to have an essentially indexical dimension, and no amount of non-indexical knowledge can entail indexical conclusions without indexical premises (Lewis 1979; Perry 1979; Seager 1990). To use the classic example, Professor So-and-so might know that "Professor So-and-so's pants are on fire", and yet be entirely unable to connect this knowledge with the action of taking off their pants unless they also know that they are Professor So-and-so. Even if I know that removing Professor so-and-so's pants would be a very good idea, I cannot act on this unless I know these red pants here on *my* legs are the pants in question, and that these hands that could easily remove them are *my* hands. So how is a selfless agent, who does not know which hands are *their* hands, or which pants are on *their* legs, supposed to do anything?

Obviously the point is not that an agent without I-knowledge would be unable to act for reasons at all – after all, we can suppose that dogs do not have the I-concept, but we would still expect a dog to react to their pants being on fire.

But this is because perception, just like action, has pre-conceptual indexical content: when the dog feels or sees the flames, it perceives that *it's* pants are on fire (in some pre-conceptual sense), and thus acts to remove them. But human agents can learn through testimony and reasoning, thus accruing a wealth of non-indexical knowledge about the locations and histories and compositions of the various things and people in the world. To connect this knowledge with action seems to require the I-concept (as well as the now-concept and the here-concept, about which many analogous things could be said). An agent would be impaired if this knowledge were cut off from action, and it seems that a selfless agent would have just that problem.

The fact that action requires indexical knowledge does not automatically imply that it requires any particular indexical term: 'we' is an indexical, and so is 'this'. So there is no *direct* argument from the convincingly established principle that non-indexical premises cannot yield indexical conclusions, to any problem for selfless agents. Presumably, in the minds of selfless agents, actions are coded as 'ours', i.e. they are set up so that actions result from practical conclusions, including those based on objective knowledge, that contain the we-concept, just as we are set up so that actions result from practical conclusions using the I-concept.

But this gives rise to the following possibility: one agent (say, Bettie) reaches some practical conclusion (say, 'let's remove these pants on (two of) our legs, because they are on fire') whose enactment would require the body of *another* agent (say, Alfie, who is wearing said pants). What will happen in this situation? The agents would clearly be impaired *qua* rational agents if this practical conclusion simply sat inertly there, producing no action simply because it was reached in the wrong head. So Bettie's resolution to act must be efficacious, which means it must produce an action of Alfie's body. For individually self-knowing agents like us, the natural way to do this is mediately, by asking or coercing or persuading the other agent into action. Yet Bettie cannot employ this solution, for how would she know to do so, if she did not know whether she was Alfie or Bettie? So she must act not 'on' but 'through' Alfie. Her 'basic action' must be 'removing these pants', which will as a matter of fact involve somehow acting on Alfie's body to make his muscles move.[8] But the mechanism by which the muscular movement is produced is invisible to Bettie in just the same way that the mechanism by which we usually reach out with our own arms is invisible to us.

We can put this by saying that selfless agents must be, as I will say, 'motor vulnerable' to each other. A first gloss on this notion is that an agent is motor

[8] A 'basic action' is often defined as an action A for which there is no other action which the agent performs in order to A, but it can be disputed whether there are such basic actions (cf. Lavin 2012). Thus I use a weaker definition: a basic action is an action A that is not preceded by any deliberation about how to A.

vulnerable to another agent if the second agent can move the first's body as a basic action. A more refined gloss, without the restriction to bodily movements, is that some of those events which the vulnerable agent can produce as basic actions are also producible as basic actions of the other agent. If two agents are completely mutually motor vulnerable, then they will have the same set of basic actions. And this, it seems, is what would need to be true of the selfless agents whose rational agency relied on the we-concept instead of the I-concept.

Wouldn't this just make the first agent's body count as (part of) the second agent's body? I am agnostic about the conditions for something to be 'my body'. If direct motor control were all that is needed, then motor vulnerability might imply a sort of 'body-sharing': the two agents' bodies overlap. But plausibly historical and biological considerations are also important, and perhaps the relative ease of removal. When someone acquires a prosthesis that they can control very skilfully, there may be a sense in which we can call it 'part of their body' (it 'feels like' part of their body; cf. Holmes and Spence 2005, 40–45; Imaizumi et al. 2016), but there is clearly also a sense in which we can distinguish it from 'their body', something which grew organically from a shared biological origin. In that sense, simply making two agents motor vulnerable to one another would not obliterate the boundary between their bodies.

Is motor vulnerability possible? It seems very difficult, perhaps impossible, for ordinary human beings, for there is no reliable causal route by which one human being can produce movements in another person's muscles, except by going through the other's sense organs. And human beings can generally only affect the sense organs of another by a more basic action, the contraction of their own muscles. But that does not mean that these restrictions are essential to rational agency: rather, they are contingent anatomical limitations of us thick-skulled beings. When we consider what can be done by contemporary neuroscience, it becomes clear that they are not inevitable.

Perceptual states produce actions by causing certain patterns of brain activity, and patterns of brain activity can be exogenously produced by the stimulation of the brain using electrodes, implants or other devices – as when experimenters make patients move their arm, or make them feel as though they have decided to move, even while no movement occurs (Fried et al. 1991; Desmurget et al. 2009). Thus with the right tools, there is no reason in principle why we cannot induce an action that uses someone else's muscles without any intervening perception. On the other side, we can construct devices which detect brain activity and are triggered by it to produce some effect, such as the flexing of a prosthetic limb or the movement of a dot on a computer screen (e.g., Grübler and Hildt 2014). So there is no reason in principle why an agent's basic actions need only be muscle contractions. Putting these points together, there is no reason in principle why one agent might not, automatically or as a basic action, cause movements

of the muscles of another. Thus there is no reason in principle why two agents could not be mutually motor vulnerable, as I have defined these terms.

So our example humans, Alfie and Bettie, should go in for some form of surgery, some sort of neural implant that can automatically detect certain of their brain states (or equivalently, can detect certain features of their complex overall brain state), and which can then automatically signal to the other's implant to induce some state in the other brain's motor areas. This would be functionally similar, in many ways, to structures like the corpus callosum which respond to brain activity in one hemisphere and induce it in the other. The implants likely will not establish motor vulnerability by themselves; they will simply remove the anatomical barriers to this relation. Actually establishing it would build on the participants' existing willingness to act on each other's requests or suggestions, using technology simply to allow for habitual forms of these practices to become genuinely automatic, without the requirement of intervening agentive steps involving muscle contractions and perceptions thereof.

Note that it is only in a very weak sense that the brain-to-brain interactions need to be 'automatic' for Alfie and Bettie to be mutually motor vulnerable. They simply need to take effect without a distinct, prior, agential decision – without the agents needing to decide to do something *in order to* make the other's body move. This essentially negative sort of automaticity is compatible with the interaction being slow or unreliable, or with each finding it more difficult to move the other's body than their own, just as some of our basic muscular actions are harder to pull off successfully than others.

I think it is clear that with communicating neural implants of the sort just discussed, Alfie and Bettie could become mutually motor vulnerable. Is there any way that they could establish this relation without such science-fictional devices? This would require a causal sequence from one's brain, through their muscles, to the other's sense organs, and into the latter's muscles, which was 'automatic' in the relevant weak sense: not involving any prior deliberation. Perhaps two people who were sufficiently 'in sync' might work like this: the one could direct the other without thinking about what they were doing with their own body to direct them. Perhaps when one frowns at someone, the other punches that person without hesitation, such that the first doesn't even have to think 'that person needs a punch, so I'll frown at them *in order to* get my partner to punch them'. Instead, the first just thinks 'that person needs a punch, so let's punch them', not bothering with whether their role in this is to frown or to throw the punch directly. (A precursor to this sort of rapport, for which there may be empirical evidence, would be for the first to perceive affordances based on the capacities of the other's body; cf. Gallotti and Frith 2013.) But it is hard to see how such a situation could be stable for *all* of the actions these two perform – at some point, it seems, they would need to think in terms of acting on the other (e.g., to give instructions more complicated than a frown).

Motor vulnerability might be necessarily a partial and temporary matter when it comes to present-day humans.[9]

Whether or not complete mutual motor vulnerability requires neurosurgery, the point is that there is nothing metaphysically privileged about the nerve and muscle fibres that connect our practical conclusions with our basic actions, and no reason why the right kind of causal chain could not run through the body of another agent, and even overlap with the causal chains that connected their practical conclusions with their basic actions.

## 5. *Could selfless agents understand their own reasoning?*

Thirdly, Tyler Burge (2000) argues that we could not reflectively understand our own reasoning without the I-concept. Burge points out that in order to reason, someone must not only recognise rational evaluations of beliefs, but must also generally amend their own beliefs in light of their own such evaluations, and do so 'immediately'. The immediacy of this amending contrasts with the way that a reasoner implements their rational evaluations of other people's beliefs: when I realise that someone else has made a mistake, I must take steps to change their mind, but when I realise that I have made a mistake, I just change my mind directly. Thus, to understand reasoning we must have a concept that "marks the … attitudes where a rational evaluation of the … attitude immediately rationally requires using that evaluation to change or maintain the attitude" (2000, 253), thereby separating such cases from those where the implementation of such an evaluation is only mediate. He claims, plausibly, that in actual human thought this role is played by the I-concept, because the cases calling for immediate implementation are just those in which both the evaluation and the belief belong to the same subject.

Note that the objection is not that one cannot *reason* without the I-concept, for the concept merely marks this distinction rather than establishing it. Yet it is necessary for the *understanding* of reasoning: a being that lacked this concept could not understand how its own reasoning worked, and thus would be unable to reason in a reflective way. Thus a selfless agent would, by Burge's lights, be incapable of reflective critical reasoning.

When we ordinary agents reason critically we must keep track of which reasoning is 'mine' and which is someone else's; selfless agents cannot do this, but can only keep track of which reasoning is 'ours' and which is not. For the we-concept to serve this function adequately, it would have to be that the cases

[9] Note that if motor vulnerability without neurosurgery is even coherent, that provides further reason to think that motor vulnerable agents' bodies do not overlap: here we would have two unaltered human bodies whose relationship had changed at an agential level, but were completely unchanged at a physiological level.

calling for immediate implementation of rational evaluations are just those in which both the evaluation and the belief belong to some members of a certain set of related subjects. Could there be subjects for which this is true?

Burge identifies two reasons why implementation of evaluations on the beliefs of others must be mediate: first, that our evidence may differ, so that "What may be a reasonable evaluation by … A of an attitude held by … B may not be a reasonable evaluation for B" (2000, 254), and second, that we always employ some means to do so: "the question of how one is to bring about any alteration must inevitably arise. One cannot simply alter the thought … with no intervening practical premisses" (2000, 253). However, while these points hold for actual humans, there could be agents for which they did not.

Consider first the worry that since agents may have different evidence available, a verdict reached by one might be unreasonable for the other to immediately adopt. This problem is avoided if the two agents are always sufficiently aware of what evidence the other has available – where 'being sufficiently aware of' evidence need not require attentively thinking about it, nor grasping every detail of it, but merely that the overall 'gist' of the evidence be poised to inform judgements that the agent makes. That is, an agent's 'sufficient awareness' of some evidence needs only to preclude cases where an evaluation based on that evidence cannot be rationally accepted on account of the agent not having the evidence that would justify it. I take it that this standard is met when, for instance, two detectives working different aspects of a case keep each other updated with summaries of relevant things they have each learnt. This can be so, even if much of the detail of what each has learnt is left out as irrelevant: each still has sufficient access to the total body of evidence to recognise the applicability of the other's judgements based on it.

However, selfless agency would require a more *immediate* pooling of evidence than is involved with detectives keeping each other updated with summaries of what they have each learnt. For in the latter case each agent must decide to provide the other with a summary, i.e. must act, and thus rely on practical reasoning. Rather, our selfless agents would need to be related so that each was aware of the other's evidence *automatically*, in the same weak sense of automaticity employed in the previous section, simply in virtue of the other having this evidence. Call agents related in such a way 'evidentially unified'.

Next consider the worry that agents cannot directly act on each other's beliefs. More precisely, whenever a normal human acts on another human's beliefs, the action is mediated both in how it affects the person acted on – it must first produce a perceptual state, which in turn affects belief – and in how it comes from the agent – it is something they do *by intentionally doing something else*, and not a basic action. Our selfless agents, then, would have to differ in that they could 'persuade each other' of things without use of any intervening perceptual state, and as a 'basic action', not preceded by any deliberation about how to do it. This

could still be something relevantly like persuasion as long as it appealed to the other's reason, rather than relying on coercion, rhetoric or other non-rational means. That is, it would still be persuasion if it functioned so that the belief produced in the other agent would take hold only as long as the other agent remained satisfied with the support for it. Say that an agent that can be acted on by another in such a way is 'cognitively vulnerable' to the other.

Are evidential unity and cognitive vulnerability possible? I see no reason to think them impossible, for the same reasons as I cited in the previous section for not thinking motor vulnerability impossible. It is simply a question of physiological engineering. Perhaps at present, human beings can only convey evidence to another, or amend their beliefs, by first deciding to contract their own muscles, in order to produce a perceptual state in their target, which will then cause a change in their beliefs. But since perceptions change beliefs by delivering electrical signals to the brain, devices by which one person's brain activity can directly deliver such electrical signals to another's brain would allow for one agent, automatically or as a basic action, to create direct changes in the evidence or beliefs of another. Thus there is no reason in principle why two agents could not be evidentially unified and mutually cognitively vulnerable.

Indeed, just as it may be possible for real-world people who are sufficiently attuned to one another to be temporarily motor vulnerable, so it might be possible for real-world people who are sufficiently good at reading each other, and who are in close enough proximity to do so, to become temporarily evidentially unified, at least on some restricted topic, or even cognitively vulnerable to one another. Consider two poker players who are both very good at reading other players, but very bad at bluffing: they might find it so natural to infer each other's hands from watching each other's faces that, on the topic of what cards are in their hands, they have equivalent evidence as each other even without either deciding to convey it to the other. Cognitive vulnerability is harder to envisage: to count as intentionally acting to amend someone else's belief, one must intend to amend that particular belief, which will typically require noting the self-other distinction and employing some deliberate means to affect the other. Even if cognitive vulnerability is possible among present-day humans, it seems very unlikely (just as with motor vulnerability) that such a close rapport could persist for very long, or cover very many topics.

This means that Alfie and Bettie will have a further use for their neural implants: establishing a richer, more systematic sort of evidential unity and cognitive vulnerability than is normally possible. As before, the implants would not themselves establish cognitive vulnerability or evidential unity; they will simply remove the anatomical barriers to those relations. Alfie and Bettie would then have to build on their existing co-operative practices of evidence-sharing and rational persuasion, until these become as automatic and involuntary as the cognitive interactions among states of a normal individual's brain. They may of

course still disagree on things – two people can reach different conclusions based on the same evidence, and although mutual cognitive vulnerability allows each to try and directly change the other's mind, it does not guarantee success, or guarantee that the other will not be changing your mind at the same time. But their disagreements will have the fluid and volatile character of a person's 'disagreements with themselves', rather than the fixed and persistent character of many inter-person disagreements.

## 6. *Could selfless agents identify the right reasons for them?*

Fourthly, I-knowledge seems essential to knowing which reasons one has, or equivalently which factors one has reason to respond to. For instance, if I do not know who I am, I do not know whose wellbeing is my wellbeing, and so if each of us has reason to promote their own wellbeing, I will not be able to do this effectively. I may know that an action would bring about an outcome where A gains and B loses, but be ignorant of whether this is a good outcome for me or a bad one.

Obviously this difficulty is heavily dependent on what the right theory about our reasons is. Some theories, such as classical utilitarianism, are 'impartial' in that they hold all agents to have the same practical reasons – e.g. reason to promote the greatest happiness of the greatest number, reason to promote the flourishing of all sentient beings, or similar.[10] Given a certain set of beliefs and available actions, then, an agent need not know who they are in order to know which actions they have reason to perform. So on some views, the function supposedly played by the I-concept turns out to be superfluous.

But let us suppose for the sake of argument that there are significant differences between agents in what ultimate practical reasons they have. I, for instance, might have special reason to promote my own wellbeing in the future, stronger than my reasons to promote other people's. Moreover, I might have special reason to promote the wellbeing of my friends and family, and not that of other people's friends and family. We might also think that I am subject to special restrictions on

[10] Of course, sameness in ultimate reasons is compatible with a derivative difference in reasons – if I am well-placed to promote A's good and you are well-placed to promote B's good, a utilitarian can regard us as having good reason to focus on these different goals. But here I am concerned only with ultimate reasons: for consideration of the way that self-knowledge affects derivative goals, see section 4.

Note that although the objections considered in sections 4 and 6 both involve indexical knowledge (of how to act, or of what reasons are one's own), they are distinct because the knowledge in question relates differently to action: the former objection concerns the need for some sort of indexical knowledge in order for action to happen at all, but the latter is about specific items of knowledge which bear on which particular action is most reasonable.

how I may treat others (e.g., not sacrificing their major interests, like life or physical liberty, for the sake of a moderate benefit to many others) but not in how I may treat myself, or that I have reason to keep my own promises and not those made by others. If this is true, then knowing who I am will be crucial to knowing whose interests to promote, who to treat in special ways and which promises to keep.

However, I believe that for suitably related subjects, the we-concept can serve this function quite adequately. For observe that the very same intuitions which support ascribing different reasons to different subjects also tend to support allowing for the character of two subject's relationship to modulate the significance of the self–other distinction in their case. This is true in at least two major ways. Firstly, certain sorts of 'close' relationship, like friendship or parenthood, increase the strength of a subject's reasons for promoting the wellbeing of certain others, reducing whatever difference of strength we might think there is between egoistic and altruistic reasons.[11] At the extreme point, this would make egoistic and altruistic reasons equal in strength, and then the strength of each subject's reasons for promoting wellbeing would be independent of whether it was their own or the other's.[12] Thus they would not need to know whose wellbeing it was that a given action would promote, as long as they knew that it was *either* theirs or the other's, and knowledge involving the we-concept allows them to do this.

Second, acts such as consent, contract and promising create a relationship between the parties in which usual restrictions or permissions may be suspended. We can, for instance, authorise others to make promises or deals on our behalf, and will then be bound by them just as much as they are, as when someone hires a negotiator or appoints an ambassador for a negotiation too complex, or too far away, for them to enter into directly.[13] For another example, suppose that one

[11] There is also some plausibility in the thought that people's relations to others in their culture or community can induce convergence of moral reasons, or even modulate the strength of moral reasons according to the norms prevailing in that culture or community. This thought might be supported by the empirical evidence of moderate systematic cross-cultural differences in moral intuitions (e.g. Triandis 1990; Bersoff and Miller 1993; Keller et al. 2005), though that evidence of course admits of other, less relativistic, interpretations (cf. Dworkin 1988, 200ff).

[12] Isn't it still morally praiseworthy to benefit a friend, but morally neutral to benefit oneself? Yes, but this difference also diminishes as we consider closer and closer friends: we praise disinterested altruism more than looking out for one's own, because it is psychologically rarer and harder, and altruistic concern for close friends is often a powerful temptation that must be resisted, just like egoistic concern is (it threatens 'favoritism' rather than 'selfishness').

[13] The structure of promissory obligations in such a case is hard to analyse exactly – if X appoints Y to negotiate on their behalf with Z, they plausibly are obligated to adhere to the terms of any deal that Y and Z make. But is this because X has a promissory obligation to Y (as part of the latter's appointment), or directly to Z (as part of their conveying to Z that Y speaks on their behalf), or indirectly to Z (created only by Y's promissory act of agreeing to the deal), or all three? Fortunately, it does not matter to my argument how we analyse the case, since it seems clearly true that people can in fact appoint others to deal and promise on their behalf.

normally may sacrifice one's own major interests (life, health, physical liberty, etc.) to secure a greater aggregate benefit for others, but may not do so with another person's major interests. Yet if one forms an agreement with someone else, permitting both to make such a sacrifice if it becomes necessary, and if we consider them competent in doing so, then this self–other distinction is removed.[14] Thus to know whether one had a special overriding reason not to perform an action that would sacrifice one person's major interests for the greater good, one would not then need to know specifically whether that person was *oneself*, but only whether they were a member of the group formed by oneself and those who were parties to such an agreement: one of 'us'. And so knowledge involving the we-concept might serve just as well as knowledge involving the I-concept.

Putting these two points together, we get the following result: knowledge involving the I-concept is dispensable in favour of knowledge involving the we-concept *if* the people referred to by that 'we' are related in a way that is both very very 'close', in the sense in which friendship and family are relationships of 'closeness', but exceeding those in degree, and also are parties to a series of contracts or agreements in which they waive the special moral protections against each other that their distinctness affords them. Call such a relationship, if it is possible, 'moral unity'. It is hard to say what moral unity would require – it might be something that human beings are generally incapable of, requiring each party to be more deeply and intimately familiar with and committed to another person than we are in even our most devoted and loving relationships. Or perhaps close relationships that human beings already do stand in qualify as moral unity: perhaps all that Alfie and Bettie need to do to establish moral unity is have a fairly standard sort of 'commitment ceremony', or perhaps a ceremony with vows much more extreme and explicitly-formulated than is customary. And if classical utilitarianism is true, perhaps we are all already morally unified. This is essentially an ethical question, about what reasons people in or out of relationships have.

One might worry that this requirement would make selfless agency depend on previous, individually-self-knowing agency. For if moral unity requires some set of 'promises' to be made, must these not ultimately rest on a sort of self-knowledge? Must not each participant know who they are and who they are

[14] Can such an agreement be valid? It seems right that certain agreements are always legally void – agreements to be someone's slave, to allow another to violently abuse you at will, and so on. Yet it also seems that certain agreements can subject people to major harms and still be valid – contracting to go to war or attempt a near-suicidal rescue if called upon, instructing others not to resuscitate you if you become permanently vegetative, agreeing to medical experimentation or a bone marrow transplant, etc. I do not know what makes the difference, but it seems to me that an agreement made equally by all parties (i.e., not subjecting one person to another) and which serves some significant goal (i.e., not made frivolously or recreationally) will generally be valid even if it subjects members to the risk of serious harms.

making their promise to? If they did not, this might seem to invalidate their promises, like contracts signed under coercive pressure or fraud. This would not show that selfless agency is impossible, but it would show that (if the true theory of reasons doesn't turn out to be wholly impartial) the only way to establish selfless agency would be to first establish individually-self-knowing agency (as Alfie and Bettie do). Thus the Indispensability Thesis would not be true, but something related to it would be – that rational agents must have individual self-knowledge at some stage in their lives.

I believe even this weaker thesis is false. Let us consider two beings who are to become selfless agents without going through any prior individually-self-knowing stage – perhaps a pair of artificial intelligences, called Alphex and Betex. Could they establish moral unity – could they, for instance, make promises to each other regarding their future behaviour? I will take promises to be part of a broad class of acts by which people seek either to give assurance about their future behaviour to someone who wants that assurance, by means of resolving to behave that way and expressing this resolution (cf. Scanlon 1998, 304), or to invite others to trust them to behave in a certain way, i.e. to act out of a regard for the importance that behaviour has to one who trusts (cf. Southwood and Friedrich 2011).

So suppose that Alphex wishes for assurance about Betex's future behaviour, specifically about whether Betex will honour promises made by Alphex on its behalf. Suppose, moreover, that both Alphex and Betex know this, though neither knows whether they themselves are Alphex or Betex. Both might wish to provide that assurance (and to invite and live up to the corresponding trust), but Alphex at least is not able to – no act it performed could count as promising Alphex that Betex will do something. What they can both do, however, is promise to the other that (i.e., form and express a resolution that, invite Alphex to trust that) *if* they are Betex, they will act a certain way (e.g., honour promises made by Alphex on Betex's behalf). One of these promises will be empty, though neither will know whether it is theirs.

The key question is: does the emptiness of Alphex's promise invalidate Betex's, just because Betex cannot know whether its promise is the empty one? I will admit that I am not sure. But if Betex's promise is invalidated, surely it is not *wholly* invalidated. It is still a morally significant act of some sort, even if an imperfect one. We might say that is not like a contract extorted by force, but rather like a contract made by a 16-year-old, or a moderately tipsy person. Some legal regimes might accept such contracts (or some subclass of them), and others might not, without either regime being clearly wrong, because the case has only partial moral force, and the law requires a yes-or-no threshold, and it is indeterminate where exactly that threshold should be.

Even partial moral validity makes a difference, if it leaves Alphex even a little more entitled than it was before to make promises on Betex's behalf. For this

means that the next time they make the same promise to each other, Alphex's promise will not be entirely empty, and so will not invalidate Betex's promise as much as before. This means that after the second round of promises, Betex's promise will be more valid, and so Alphex will be more entitled than before to make promises on Betex's behalf. Successive rounds of such imperfect mutual commitments might thus eventually establish a fully valid mutual commitment, and thus the moral preconditions for selfless agency.

This may sound like an odd and contorted sort of process, but it is not that different from the way human children become rational agents. They must go through the experience of choosing, acting and taking responsibility, but will inevitably do so imperfectly at first. Yet their imperfect efforts at agency are enough to make the next effort a little less imperfect, and so on until adulthood. Our own rational agency cannot arise fully-formed; we should not expect selfless agency to do so.

## *7. Would selfless agents still exist as individuals?*

What about the concern that selfless agents would supersede themselves, vanishing into a group agent with no members? After all, the relations that would have to obtain among selfless agents using coreferential 'we's seem to be, in many respects, like those between the lobes or hemispheres of a single human brain. I would like to use this fact to make us more inclined to take seriously the possibility of lobes or hemispheres being rational agents, but a critic might equally well use this fact to cast doubt on the possibility of selfless agency. Let us distinguish two steps in this objection:

1. When there is selfless agency, the selfless agents compose a group agent.
2. If selfless agents compose a group agent, they are not individual agents.

I am happy to accept the first of these two claims: wherever there are selfless agents using coreferential 'we's, the set of them all will constitute a group agent. But I want to reject the second – the presence of this group agent does not deprive the individual members of their own status as agents. On the contrary, cases of selfless agency exhibit a very close connection between two layers of agency, neither of which undermines or threatens the other.

Let me first say a bit to motivate claim 1. When there is a group of selfless agents who use the we-concept to refer to that group, the relations among these agents ensure the existence of the following things: a shared total evidential state (because the agents are evidentially unified), a shared total set of applicable reasons for action (because the agents are morally unified), a shared total set of

available actions (because the agents are mutually motor vulnerable), and a shared total set of cognitive states which can be immediately adjusted in light of rational conclusions based on the others (because the agents are mutually cognitively vulnerable). These make it easy to describe and explain the group's behaviour by appealing to the way that this shared cognitive process identifies the particular action from among the shared set of available actions which, according to the shared evidential state, best responds to the shared set of reasons. And this is paradigmatically agential.

Philosophical defences of group agency typically proceed by first showing the *prima facie* reasonableness of treating groups as agents, and then arguing that the agency involved is in some sense 'irreducible' to the agency of individual members (e.g. Pettit and List 2011, 5ff; Huebner 2014, 126ff). To follow this style, I would next try to show that the appearance of agency in a group of selfless agents could not be properly accounted for by any description in terms of individual agents. But I do not believe this: the individuals share the agential features (set of reasons, set of actions, etc.) that characterise the group agent. Their selflessness means precisely that their own agential perspective 'opens outwards' to include everything that the group's does – to include everything they can call 'ours'.

For example, here is a story told at the level of individuals: Bettie thinks about getting bread, and concludes 'we don't need bread', only to have this judgement immediately amended by Alfie's rational evaluation of it as mistaken, into 'actually we do need bread'. Since it is Alfie who is near the store, this conclusion leads to the further conclusion 'Alfie should get bread' (or perhaps better, 'we should get bread with the Alfie body'). This conclusion is in fact reached by Bettie, but directly activates Alfie's motor cortex so that he walks towards the bread aisle. We could re-tell the same story at the level of 'the couple': it thinks about getting bread, first judges that it doesn't need any, then changes its mind, and decides to get bread using its Alfie-part, which it consequently moves towards the bread aisle. These two stories are describing the very same events, and each could be readily derived from the other. But this leads directly into the objection we are here considering: if the individual agents and the group agent share so much of their agency, is it really appropriate to count them all as existing? Haven't we actually reduced the number of agents present?

One way to press this concern would focus on the meanings of pronouns. Suppose that some English-speaking selfless agents utter the word 'we' from many individual mouths (or think it in many individual brains): this word functions in many ways like the word 'I' would if uttered by a single agent with many mouths (or many brains). So perhaps we should say that really it is an 'I', which for contingent historical reasons happens to have carried over the sound and spelling of the word 'we'. Perhaps the word 'we' now expresses the I-concept.

However, we can both agree that the word 'we' now expresses the I-concept, and maintain that it continues to express the we-concept. To see how this is possible, consider this figure, taken from Heller (2000):

| This | sentence | contains | exactly | six | words. |
|---|---|---|---|---|---|
| sentence | | | | | |
| contains | | | | | |
| exactly | | | | | |
| three | | | | | |
| words. | | | | | |

Heller writes that "the across-sentence is true. The down-sentence is false. The single token of 'this' has two contents, referring to both the across-sentence and the down-sentence" (2000, 376). This is possible because its meaning is context-sensitive, and it occurs in two contexts simultaneously, allowing for two assignments of meaning. I claim that when a group agent says or thinks something, these token utterances or thoughts are simultaneously uttered or thought by one or more individual agents. Of course it is a significant assumption that thoughts, like utterances, can coherently belong to more than one agent; in other work I have argued at greater length for the more general thesis that mental states, including not just thoughts but conscious experiences, are shareable within certain limits (Roelofs Forthcoming-b). Not wanting to inflate this paper with that discussion, let us suppose that thoughts can have two thinkers, and ask how that would resolve the worry about 'we' having come to mean 'I'.

Being uttered or thought by both individual and group provides two contexts to look to in assigning meaning to the utterance or thought, and if particular elements thereof are context-sensitive, they may come to have two distinct meanings relative to the different agents. This applies, I believe, particularly in the case of first-person pronouns. When used to refer to the group, their meaning will be the 'we-concept' relative to the individual member using them, but will be the 'I-concept' relative to the group using them.

This follows from how I have defined the I-concept: a being which refers to something non-identical with itself is not using the I-concept. Thus if the individual agents are using a concept to refer to the whole group of them, a group with which they are not individually identical, they cannot be using the I-concept. But the group agent, which is simultaneously the subject of the same referential

act, is referring strictly to itself, and doing so independently of any objective specification of its properties. Thus it is using the I-concept, while they are not.

The more fundamental question, though, is not what words mean but whether the right way to individuate agents will allow for the co-existence of group and individuals in cases like this. If the individuals share with the group, and with each other, whatever it is that individuates agents, then they will be (or support, or constitute, or realise) the same agent as it, and as each other. This might also mean that there is only one *person* – that Alfie and Bettie, in our example, are no more and have ceased to exist when they succeed in becoming selfless. Alternatively, it might mean that while Alfie and Bettie, the people, have persisted, they have lost their status as agents, just as a human person might continue to exist but lose their agency by suffering massive brain damage.

The most explicit defender of a criterion that would disqualify selfless agents from being agents is Rovane (1998, 2012, 2014a, 2014b), for whom agents are individuated by their 'rational point-of-view', the perspective from which they reason and act. Many other theorists say things that could be taken to disqualify selfless agents – indeed any 'psychological theory' of personal identity might have this implication (see, e.g., Shoemaker 1970; Parfit 1984; Noonan 2003). But for most of these theorists there are hard interpretive questions that would need to be answered, about how psychological continuity relates to consciousness, about how diachronic identity criteria relate to synchronic individuation criteria, and about what kinds of causal mechanisms need to bring about psychological relations for those relations to establish identity. To side-step these questions, I will focus on Rovane. She characterises a rational point of view as a set of intentional episodes which include a commitment to some project which requires many activities coordinated either at a time or across time, where that coordination is enabled by the rational relations among the intentional episodes in the set, and where this 'unifying project' brings with it "a commitment to achieving overall rational unity within the set" (Rovane 1998, 64).

This might not be the only way to characterise the relevant sort of point-of-view: as well as intentional episodes like beliefs and intentions, we might also appeal to sets of objective reasons, or to sets of available basic actions, etc. But however exactly we characterise rational points-of-view, it is clear that selfless agents like Alfie and Bettie are good candidates for people who share one. Because they are mutually motor vulnerable, there is a single set of available basic actions for the pair, and for Alfie, and for Bettie. Because they are both evidentially and morally unified, there is a single set of reasons bearing on the belief and action of the pair, and of Alfie, and of Bettie. And because they are evidentially unified and cognitively vulnerable, there is a single set of 'intentional episodes' (thoughts, plans, decisions, etc.) whose rational relations can govern the activities of the pair, and of Alfie, and of Bettie. So it seems likely that there is only one rational

point-of-view here. We can set aside the question of whether Alfie and Bettie themselves have ceased to exist (being essentially agents), or whether they persist as something less than agents; either way, the charge is that there remains only one agent.

One option would be simply to deny this criterion of individuation, and appeal to living bodies, or human organisms, or even streams of consciousness (assuming that Alfie and Bettie remain conscious). There is something plausible about the thought that I myself am both a rational agent and a human animal, and that we should individuate agents by first individuating things (e.g., organisms) which are candidates for being agents, and then evaluating them by some criterion of agency. Then we would say, for instance, that because there are two human beings, and both are capable of rational thought and action, we have two rational agents (plus any group agents). But if we wanted a more conciliatory response to this line of objection, we might try accepting the Rovanian view, but then arguing that individual agents can still co-exist with group agents, by having rational points-of-view which are component parts of its point-of-view (for fuller development of this idea, see Roelofs Forthcoming-b).

This notion of component and composite points-of-view is unfamiliar, but easy to sketch in outline: whatever the important relations among the elements of a rational point-of-view (dispositions to causally interact, for simultaneous thoughts; rationally conferring motivation on, for reasons and available actions, etc.), they are likely to often be matters of degree, and so one subset of the elements of a point-of-view might be more tightly connected to each other than they are to the other elements. They would then qualify as a point-of-view in a slightly stronger sense than the more encompassing set – not enough to warrant demoting it from its status as a point-of-view, but enough to warrant recognising them as well as it.

Component points-of-view make sense especially in light of the various forms of conflict that remain possible between selfless agents. Cognitive and motor vulnerability ensure that selfless agents sharing a 'we' have a single set of basic actions available, and a single set of beliefs they can directly adjust, but do not ensure that they will always attempt the same actions, or accept each other's belief-revisions. And moral and evidential unity ensure that their actions and beliefs are accountable to a single set of reasons, but do not ensure that they will reach the same practical and theoretical conclusions based on those reasons. Understanding such conflicts as still arise will be easier if we can talk about distinct rational points-of-view that tend to produce the two conflicting viewpoints.

For an example, consider Alfie and Bettie with their communicating implants. These implants allow for mutual cognitive vulnerability and evidential unity, which is to say they allow all intentional states across both brains to bear a certain relation to each other: that of potentially directly influencing each other in a certain rational way. But communication across the implants may still be slower and less

reliable than communication within a brain, making the states in a single brain slightly more functionally integrated than they are with states in the other brain.[15] Information stored in one brain is accessible to processes in the other, but probably not as *readily* accessible. Moreover, the two brains might differ in how well, or how quickly, or how thoroughly, they consulted different subsets of the accessible information. If Alfie prioritises speed and Bettie prioritises certainty, they might often disagree, with Alfie jumping quickly to an erroneous conclusion.

It is also likely that, for instance, Alfie can amend Bettie's thoughts or move Bettie's body with more difficulty, more slowly, less reliably, than Bettie can. As a result, when they do reach different conclusions about what to do or think, the correction of this conflict may be slow, difficult, or even impossible: there may be a shorter or longer period during which one is trying to move the other's body but failing because the other is moving it more forcefully, or during which both succeed in inhibiting or reversing the other's efforts.

By focusing on the closer integration of processes in each brain, we might identify two 'clusters' of intentional states which are interlinked in (some of) the ways constitutive of a rational point of view, and which are so interlinked to a higher degree than the whole set of states across both brains. We could then reasonably regard those two clusters as rational points-of-view, and thus as belonging to two distinct agents, each part of the group agent.

This proposal to recognise agentially-integrated subclusters as 'component agents' might seem rather cheap and sophistical. Even if it is internally consistent, surely this just isn't the way we tend to talk – in particular, isn't it clear that our actual practice is to refuse to count something as an agent whenever it is entirely contained within a more encompassing agential system? That is, we seem to impose a 'maximality' condition on our term 'agent', where only the 'largest' of any set of overlapping agential systems is regarded as an agent (cf. Sider 2001). I have argued that we could talk differently, without the maximality condition, but why should we adopt some deviant dialect in preference to our existing practices?

It is at this point that we should recall the thesis of Agentive Anti-Homuncularism, introduced in section 2. That thesis said that we could rule out a whole class of theories about our own nature – any on which we are group agents, composed of subsystems which are literally rational agents in their own right. This is a significant claim, and seems to reflect some deep fact about us. But if my arguments so far are correct, then there is no deep conceptual, metaphysical or empirical basis for such a confident rejection of 'homuncular

[15] To the extent that all information is redundantly stored in multiple platforms, these clustering effects might be reduced or eliminated. But to that extent we might also begin to wonder if we now have multiple implementations of the same abstract point-of-view-type. The question of what happens, identity-wise, to a person who is exactly duplicated is a tricky one.

realism'. Of course, I have conducted my discussion in terms of 'rational agents', and homuncular analyses of human cognition might not require the homunculi to be *rational* agents but merely agents: I take it that if there is no principled basis for denying that they can be rational agents, there is *a fortiori* no principled basis for denying that they can be agents.

If we continue to accept Agentive Anti-Homuncularism, simply on the basis of the maximality condition imposed by everyday language, despite the availability of equally consistent other ways of talking, we will be ruling out a certain way of understanding ourselves, a certain way of making theoretical sense of the structures we find to underlie our own agency, on essentially semantic grounds. That seems to me to be something we should avoid doing.

That is not to say that the whole question of this paper turns on a semantic point. As noted above, there are plenty of ways of individuating agents which do not allow the Rovanian challenge to proceed, such as by bodies, brains or streams of consciousness. It is a substantive claim about the metaphysics of agency that agents are individuated quite independently of these factors. But the claim that a being which is intrinsically just like an agent, which can perform all the functions of agency, should not be called an agent if it is contained within a more encompassing agential system, is a semantic stipulation which can easily be done without.[16]

The Rovanian might at this point make the following reply: what we recognise as an agent should depend on what our aims are in recognising agents – what purpose does the concept serve for us? One crucial purpose is a social and ethical one, of identifying beings that can be addressed and inter-personally engaged with through "relations in which one agent attempts to influence another and yet aims not to hinder the other's agency" (Rovane 1998, 5). And for this purpose, it is no use to recognise component points-of-view: we cannot engage Alfie without Bettie being involved, or vice versa, for anything we say to one will be heard by both, and any response will be formulated by both.

But this social-ethical purpose, though important, is not the only purpose served by the concept of an agent. We also employ it for explanatory purposes: we explain the occurrence of beliefs, thoughts and actions by seeing them as guided by the reasons available to an agent. And for this kind of purpose, positing component agents within an agential system, with component points-of-view, is potentially very useful. To explain the functioning of an agential system, we often

[16] This point could also be made using Sider's notion of the intrinsic counterparts to maximal concepts – while 'rock' is maximal, its counterpart concept 'rock*' is exactly the same but not maximal, so that many parts of a rock will be 'rocks*' but not 'rocks'. My point would then be that 'agents*' are just as interesting, from a scientific and philosophical perspective, as 'agents', and claims which are true of 'agents' but not of 'agents*' are not deep or interesting claims.

find it useful to break it down into parts, and to explain the behaviour of those parts we often find it useful to describe them as having aims and representations of various kinds. This is particularly true in light of the various forms of internal conflict that can occur between selfless agents, which will often be best explained as resulting from two somewhat independent agents reaching different conclusions within the same system.

Alternatively, it might be objected that these clusters cannot be regarded as points-of-view because they lack various kinds of 'autonomy'. There may, for instance, be decisions in one cluster whose rationale is not fully represented by thoughts in that cluster; that cluster has only the communicated 'gist' of a rationale whose full understanding is in the other cluster. More broadly, the information processing going on in each cluster is 'porous', opening out towards the contents of the other in a way that prevents us from understanding what happens in that cluster exclusively in terms of other goings-on in that cluster. But I do not think we can reasonably say that this kind of openness is incompatible with being the point-of-view of a rational agent, for we all display greater or lesser versions of it in our own everyday lives. We use terms whose meanings we are unsure of, relying on a broader linguistic community to specify their significance. We reach conclusions based on trusting the reasoning which has gone on in other heads. We embark on actions because commanded, requested or mandated to do so. Only by unrealistically exaggerating the 'autonomy' of ordinary human agency can we treat it as a necessary condition for rational agency.

Of course, it is an empirical question, for any particular seamless group agent, whether there will be clusterings that deserve to be singled out as component agents, and which they are. Discovering these lines of cleavage – which cognitive processes depend on which others, and how closely, which are dissociable or capable of coming into conflict, is a major part of the goal of cognitive science. There might be ways of connecting agents that do not admit of any clusterings at all, in which case the case for recognising both parts and whole as agents will be weaker. But it is at least a clear possibility for clusterings like this to exist, and thus we should not think it necessary for all rational agents to have individual self-knowledge. For the agents whose points-of-view are these subclusters will lack any such knowledge. And since seamless agency in a group does not rule out the simultaneous existence of component agents, we cannot infer from the seamlessness of our own agency – our lack of awareness of any self except our unitary whole self – to the literal non-existence of agential parts of us. We should thus reject Agentive Anti-Homuncularism, and take homuncular realism as a live option. As far as we know, subsystems within us may be literally and genuinely rational agents, despite not being conscious of themselves as such.

The above arguments, of course, apply to agential subsystems themselves: they may in turn be composed of genuinely agential parts performing simpler tasks,

who may also be composite, and so on. At the bottom level may be agential parts that are far from being rational, barely deserving the label of 'agent': at the higher levels may be agential parts that are similar to us in their sophistication.

One particularly promising case that may involve moderately selfless agents is the split-brain phenomenon, where cutting the corpus callosum that normally connects the two cerebral hemispheres seems to lead to the presence of 'two minds in one head' (for defences of the two-minds diagnosis, see Sperry 1965; Nagel 1971; Schechter 2015; for criticism, see Tye 2003; Bayne 2008; and also Nagel 1971). As Schechter (n.d.) shows, if there are two minds, they are clearly both rational agents, but lack any individual self-knowledge: their self-knowledge is entirely about the patient as a whole ('we'). In my terms, they are mostly mutually motor vulnerable, almost certainly morally unified, and accomplish evidential unity and mutual cognitive vulnerability by means of subtle 'indirect' environmental, corporeal and sub-cortical cues. These methods, and their resultant agential unity, can be made to fail in carefully controlled experimental settings, leading to impairment as a rational agent (e.g., doing things but being completely in the dark as to whether and why one did them). But most of the time, they are able to establish sufficient unity and mutual vulnerability by these indirect means that they function as unimpaired rational agents (Schechter 2012).

For a good example of a decomposition into agents only some of which possess reason, consider the various views falling under the umbrella of 'dual-systems theory' (see, e.g., Evans 2003; Frankish 2010; Kahneman 2011). These propose that humans have two somewhat distinct systems for interpreting and responding to the world, one that operates quickly, produces relatively inflexible, stereotyped responses, and is evolutionarily old, and another that operates more slowly and flexibly, and is a more recent, or even distinctively human, evolutionary development. Insofar as this proposal identifies two parts of the human being which both perform agent-like tasks – interpreting the world, reaching conclusions about it, initiating actions in response to it – it can be read as one account of the selfless agents that compose us. If we read it in that way, it seems likely (depending on how exactly we understand the two systems, which are distinguished in different ways by different theorists) that the second system but not the first will be a rational agent, capable of abstract reasoning and reflection. Of course, both systems may themselves admit of further decomposition into yet-simpler sub-agents, but this no more impugns their status as genuine agents than it does ours.

## *8. Conclusions*

The Indispensability Claim seems like a truism: rational agents must be self-conscious, must know themselves individually. But I have argued that this view

is parochial: we humans happen to use the I-concept to organise our agency but other beings might not. These beings might be artificial intelligences moving in a cybernetic world, or human-derived cyborgs co-ordinating by radio-mediated telepathy. They might even be the component parts of our own brains. But all I have sought to show is that they are not impossible in principle – they simply require certain preconditions, namely the relations of moral unity, evidential unity, mutual cognitive vulnerability, and mutual motor vulnerability among all those agents who are referred to by the relevant uses of 'we'.

## References


Alter, T. 2013, "Social Externalism and the Knowledge Argument", *Mind*, **122**, 486, pp. 481–496.

Anscombe, E. 1963, *Intention* (2nd edition). Oxford: Blackwell.

Baars, B. 1997, *In the Theater of Consciousness: The Workspace of the Mind*. Oxford: Oxford University Press.

Ball, D. 2009, "There Are No Phenomenal Concepts", *Mind*, **118**, 472, pp. 935–962.

Bayne, T. 2008, "The Unity of Consciousness and the Split-Brain Syndrome", *Journal of Philosophy*, **105**, 6, pp. 277–300.

Bechtel, W. 1994, "Levels of Description and Explanation in Cognitive Science", *Minds and Machines*, **4**, pp. 1–25.

Bennett, M. and Hacker, P. 2003, *Philosophical Foundations of Neuroscience*. Chichester: Wiley-Blackwell.

Bersoff, D. and Miller, J. 1993, "Culture, Context, and the Development of Moral Accountability Judgments", *Developmental Psychology*, **29**, pp. 664–676.

Blakemore, S., Wolpert, D. and Frith, C. 2002, "Abnormalities in the Awareness of Action", *Trends in Cognitive Sciences*, **6**, 6, pp. 237–242.

Block, N. 1995, "On a Confusion about a Function of Consciousness", *Behavioral and Brain Sciences*, **18**, 2, pp. 227–287.

Bratman, M. 2009, "Modest Sociality and the Distinctiveness of Intention", *Philosophical Studies*, **144**, pp. 149–165.

Burge, T. 2000, "Reason and the First-Person", in: C. Wright, B. Smith and C. Macdonald, eds, *Knowing Our Own Minds*. Oxford: Oxford University Press, pp. 243–271.

Carruthers, G. 2012, "The Case for the Comparator Model as an Explanation of the Sense of Agency and its Breakdowns", *Consciousness and Cognition*, **21**, pp. 32–47.

David, N., Newen, A. and Vogeley, K. 2008, "The 'Sense of Agency' and its Underlying Cognitive and Neural Mechanisms", *Consciousness and Cognition*, **17**, pp. 523–534.

De Sousa, R. 1976, "Rational Homunculi", in: A. Rorty, ed., *The Identities of Persons*. Berkeley, CA: University of California Press, pp. 217–238.

Dennett, D. 1991, *Consciousness Explained*. Harmondsworth: Penguin.

Desmurget, M., Reilly, K., Richard, N., Szathmari, A., Mottolese, C. and Sirigu, A. 2009, "Movement Intention after Parietal Cortex Stimulation in Humans", *Science*, **324**, pp. 811–813.

Dworkin, R. 1988, *Law's Empire*. Cambridge, MA: Belknap Press.

Evans, J. S. 2003, "In Two Minds: Dual Process Accounts of Reasoning", *Trends in Cognitive Sciences*, **7**, pp. 454–459.

Fernyhough, C. 1996, "The Dialogic Mind: A Dialogic Approach to the Higher Mental Functions", *New Ideas in Psychology*, **14**, 1, pp. 47–62.

Fernyhough, C. 2004, "Alien Voices and Inner Dialogue: Towards a Developmental Account of Auditory Verbal Hallucinations", *New Ideas in Psychology*, **22**, pp. 49–68.

Frankfurt, H. 1971, "Freedom of the Will and the Concept of a Person", *The Journal of Philosophy*, **68**, 1, pp. 5–20.

Frankish, K. 2010, "Dual-Process and Dual-System Theories of Reasoning", *Philosophy Compass*, **5**, 10, pp. 914–926.

Fried, I., Katz, A., McCarthy, G., Sass, K., Williamson, P., Spencer, S. and Spencer, D. 1991, "Functional Organization of Human Supplementary Motor Cortex Studied by Electrical Stimulation", *Journal of Neuroscience*, **11**, 11, pp. 3656–3666.
Frith, C. 2012, "Explaining Delusions of Control: The Comparator Model 20 Years On", *Consciousness and Cognition*, **21**, pp. 52–54.
Gallotti, M. and Frith, C. 2013, "Social Cognition in the We-Mode", *Trends in Cognitive Sciences*, **17**, 4, pp. 160–165.
Gilbert, M. 2000, "What Is It for Us to Intend?", in: *Sociality and Responsibility.* Lanham, MD: Rowman and Littlefield, pp. 14–36.
Gilbert, M. 2002, "Collective Guilt and Collective Guilt Feelings", *The Journal of Ethics*, **6**, pp. 115–143.
Gregory, D. 2017, "Is Inner Speech Dialogic?" *Journal of Consciousness Studies*, **24**, 1–2, pp. 111–137.
Grübler, G. and Hildt, E. eds. 2014, *Brain-Computer-Interfaces in their Ethical, Social and Cultural Contexts*. Berlin: Springer.
Heller, M. 2000, "Temporal Overlap is Not Coincidence", *The Monist*, **83**, 3, pp. 362–380.
Holmes, N. P. and Spence, C. 2005, "Beyond the Body Schema: Visual, Prosthetic, and Technological Contributions to Bodily Perception and Awareness", in: G. Knoblich, I. M. Thornton, M. Grosjean and M. Shiffrar, eds, *Human Body Perception from the Inside Out*. New York: Oxford University Press, pp. 15–64.
Huebner, B. 2014, *Macrocognition: A Theory of Distributed Minds and Collective Intentionality.* Oxford: Oxford University Press.
Imaizumi, S., Asai, T. and Koyama, S. 2016, "Embodied Prosthetic Arm Stabilizes Body Posture, While Unembodied One Perturbs It", *Consciousness and Cognition*, **45**, pp. 75–88.
Kahneman, D. 2011, *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux.
Keller, M., Edelstein, W., Krettenauer, T., Fang, F. and Fang, G. 2005, "Reasoning about Moral Obligations and Interpersonal Responsibilities in Different Cultural Contexts", in: W. Edelstein and G. Nunner-Winkler, eds, *Morality in Context*. Amsterdam: Elsevier, pp. 317–337.
Lavin, D. 2012, "Must There Be Basic Action?" *Noûs*, **47**, 2, pp. 273–301.
Lewis, D. 1979, "Attitudes De Dicto and De Se", *The Philosophical Review*, **88**, 4, pp. 513–543.
Locke, J. 1975, *An Essay Concerning Human Understanding*, ed. P. Nidditch, Oxford: Clarendon Press (originally published 1694, 2nd edition); partly reprinted in: J. Perry, ed., *Personal Identity.* Berkeley, CA: University of California Press.
Lowe, E. 1996, *Subjects of Experience*. Cambridge: Cambridge University Press.
Moore, J. and Haggard, P. 2008, "Awareness of Action: Inference and Prediction", *Consciousness and Cognition*, **17**, pp. 136–144.
Naam, R. 2012, *Nexus*. Oxford: Osprey Publishing.
Nagel, T. 1971, "Brain Bisection and the Unity of Consciousness", *Synthese*, **22**, pp. 396–413.
Noonan, H. 2003, *Personal Identity* (2nd edition). London: Routledge.
Olson, E. 2007, *What Are We? A Study in Personal Ontology*. New York: Oxford University Press.
Parfit, D. 1984, *Reasons and Persons*. Oxford: Oxford University Press.
Parfit, D. 1999, "Experiences, Subjects, and Conceptual Schemes", *Philosophical Topics*, **26**, pp. 217–270.
Paul, S. 2009, "Intention, Belief, and Wishful Thinking: Setiya on 'Practical Knowledge'", *Ethics*, **119**, 3, pp. 546–557.
Perry, J. 1979, "The Problem of the Essential Indexical", *Nous*, **13**, pp. 3–21.
Pettit, P. and List, C. 2011, *Group Agency: The Possibility, Design, and Status of Corporate Agents*. Oxford: Oxford University Press.
Roelofs, L. 2014, "Phenomenal Blending and the Palette Problem", *Thought*, **3**, pp. 59–70.
Roelofs, L. 2016, "The Unity of Consciousness, Within and Between Subjects", *Philosophical Studies*, **173**, 12, pp. 3199–3221.
Roelofs, L. Forthcoming-a, "Can We Sum Subjects? Evaluating Panpsychism's Hard Problem", in W. Seager, ed., *The Routledge Handbook of Panpsychism*. New York: Routledge.
Roelofs, L. Forthcoming-b, *Combining Minds*. Oxford: Oxford University Press.
Rosenberg, J. 1994, "Comments on Bechtel, 'Levels of Description and Explanation in Cognitive Science'", *Minds and Machines*, **4**, 1, pp. 27–37.

Rovane, C. 1998, *The Bounds of Agency: An Essay in Revisionary Metaphysics*. Princeton, NJ: Princeton University Press.
Rovane, C. 2005, "Alienation and the Alleged Separateness of Persons", *The Monist*, **87**, 4, pp. 554–572.
Rovane, C. 2012, "Does Rationality Enforce Identity?", in: A. Coliva, ed., *Self and Self-Knowledge*. Oxford: Oxford University Press, pp. 17–40.
Rovane, C. 2014a, "Forward-Looking Collective Responsibility: A Metaphysical Reframing of the Issue", *Midwest Studies in Philosophy*, **38**, 1, 12–25.
Rovane, C. 2014b, "Group Agency and Individualism", *Erkenntnis*, **79**, pp. 1663–1684.
Scanlon, T. 1998, *What we Owe to Each Other*. Cambridge, MA: Harvard University Press.
Schechter, E. 2012, "Intentions and Unified Agency: Insights from the Split-brain Phenomenon", *Mind and Language*, **27**, 5, pp. 570–594.
Schechter, E. 2015, "The Subject in Neuropsychology", *Mind and Language*, **30**, 5, pp. 501–525.
Schechter, E. n.d., "Bodies, Selves, and Self-Consciousness". Unpublished manuscript.
Seager, W. 1990, "The Logic of Lost Lingens", *Journal of Philosophical Logic*, **19**, pp. 407–428.
Searle, J. 1990, "Collective Intentions and Actions", in: P. Cohen, J. Morgan and M. Pollack, eds, *Intentions in Communication*. Cambridge, MA: MIT Press, pp. 401–415.
Selfridge, O. 1959, "Pandemonium: A Paradigm for Learning", *Symposium on the Mechanization of Thought Processes*, London: HMSO.
Setiya, K. 2008, "Practical Knowledge", *Ethics*, **118**, 3, pp. 388–409.
Setiya, K. 2009, "Practical Knowledge, Revisited", *Ethics*, **120**, 1, pp. 128–137.
Shanahan, M. and Baars, B. 2004, "Applying Global Workspace Theory to the Frame Problem", *Cognition*, **98**, pp. 157–176.
Shoemaker, S. 1970, "Persons and Their Pasts", *American Philosophical Quarterly*, **7**, 4, pp. 269–85.
Sider, T. 2001, "Maximality and Intrinsic Properties", *Philosophy and Phenomenological Research*, **63**, 2, pp. 357–364.
Smith, C. 2011, *What Is a Person?: Rethinking Humanity, Social Life, and the Moral Good from the Person Up*. Chicago, IL: University of Chicago Press.
Southwood, N. and Friedrich, D. 2011, "Promises and Trust", in: H. Sheinman, ed., *Promises and Agreements: Philosophical Essays*. Oxford: Oxford University Press, pp. 277–290.
Sperry, R. 1965, "Mind, Brain and Humanist Values", in: J. Platt, ed., *New Views of the Nature of Man*. Chicago, IL: University of Chicago Press, pp. 71–92.
Stapledon, O. 1930, *Last and First Men: A Story of the Near and Far Future*. London: Methuen.
Synofzik, M., Vosgerau, G. and Newen, A. 2008, "Beyond the Comparator Model: A Multifactorial Two-Step Account of Agency", *Consciousness and Cognition*, **17**, pp. 219–239.
Triandis, H. C. 1990, "Cross-Cultural Studies of Individualism and Collectivism", in: J. J. Berman, ed., *Nebraska Symposium on Motivation 1989: Vol. 37. Cross-cultural Perspectives*. Lincoln, NB: University of Nebraska Press, pp. 41–133.
Tuomela, R. 2005, "We-Intentions Revisited", *Philosophical Studies*, **125**, pp. 327–369.
Tuomela, R. and Miller, K. 1988, "We-Intentions", *Philosophical Studies*, **53**, pp. 367–389.
Tye, M. 2003, *Consciousness and Persons: Unity and Identity*. Cambridge, MA: MIT Press.
Vinge, V. 1992, *A Fire Upon the Deep*. New York: Tor Books.