Selfhood as Self-Representation

Kenneth A. Taylor

Stanford University

**1. Preliminaries: Self as the Bearer of Selfhood**

In this essay, I develop and defend the view that a “self “ is nothing but a creature that bears the property of what I call selfhood. Bearing selfhood is, in turn, nothing but having the capacity to deploy what I call self-representations. Self-representations are very special things. They are distinguished from other sorts of representations, I shall argue, not by *what* they represent – mysterious inner entities called selves, say -- but by *how* they represent what they represent. Both you and I may represent the living human animal, Ken Taylor. But only I may represent that very animal via self-representations of Ken Taylor. All I am doing in representing my “self” is representing the living human animal, Ken Taylor, via a distinctive kind of representation – a self-representation.[[1]](#endnote-2)

I shall have a great deal more to say about the nature of self-representations in the course of this essay.[[2]](#endnote-3) I shall argue that self-representations are distinguished from other sorts of representations by their psychofunctional roles, rather than by their representational contents. To a first approximation, self-representations play three broad roles in our mental economy. First, they subserve the synchronic integration of inner states into rationally interconnected wholes. Second, they are central to the diachronic integration of past, present and future states into a temporally unfolding self-consciousness. Third, they serve somehow to set the boundaries that separate one center of selfhood from others centers of selfhood and serve to locate the self as one self among others.

According to the picture of the self that I shall develop and defend here, then, a “self” is just a living animal or artifact that can represent itself via self-representations of that very animal or artifact. If so, then selves should not be thought of as mysterious inner entities that living human animals somehow “have” or “contain.” Selves are simply what living human animals – and possibly other animals and artifacts too –sometimes “are.” Or to put it differently, a self exists when and only when a living human animal (or other self-representing animal or artifact) is psychically organized in a certain characteristic way. Nothing but the living human animal (or other artifact or organism) psychically organized in the relevant way deserves the title of a “self.” When a living animal is organized in the relevant way we may say of that animal that it exist *as* a self.

To the maximum extent that a due respect for both grammar and style allows, I will speak throughout this essay not of ‘the self’ but of ‘selfhood.’ In so doing, I mean to forestall the temptation to reify the self. My aim is thereby to shift our focus from understanding “the self” as some sort of inner agent to understanding selfhood as an overall property of a suitably organized organism or artifact. Doing so well also enable us to chart a principled middle ground between, on the one hand, a Cartesian view of the self, according to which a self is a thinking, self-knowing, simple, possibly immaterial substance, known through, but still prior to its own self-representations and, on the other hand, an eliminativism (a la Hume) or fictionalism (a la Dennett) that outright denies, or at least takes to be problematic and doubtful, the serious and robust existence of “the self.” [[3]](#endnote-4) On my view, there are no Cartesian thinking substances. To that extent, I side fully with the fictionalist cum eliminativist. But against the eliminativist cum fictionalist, I insist that there really and truly are beings *organized as selves*.[[4]](#endnote-5) We human animals are, for the most part, psychically organized in ways that exhibit selfhood. If self-representing human animals are robustly real, then so too are selves.

Though I hold that creatures that enjoy selfhood are, in one sense, ontologically special – they are among evolution’s most intricate designs – I hope it is clear that to grant the robust reality of selfhood enjoying creatures is not to reify the self into an entity of a metaphysically sui-generis kind. Once selfhood is conceived of as a property enjoyed by nothing but a living animal or artifact psychically organized in a certain way, selfhood takes on a decidedly non-Cartesian character. A living human animal psychically organized as a self is clearly no simple or indestructible thing.[[5]](#endnote-6) We human animals are, in fact, markedly fragile and fleeting things. The psychic organization that constitutes a human animal as a self does nothing to diminish that fragility. In one way, our existence as selves may be thought to be more fragile and fleeting than even our basic biological existence. As Locke noted long ago, the very same living human animal may persist over time, perhaps while losing, either in whole or in part, the inner psychic organization that is characteristic of selfhood. When that happens, though the same living human animal will persist, she may cease to exist as a self -- or at least as a single or unified self. Confronted with the concrete reality of the ways in which our selfhood may fracture or dissolve, we may sometimes feel at a loss to say whether one or more distinct “selves” persist in the same living human animal or whether the same persisting human animal persists as the same self throughout the entire life span of that human animal. [[6]](#endnote-7)

To the Cartesian my approach will seem to enjoy all the benefits of theft over honest toil. The self, she will say, must be simultaneously that thing, whatever it is, that is doing the representing and the thing, whatever it is, that is being represented in episodes of self-representation. No mere plenum of self-representations, however organized or held together in the mind-brain, could possibly constitute, from the ground up, either the *representing* self or the *represented* self. The self is, rather, a thing prior to and independent of the self-representations that that very self both deploys and answers (or fails to answer to) in its episodes of self-representation. From this perspective, denying that there exists a metaphysically antecedent self may seem rather like trying to build a house without first laying its foundation.

I undertake to toil honestly in what follows. But I am prepared to find that my arguments may not move minds firmly in the grips of entrenched Cartesian convictions. That concession should not, however, be taken as an upfront admission of failure. My arguments are primarily addressed to the agnostic, who comes with no entrenched convictions about the metaphysical basis of our selfhood. No doubt such agnosticism may itself represent something of a philosophical achievement. That is because our pre-reflective intuitions may already be conditioned to some extent by a certain tacit Cartesian outlook. And such pre-reflective intuitions serve as the initial point of departure for much philosophizing. To that extent, my claims and arguments may appear to be at odds not just with explicit and self-conscious Cartesianism, but also with common sense itself. I do not, however, take that as a sign of the inadequacy of my arguments. Philosophy must sometimes disturb, rather than accommodate or justify common sense. To the extent that common sense does side with the Cartesian, disturbing it is what I must seek to do here. Whatever our inherited common sense may seem to imply to the contrary, I shall argue that on deeper reflection there are no good reasons for positing a substantial self that lies, as it were, “beneath” our self-representations. Beneath our capacity for self-representations is nothing but a bundle of further psychic capacities and structures that amount to nothing like a Cartesian ego. Absent antecedent Cartesian convictions, there is really no pressure to think otherwise. We can coherently maintain both that our selfhood is grounded in nothing but a certain psychic organization of our mental states and that that very psychic organization is the ultimate source of our capacity for self-representation and thus the ultimate ground of our selfhood. There need be no further, metaphysically simple “substance” that sits “beneath” the plenum of representations produced by exercises of that capacity such that it may be identified as the one, true “self” of Cartesian (and perhaps commonsensical) imagining.

But I stress again that in denying the reality of anything like a Cartesian thinking substance, I do not mean thereby to side fully with the fictionalist. In refusing or reify the self, I do not aim to dissolve, diminish, deflate or dismiss our selfhood. I aim merely to reconceive it. Now unlike the Cartesian, the eliminativist cum fictionalist should find much that I say to some degree congenial. After all, she too rejects the reification of our selfhood. She too denies the existence of a metaphysically sui generis entity that a living human being somehow contains within. Moreover, only a certain confusion – about which I shall say more shortly – could keep her from conceding that living human beings and possibly other animals and artifacts are sometimes psychically organized in ways that I contend fully suffice to constitute our selfhood. Yet despite acknowledging that humans are sometimes organized in ways that suffice, at least by my lights, for selfhood, the eliminativist cum fictionalist may still be likely to dismiss the “self” as less than robustly real -- as at best a sort of fiction, abstraction or reification and, at worst, as entirely non-existent.

What may be at issue between the fictionalist and myself are competing assessments of just what it takes to vouchsafe the robust instantiation of selfhood. The fictionalist is, I suspect, partly driven by the realization that the brain contains no such thing as a localized, central executive that functions as the mind’s chief operating officer. Perhaps she believes that only such a thing could fully merit the title of a robustly real ‘self’. And with this last thought, Cartesians would perhaps agree. But that way of thinking demands too much. Though I endorse the robust reality of selves, at least when talk of the ‘self ‘ is properly understood as referring merely to that entity, whatever it is, that instantiates the property of selfhood, no argument or claim of mine is intended to lend weight to the view that the self understood in this sense is well-regarded as the mind’s inner chief executive. Indeed, if that were what selves are required to be, I too would deny their robust reality.[[7]](#endnote-8)

It is important to stress that on my view the property of selfhood enjoys metaphysical priority over “the self.” It is solely in virtue of instantiating selfhood that an organism or artifact exists as a self in the first place. But it is also important to stress that selfhood is not a fundamental property. Indeed, our selfhood is merely the emergent tip of a mostly submerged mental iceberg – and it is, I think, in part the recognition that this must be so that leads the eliminativist cum fictionalist to (mistakenly) doubt the robust reality of our selfhood. The psychic organization that suffices to constitute a living human animal as a self is, I conjecture, caused and sustained by sub-personal structures and dynamics of which we are mostly, if not entirely, introspectively unaware. Moreover, these sub-personal structures and dynamics are largely beyond our direct conscious control. When we begin to think of our selfhood in this manner, we have moved very far from viewing selves as the mind’s inner chief operating officer. Still, it does not follow from the fact that selfhood is an emergent rather than a fundamental feature of certain organisms or artifacts that therefore “selves” are not robustly real entities. Normal intact human beings are psychically organized in ways that instantiate selfhood. We humans thus mostly exist as selves. And we are as robustly real as anything non-fundamental could possibly be.

Both the dualist and the fictionalist are led astray by their restricted fields of vision. The dualist is, I think, too narrowly focused on the deliverances of introspection as the key to the metaphysics of the self, while the fictionalist is overly impressed with the deliverances of third personal or scientific reflection. No doubt there is a contrast between, as it were, the manifest image of the self as revealed both via introspection and via our further philosophical elucidation and articulation thereof, on the one hand, and the emerging scientific image of the mind-brain, on the other hand. And I think it is entirely right to think that this admittedly stark contrast is the main source of a certain deep tension and instability in our overall philosophical understanding of our selfhood.

Consider first the deliverances of introspection. Via introspection, we seem to ourselves to cognize an enduring “inner” selfhood. And that selfhood is introspectively present to us as something plausibly primitive and sui generis -- though still to a certain extent “determinable” in a sense outlined more fully below. Introspection decidedly does not present the self as a thing grounded in lower level subconscious processes. [[8]](#endnote-9) That is because introspection bears no direct witness to the ground up “generation” of what I below call our “raw” or ”basic” selfhood. Via introspection, selfhood is experienced as always and already there, from the first dawning of reflective self-awareness –though it also presented as “determinable” in a sense to be outlined below. Just this introspectively certified, apparently groundless, ever-present, though still determinable selfhood serves, I think, as the tacit starting point for many misbegotten arguments for substance dualism, as well as for such metaphysically kindred spirits of dualism as agent causation, radical autonomy, and libertarian freedom.

Though I do not deny the felt intuitive pull of the shared, if tacit, starting points of such arguments, I find none of them ultimately convincing. When our brains are humming along smoothly enough to preserve the felt unity and integrity of what we are tempted to dub the self, we do indeed seem to have the inner experience of something like an ever-present, ungrounded selfhood. But let the brain misfire or malfunction in ways that disrupt that felt unity and integrity, and this supposedly groundless, ever-present selfhood begins to dissipate. It is only. I think, with that dissipation, that the metaphysical dependence of our selfhood on the hidden dynamics of the brain first begins to come more sharply into focus. I stress, though, that even when the unity and integrity of our selfhood is disrupted in this way, it is not through introspection that the metaphysical basis of our selfhood is made epistemically manifest for us. In fact, the ground up generation of our selfhood self via the hidden psychodynamics of the brain never comes directly into focus via introspection alone – not even as a phenomenon in need of further explanation. Rather, such epistemic hold on the metaphysical dependence of our selfhood on the hidden architecture of the mind-brain as we are able to achieve will inevitably be of a third-personal, rather than first-personal nature.

Some may be tempted to conclude that we simply have two irreconcilable standpoints from which to cognize our selfhood – perhaps with neither being eliminable or correctable in light of the other.[[9]](#endnote-10) Each independent standpoint, it may be thought, reigns supreme within its own domain. Each gives us a valid, even if merely partial, window onto the nature of our selfhood. One might think, for example, that when we are deliberating about what to do, be, or believe, we adopt the first-personal or introspective standpoint. Nothing we now know or could learn from the third person standpoint is sufficient to undermine what we know or undertake from the first personal standpoint. Indeed, one might even argue that to attempt to correct or augment the deliverances of the first personal standpoint in light of third personal information is akin to a category mistake. By contrast, when are doing the science of the brain, we do so from the third personal standpoint. Perhaps such a standpoint can claim to give us objective insight into the deep metaphysical foundations of our selfhood. But studying the self from a scientific perspective is one thing. Living as selves is an entirely different thing.

I confess that I find the appeal to two irreconcilable standpoints, forever passing each other by like ships on a night sea, neither alluring nor illuminating. Such an approach seems to me to punt prematurely, just when the going gets tough. Though I shall not offer an explicit argument against that approach, I will try to do better. I shall try to show that we can do justice enough, from within a single overall theory of selfhood, to both the first person standpoint and the third person standpoint, while neither reifying nor dismissing our selfhood. I stress, though, that doing justice enough is not the same as doing full and complete justice. We cannot simply take on board the results of each standpoint and merge them into a perfectly coherent whole. We must do precisely what the advocate of two independent standpoints says cannot be done. We must re-interpret the deliverances of each of the standpoints in light of the other, in a way that both corrects and reconciles those reinterpreted deliverances. In that process, something will no doubt be lost. But more, I suggest, will be gained.

**2. Advertisement for Three Layers of Selfhood**

Now to fully carry out this reconciliation and reinterpretation, we need some further and admittedly novel distinctions. The distinctions I have in mind are not fully and naturally at home in either standpoint, but neither are they entirely alien to either standpoint. My distinctions are perhaps best thought of as attempts at what might be called moderately conservative conceptual innovation. My aim is to reform, rather than to merely analyze and explicate, our initial, commonsense understanding of the self. The ultimate goal of such a philosophical exercise is to morph our antecedent understanding of our selfhood into something that is more nearly and obviously compatible with a gradually emerging third-personal scientific understanding of the real basis of our selfhood. In order to do that, we must somehow bridge and reduce extant conceptual and epistemic gaps between our two competing epistemic standpoints. Because my argument involves an attempt at conservative conceptual innovation, it is intended not as an argument from first principles, nor as an argument that begins and ends with conceptual-analytic truths, nor as an argument designed to merely “explicate” and clean up the rough edges of our initial untutored philosophical intuitions. Though it does bear a rough affinity to approaches that seek to “explicate” our rough and ready initial concepts, it goes further than any such approach in inviting us to re-imagine what our selfhood might possibly be. My approach grants that we have an initial referential hold, as I call it, via certain initial concepts and ideas, on the phenomenon of selfhood. Those initial ideas and concepts suffice to make it the case that it is genuine selfhood, rather than some other phenomenon, that we initially cognize. But that is no reason to presume that the concepts and ideas that give us our initial referential hold on the phenomenon of selfhood are fully adequate to the true and full nature of our selfhood. Success in our attempt to reimagine our selfhood should be measured by the degree to which selfhood as reimagined ultimately proves to fit well or badly with our best and most complete overall theory of things. It should not be measured by the degree to which we leave our antecedent or initial concepts of our selfhood standing just as they antecedently were. My guiding conjecture is that we gain a fuller understanding of the nature of our selfhood if we view our selfhood as something of a three-faceted phenomenon, involving what I call raw or basic selfhood, what I call determinable selfhood, and what I call determinate selfhood. Though I take these three facets of selfhood to be facets of one and the same thing, at least in us humans, they each deserve their separate notice.

Begin with raw or basic selfhood. Raw or basic selfhood, I claim, is ultimately grounded in nothing but that distinctive form of psychic organization, whatever it is, that ultimately causes and sustains the bare capacity for self-representation. For us humans at least, raw selfhood is introspectively opaque in the sense that the true metaphysical basis of our raw selfhood is never internally presented to us via introspection. I have already alluded to this opaqueness above and I have suggested that it is this very opaqueness that tends to tempt some to dualism.

We also enjoy an apparently more transparent form of selfhood --determinable selfhood. Through this form of selfhood, we self-represent ourselves as determinable beings. To enjoy determinable selfhood is to have the capacity to ask oneself what I call “Who am I” questions. Such a creature can ask of herself, and in the asking thereby self-represent herself as asking, such questions as: “Who am I to be?” “What am I to do?” “What am I to believe?” We might say that through our determinable selfhood, we represent the “self” not as something fixed and determinate, but as a problem to be solved, a question to be answered, or a project to be undertaken.

In us humans, at least, determinable selfhood is best understood as a dialectical moment on the journey toward more determinate selfhood. Determinate selfhood is a further achievement of a determinable self. To achieve determinate selfhood is to become, partly through the efficacy of one’s own self-representations, the answer to the “who am I?” questions that are the hallmarks of determinable selfhood. In achieving determinate selfhood, we self-represent ourselves as beings of a certain kind and in the self-representing we make an often indispensible contribution to constituting ourselves as beings of the relevant kind. Part -- though only part --- of what constitutes me as an American is that I self-represent myself as such.

Now I hasten to add that it takes more than mere powers of self-representing to achieve determinate selfhood. To self-represent oneself in a certain way is not ipso facto thereby to become or be what one self-represents oneself as. And this means that our self-representations may fail to be fully veridical. A faithless spouse, with a wandering eye and heart, may, when caught up in enduring episodes of bad faith, self-represent himself as true blue and faithful till the end. But his self-representing himself as such clearly does not ipso facto make him to be such. In order to achieve determinate selfhood, one needs not just the capacity to give oneself answers to the who am I questions but also the capacity to govern one’s life in accordance with those self-given answers. And this requires not just the power of self-representation, but also the power for what we might call self-determination – which is nothing but the power to become what one self-represents oneself as being, at least in part through the causal efficacy of those very self-representations.

The main burden of the remainder of this essay is to lay the groundwork on which a full defense of this multi-faceted conception of our selfhood ultimately rests. For lack of sufficient space, I do not pretend to execute that full defense of my view in all of its complexity here, however. What I offer here is something more on the order of down payment toward the discharging of on my ultimate argumentative and explanatory debts. For that reason I will primarily be focusing on explaining the kinds of foundational representational powers and structures that I take to be characteristic of raw or basic selfhood. I will have comparatively little further to say, except toward the end, about the nature of determinable and determinate selfhood – not because there isn't a great deal to say. There is rather too much to say to be squarely contained within a single essay.

Since I hold the raw or basic selfhood is a matter of being a creature with the capacity to deploy self-representations, I begin in section three by saying a bit more about the distinctive nature of self-representations – in particular, on their distinctive psycho-functional roles. In section four, I indulge in a brief historical interlude by way of comparing and contrasting my views about the nature of our selfhood with the views of Lock, Hume and Kant. In section five, I take up the unity of the self. I argue, very much in the spirit of Kant, that the unity of the self is a form of what I call representational unity, rather than a form of substantial unity. I conclude with some forward-looking reflections about the relations among raw or basic selfhood, determinable selfhood and determinate selfhood.

**3. On the Psycho-Functional Roles of Self-Representations**

I began this essay with the claim that self-representations are distinguished from other sorts of representations not by what they represent, but by how they represent what they represent. Such representations, I claim, play distinctive psycho-functional and social-functional roles in both our mental and social lives. My aim in this section is to try to characterize those roles just a bit more fully – though I do not aim at exhaustion. In us humans at least, self-representations play three broadly distinct, but closely interconnected roles in our mental economies. First, they subserve the synchronic integration of inner states into rationally interconnected wholes. Second, they are central to the diachronic integration of past, present and future states into a temporally unfolding self-consciousness. Third, they serve to set the psychic boundaries that separate one center of selfhood from others centers of selfhood.[[10]](#endnote-11) I will discuss each of these roles briefly in turn.

Begin with the synchronic integration of simultaneous mental states into rationally interconnected networks. It is one thing for two or more mental states to occur simultaneously in a given brain. It is another thing entirely for a self-representing being to self-represent her simultaneous states as simultaneously states of her own. It is no doubt possible for a creature to have the thought that *p*while simultaneously also having the thought that q, but without thereby self-representing those two thoughts as simultaneous thoughts of her own. A self-represented thought, I want to suggest, “belongs” to a self-representing creature in a way that a thought that is not yet self-represented by that creature does not – even when the unrepresented thought occurs within the spatial confines of that creature’s mind. Think of self-representing a thought as a way of taking a kind of ownership of a thought.[[11]](#endnote-12)

My claim is that thoughts of which an agent has taken ownership through her powers of self-representation subsist together in a new way. The intuitive idea is that they subsist together as simultaneous thoughts of a single self-consciousness. And the substantive claim about what this “subsisting together” consists in is nothing but being represented together via the power of self-representation. To be sure, thoughts that are not yet united into a single self-consciousness via the mediation of self-representations may still be said to “belong” to a given self-representing creature. But that form of belonging is merely a matter of persisting within the spatial confines of a given mind-brain. When thoughts persist together in a single brain simultaneously, but without being self-represented as so persisting, there is a sense in which they may be said to be isolated from each other. But I do not mean to suggest that such states will necessarily be *causally* isolated from each other. There may well be causal pathways that connect states that are not yet self-represented to one another without the mediation of our powers of self-representation. My point is rather that when transitions and relations among our thoughts are mediated by our powers of self-representations, those transitions and relations take on a very different character. Via our powers of self-representation, what would otherwise persist as merely causal and/or temporal relations of simultaneity and succession among a collection of thoughts may, at least in the right sort of mind, with the right sort of additional capacities and powers, be promoted into a collection of states that stand in rational relations to one another. Merely causal and/or temporal relations of simultaneity and succession among our thoughts are not yet rational relations. That is because states that are not yet self-represented do not implicate a single self-consciousness or a single center of rational self-management. Or so I shall argue.

In distinguishing the bare causal interanimation of our mental states from their distinctively rational interanimation, I do not mean thereby to deny that rational interanimation itself involves a distinctive sort of causal interanimation. Human reason is not merely an evaluative capacity that critically evaluates and normatively assesses transitions and relations that it is powerless to causally influence. Reason is a distinctive causal power of the human mind-brain. A fully rational mind is not just a mind that moves *in accordance with* reason. A rational mind is actually *causally driven* by reason. When reason is causally operative in a human mind-brain, it will no doubt have a distinctive causal profile. But trying to say, in non-normative, fully naturalistic terms jut what the distinctive causal profile of rational movements of the mind consists in is a daunting task. I shall not attempt to undertake that task here. My current aim is the rather more modest one of showing that our capacity for self-representations plays a decisive role in making distinctively rational movements of the mind possible in the first place.

Some philosophers will no doubt deny that sort of task I set aside for another occasion could possibly be carried out in any case. Philosophers of that ilk tend to believe that reason and rationality have “no echo” in the natural order -- as Davidson (1970) once put it. They are likely to interpret talk of reasons and rationality as somehow intrinsically and irreducibly normative and to insist that nothing normative can be fully understood in merely causal terms. To be sure, in addition to arguing that rationality fails to have an echo in the brute causal order, Davidson also held that reasons must also be causes. And in order for this to be true, he argued, each mental particular must be nothing but a physical particular. Granted, he seemed to believe at the same time that locating a given intentional mental state in, as it were, the space of reasons is an entirely different matter from locating that very state in the brute causal order of physical nature. We cannot possibly derive, he seemed to believe, a characterization of what we might call the rational profile of a mental cum physical state from just a characterization of its brute causal profile. Nor could we go the other way and derive a description of a state as mental or rational from a description of it as merely physical. Many subsequent philosophers have followed Davidson in holding views of roughly this sort. Though I disagree with this approach on many different grounds, I will not attempt to systematically adjudicate those disagreements here. I will instead just help myself to an understanding of reason according to which, whatever else it may be, it is one among the mind-brain’s distinctive causal powers -- on a par with perception and the capacity for language, say. For those of a different cast of mind, I can only ask their forbearance in bracketing the many and deep issues that may divide us.

Once it is acknowledged, if only for the space of the current argument, that reason is a distinctive causal power of the mind-brain that is causally and not just normatively implicated in at least some of our real time rational cognition, it is relatively straightforward to show that the capacity for self-representation plays a distinctive role in making real time rational cognition possible in the first place. To see this, begin by considering the belief that p and the belief that not p. Nothing prevents two *distinct* rational minds from being such that while the one believes that p the other believes that not p.[[12]](#endnote-13) In a *single* rational mind, however, believing that p would seem to rationally preclude believing that not p. To say this is not to deny that there are individual minds that do in fact manage to hold, somehow or other, such pairs of beliefs simultaneously. We do tend to condemn such minds as less than fully rational, but less than fully rational minds are real psychological possibilities.

Now the fact that we do condemn those with contradictory beliefs as not fully rational may seem to suggest that we must understand rational preclusion in normative rather than in merely causal terms. Such preclusion, one might think, amounts to nothing but the fact that one *ought* not simultaneously believe both that p and that not p. But an ought is not a prediction. To say that one *ought* not believe p and not p simultaneously is decidedly not to predict that one *will* not so believe. But recall that our guiding assumption is that reason is more than a passive normative evaluator of the mind’s movements. Reason is also an active causal power that is actually causally operative in paradigmatic episodes of rational cognition. If that is so, then at least to the extent that the mind is actually causally driven by reason, rather than something other than reason, rational preclusion will typically involve causal preclusion as well.

In virtue of what do two rationally incompatible states tend to causally preclude each other in a mind actively driven by reason? Though I will not propose a full account of the causal nature of rational compulsion and preclusion here, I do want to suggest that the capacity for self-representation plays a central role in enabling both real time, online rational compulsion and rational preclusion. In particular, I claim that it is through the mediation of our powers of self-representation that our mental states first come to exert rational cum causal pressure on one another at all. It is only to the extent that the diverse mental states of a self-representing rational agent are, as it were, gathered together into rationally and causally interanimating networks that causally efficacious, real-time rational cognition is possible at all. Or to put the point in a somewhat Kantian vein, it is only when a diversity of states are thought together, as one and all states of a *single* self-consciousness, that those states begin to exert rational, rather than merely causal pressure on one another. It is because the thought that p and the thought that –p are both self-represented as *mine* -- and this through the deployment of my powers of self-representation -- that there is any rational pressure to resolve the evident conflict between the two beliefs. For example, were one of those beliefs (self-represented as) mine and the other (self-represented as) yours, we might happily agree to disagree. Alternatively, if neither state was self-represented at all there is no obvious reason why they could not simply co-exist in splendid rational isolation from one another, even while co-resident in a single brain.

Now it may be thought that this approach requires something like a Cartesian inner agent, one that can perhaps survey, compare and contrast its inner states and rationally manage them in the manner of a central inner chief executive. I have already denied that any such thing is required to vouchsafe the robust reality of our selfhood. But let me say more about why this is so. First, it would be a mistake to think of the deployment of a self-representation as a consciously or deliberately performed act of an (inner) agent. Mental states that function in the way of self-representations are just more mental states among others in the total stew of mental states. I do hold, to be sure, that such states play quite distinctive psycho-functional roles in the total mental economies of self-representing agents. I have been arguing here that such states function as a kind of inner psychic glue in that they serve to bind collections of mental states together into rationally (and therefore causally) interanimating wholes. But there is no reason to suppose that in order for some states in the total stew of mental states to play that role, there must be anything like a Cartesian ego that sits, as it were, beneath those states.

**4. A Brief Historical Interlude on Locke, Hume and Kant**

It may help to note that my view bears a certain rough affinity both to elements of Locke’s (1959) theory of personal identity and, to a greater extent, to elements of Kant’s (1999) views about the transcendental unity of apperception – though there are, to be sure, aspects of both views that I either outright reject or remain neutral about. To their credit, both Locke and Kant are steadfastly anti-Cartesian in their thinking about the self. Nor could either justly be called an eliminativist or fictionalist. Both grasp – though Locke does so through a glass darkly, while Kant does so more clearly -- that self-representational states are states of a quite distinctive kind, with a quite distinctive role in our cognition and connation. Kant seems to have anticipated a broadly functionalist view about the nature of self-representations – though he seemed also to think that they were devoid of representational content.[[13]](#endnote-14) Locke famously suggests that a self or person is “a thinking intelligent being that has reason and reflection, and can consider itself as itself, the same thinking thing, in different times and places.” But Locke does not offer a developed account of just what a person’s capacity to “consider itself as itself” consists in. He does say that a person considers itself as itself by means of “that consciousness which is inseparable from thinking.” And it is this consciousness, he claims, which “makes everyone to be what he calls a self.” But we are told nothing at all about what such consciousness could possibly consists in.

This charge may, at superficial first glance, seem unfair to Locke. Locke does hold, after all, that we have a capacity for what he calls reflection. He clearly takes reflection to be the capacity whereby the mind is aware of its own operations. So it might be thought that our self-representations are somehow grounded in the deliverances of reflection. But this cannot be the whole story. On Locke’s view, reflection is essentially a perceptual modality – what we might call inner perception. And like all perceptual modalities, it is, on his view, the source of an array of simple ideas. Now simple ideas, according to Locke, enter the mind “simple and unmixed” with each containing just “one uniform appearance.” Moreover, he claims that no “combination,” as Kant would put it, is yet thought through our merely simple ideas. But just these features of simple ideas – whether derived from inward reflection or from outer sensation -- imply that no Lockean simple idea, nor any sequence of such ideas, is fit to play the sort of unifying functional role that is characteristic of our self-representations.

To his credit, Locke does seem to be aware that *something* in the mind must play the unifying (functional) role that I assign to self-representations. And he acknowledges that whatever it is that does unify our mental contents in the relevant way is essential to our selfhood. He seems even to acknowledge, if only tacitly and confusedly, that whatever it is that unifies mental contents into a single self-consciousness cannot simply be one among the simple deliverances of either inner reflection or outer sense. He recognizes, for example, that this unifying thing, whatever it is, is always and already present in all (self-represented) episodes of thinking, reflection, volition or sensation. As he puts it:

When we see, hear, smell, taste, feel, meditate, or will anything, *we know that we do so*. Thus it is always as to our present sensations and perceptions and *it is that which makes every one to be what he calls self, and thereby distinguishes himself from all other thinking things.*

This passage is rich with insight. But there are also difficult question raised by it that Locke nowhere addresses, as far as I can tell. To begin, it is not at all clear that it is right to say that *whenever* we see, hear, taste, feel, meditate, or will, we are consciously aware that we are doing so. This is not to deny that there is a truth that Locke is getting at here. Suppose I see a cup. Quite often, even if not always, I will be aware not just of the cup, but of my seeing of the cup. Clearly my awareness of seeing the cup is quite closely bound up with my seeing of the cup. When I see the cup and am aware of doing so, it is not as though the seeing of the cup comes before my mind with a question mark attached. “Ah, there is a seeing of the cup, but just whose seeing of the cup is it? “ One might be tempted to conclude – as Locke seems to be – that my seeing of the cup is somehow *automatically* accompanied by my consciousness of seeing the cup.

But Locke is clearly wrong here. We clearly perceive a great deal without being consciously aware of doing so. Think, for example, of blind sight. The point here is not so much to argue that Locke overestimates the extent to which we actually self-represent our perceptions. The point is rather that Locke’s approach naturally gives rise to a set of questions to which he himself provides inadequate answers. For example, Locke tells us almost nothing about what that consciousness that supposedly accompanies every episode of sensation, reflection, or volition consists in. Perhaps the natural first thought for a Lockean would be that our conscious awareness of our own mentation must be a product of the capacity for reflection. But given what Locke takes reflection to be and given what sorts of contents and ideas he takes it to deliver, this thought cannot be sustained. This is not to deny that something like reflection may play a role in making me aware that I am *seeing* *the cup now.*  But it does not seem at all plausible that reflection, as understood by Locke, is the ultimate source of my awareness that it is *I* who am seeing the cup now. The point is that while something like Lockean reflection may well be the ultimate source of some, perhaps even much, of the *contents* of our self-consciousness, it is not at all plausible that reflection just is “that consciousness which is inseparable from thinking.”

The problem, I think, is that in a mere sequence of simple ideas of reflection, there is as yet nothing that could ground the *unity* that is characteristic of self-consciousness. In the bare deliverances of reflection, there is nothing that explains what it is for that various mental acts and states to be and to be self-represented as one and all *mine.*  For again, each simple idea in a sequence of such ideas – whether derived from reflection or external sense – is, on Locke’s view, separate and distinct from every other idea in that sequence, with no intrinsic ties among them to, as it were, bundle them together.

Locke faced a similar difficulty when it came to our ideas of *external* objects. Since, again, ideas enter the mind “simple and unmixed” and, as such, bear no intrinsic connection to each other, it is not immediately obvious how we ever manage to cognize two simple qualities as cohering or subsisting together in a single underlying thing. Locke’s infamous answer, of course, is that in addition to the ideas of various sensible qualities – through which, again, no “combination” is directly cognized -- we also have the (simple) idea of a substance. The idea of a substance, according to Locke, is just the idea of an “unknown somewhat” in which qualities that are capable of producing ideas in us inhere. It is through and only through deploying the idea of substance, he thinks, that we can cognize distinct qualities as united or co-existing together in a single underlying thing.

Whatever the merits or demerits of this approach as applied to external objects, it is a solution of which Locke refuses to avail himself in the case of the self and self-consciousness. Yet it is clear that Locke thinks that something must enable us to cognize a diverse collection of mental states and processes as somehow belonging together in unified fields of self-consciousness. Now Locke is rightly famous for proposing the so-called memory criterion of personal identity and selfhood. He says, for example, that it is only “in and as far as [the same consciousness] can be extended backwards to any past action or thought, so far reaches the identity of that person.“ So it may be thought that memory is the thing that does for the self what the idea of a substance is supposed to do in the case of external objects – undergird our capacity to think of diverse contents as co-existing within a “single” field. But although memory is clearly thought to play a role in constituting diachronic personal identity over time for Locke, it would seem to play no role at all in settling questions about synchronic identity at a time. In fact, except for his not terribly illuminating talk of “the same consciousness” Locke has hardly a word to say about just what it is that makes two simultaneously occurring mental states state of one and the same self-consciousness. Moreover, given the kit of philosophical tools at his disposal, it is hard to see what he could say. In fact, his focus on memory and identity over time just obscures, it seems to me, how little he actually has to say – how little he could say -- about how the instantaneous boundaries of our selfhood are constituted.

It was left to the more resolute and clear-eyed Hume (1968) to see that if we begin with essentially Lockean starting points and restrict ourselves to the kit of philosophical tools that Locke had at his disposal, then the nearly irresistible conclusion must be that we have no idea of an enduing self at all. For as Hume rightly argues there is no constant inner “impression” that accompanies the diverse and ever fleeting impressions of inner sense that might plausibly serve as the “original” of the very idea of an enduring self. To be sure, though Hume himself concludes that he very idea of such a self must be empty and without application to any real existent, he also provides the beginnings of what he does not seem to fully realize might reasonably be taken as a tantalizing alternative way of thinking about our selfhood. He recognizes that one might well be tempted, in the face of his arguments, to identify the self not with some inner enduring thing that sits, as it were, beneath our diverse and fleeting inner impressions, but rather with the bundle of impressions itself. Doing so would, of course, require one to tell some story about the bundling -- that is, about what it is that makes the impressions in a given bundle to count as one and all members of that very bundle, rather than of any numerically distinct bundle. Hume himself even suggested such a story, though only to dismiss it. Perhaps the members of the bundle of perceptions that constitute this very self rather than any other self are simply interrelated to one another in some distinctive way. As he puts it, perhaps “... the true idea of the human mind, is to consider it as a system of different perceptions or different existences, which are linked together by the relation of cause and effect, and mutually produce, destroy, influence, and modify each other. “ But this cannot be so, Human argues, because robust sense cannot be made of the idea of “real connections” among intrinsically separate and distinct existences. “If perceptions are distinct existences,” he says, “they form a whole only by being connected together. But no connections among distinct existences are ever discoverable by human understanding.”

I will not dwell at length on Hume’s arguments here. I will say that he is right to worry about what exactly it is that bundles some or all of our diverse representations together. He rightly sees that though Locke has a name for the bundler -- self-consciousness – he has no account of what the bundler could possibly consist in. Ultimately, though, I think that Hume’s arguments suffer from an overly cramped picture of the nature and function of mental representations. Rather like Locke, he seems to be believe that because our original impressions, as he calls them (or simple ideas, as Locke calls them) are separate and distinct existences, it follows that no real combination or connection can possibly be thought through them. What Hume fails to appreciate, it seems to me, is that the mind contains many representations that have precisely the function of enabling us to think combination and connection through them. He is likely to have missed this fact because he was looking for all ideas to be derived from the given of sensation. But it is not at all plausible that representations deigned to allow us to think combination and connection will be directly derived from perception nor wholly composed of representations that are so derived. It is more likely that they are part of the antecedent furniture of the mind that enables the mind to construct a world for itself to cognize in response to the inward rush of outward energy upon the portals of sensation. Indeed, I hold-- but will not attempt to argue the point here -- that a mind that was not equipped with a plethora of such representations could not possibly learn the vast variety of things that experience has to teach us.

I stress this point because I take the capacity to produce self-representations to be one of those basic capacities of the human mind-brain that prepares it to make something out of the inward rush of energy upon the portals of sensation. And I take the representations that are produced through the exercise of that capacity to be one among the many varieties of representations through which combination and connection are made thinkable. Self-representations, on my view, are dedicated mental devices for uniting the multiple and diverse contents of consciousness into, as it were, a single or unified field. As such, self-representations are somewhat akin to syncategorematic expressions or, perhaps, logical expressions. Syncategorematic and/or logical expressions are representations that have the distinctive role or function of combining representations of various sorts to yield further representations. It is sometimes said, of course, that while syncategorematic expressions have grammatical or logical *functions,* they lack independent *reference --* though it is in fact possible to define a sort of abstract reference even for such expressions via the techniques of Montague Grammar, for instance. One might object reference of this sort is too cheap and easy to count as genuine or real reference. Be that as it may. For self-representations refer in a quite robust, non-derivate sense. A self –representation refers to the living human being, animal or artifact whose self-representation it is. In this respect, self-representations are perhaps unlike garden-variety syncategorematic or logical expressions. Self-representations are, in a sense, Janus faced. They both enable us to think combination and connection, in the way that syncategorematic expressions do, but they also refer. Explaining how this is possible is the main burden of much of the remainder of this essay.

I begin by noting that something akin to my view of self-representations and their role in constituting our selfhood seems to have been anticipated by Kant (1999). Witness in this connection his claim that what makes my representations count as one and all *mine* is the potential for what he calls the bare ‘I think’ to attach to them all. To many contemporary ears, this approach will no doubt have the ring of a kind of functionalism that is on offer in this essay. Whether this is how Kant himself would construe his own approach, we can, I think, rightly construe him as claiming, in effect, that the ‘I think’ has the peculiar functional role of uniting our diverse representational states into a single, mutually interanimating field of representational states. Belong to such a interanimating field of states is decidedly not, on his view, a matter of co-existing as simultaneous states of a Cartesian substantial ego. Indeed, Kant seems to want very much to find a way to deny the existence of such a thing – though it is unclear, as I shall suggest below, whether he ultimately succeeds in finding a principled basis for doing so. At the very least, it is clear that he does explicitly deny that ‘I” as it occurs in the bare ‘I think’ refers to any such thing. He says, in fact, that the ‘I’ in the bare ‘I think’ has reference to a “merely logical” subject. It is not entirely clear what he means by this. Somewhat tentatively, I will take him to be making something like the claim that the bare ‘I think’ expresses something like the universal “form” of self-consciousness, without containing any reference to any particular self-consciousness. If that is right, then at a minimum we can take Kant to be saying that through the bare ‘I think’ no particular and definite self is, as such, directly constituted as a concrete real existent.

Now Kant does distinguish what we might call empirical selfhood from what we might call transcendental selfhood. The former, he seems to think, somehow results from the actual or potential attachment of the ‘I think’ to the manifold deliverances of inner and outer sense, via the synthesizing activity of the understanding. Perhaps it is the achievement of empirical selfhood which first brings with it what we might call thick selfhood, that is, thick self-representation of oneself as both a being in time, a being that somehow stands apart, as subject, from the external world as object, and a being that subsist as merely one center of selfhood among others. But it is unclear, to say the least, just how this is supposed to work. The problem is that once one concedes that the bare ‘I think’ contains nothing that suffices to distinguish one center of selfhood from another, it is not clear how “attaching “ the ‘I think’ to the deliverances of intuition via the synthetic activity of the understanding helps to achieve what I will call localized selfhood. One possibility is that localized selfhood results from the particularity or locality of the representational arrays to which the bare but generalized ‘I think’ comes to be attached in the act of synthesis. What would distinguish one center of selfhood from another, on this view, is not anything contained within the ‘I think’ itself, but something like the bare numerical distinctness of the diverse representational arrays to which the ‘I think’ is in turn attached in diverse cases. One center of selfhood is constituted by the combination of the bare ‘I think’ with an array of representations R. A distinct center of selfhood is constituted by the combination of the “same” ‘I think’ with a wholly distinct, non-overlapping array of representations R’.

Though the sort of view gestured at here does not yet amount to a bodily criterion of personal identity of the sort typically rejected by Lockeans, it does perhaps begin to move us some steps toward that direction. One possible way to make this move work, perhaps, is to take the concrete body as something like the local material substratum of our selfhood. This is especially true if we think of the deliverances of intuition as, in some sense, deliverances of the body. The point is not merely that the self does not subsist as a disembodied thinking substance. The point is that the locality of the self might somehow be rooted in the fact of concrete embodiment. There are elements of Kant’s thinking -- especially about the empirical self -- that are perhaps consistent with a view of roughly this sort. I am thinking in particular of the Refutation of Idealism and the role of relatedness to outer objects in constituting the self as being in time. But I do not mean to suggest that Kant endorses the sort of view here gestured at. I mean merely to suggest that this is one way in which he might attempt to introduce what I am calling the locality of the self into his overall picture.

Still, I should say that it is not obvious that the suggested approach would really suffice to answer our original worry about the boundaries of selfhood. It is not enough to locate the boundary by appeal to the body alone, even if we grant that bodies have a kind of locality. The problem is that even if we think of the deliverances of intuition as the deliverances of this or that body, once the bare ‘I think’ is said to contain nothing that pertains to the locality of our selfhood, it is not immediately obvious why the application of a generalized ‘I think’ to now a local array R and now a local array R’ should constitute two distinct centers of consciousness rather than just numerically distinct occurrences within a single but non-localized self-consciousness. And this is so even if we think of R and R’ themselves as deliverances of separate and distinct localized bodies. We need some further argument that bodily locality, as it were, is either necessary or sufficient for what I shall call subjective locality. There may well be some sense in which the bare numerical distinctness of R and R’ -- which bodily locality arguably suffices to guarantee – might suffice to ground claims about the distinctness of the *contents* of consciousness A and consciousness B. But it is not at all obvious that distinctness of content suffices for the numerical distinctness of consciousness A and consciousness B. Indeed, it seems rather more likely that that which makes for the distinctness of consciousness A and consciousness B must precede and make possible the distinctness of their contents.

These last remarks may seem to suggest that in order to mark out clear boundaries between different centers of selfhood, we may need to drift our way back toward a rather more Cartesian understanding of selfhood and its boundaries. I doubt Kant intended such a drift. Nor do I contend that such drift is inevitable. Indeed, this threatened Cartesian drift is clearly at odds with what I take to be the central lesson of Kant’s views about the nature of our selfhood. That lesson hinges on a distinct between what we might call representational unity and what we might call substantial unity. Even if he did not have a fully worked out theory of our selfhood that made good on that lesson, Kant was the first to see, I think, that the unity through which our selfhood is constituted is not a form substantial unity at all, but a representational unity, achieved through noting but the exercise of our capacity for self-representation.

I do not profess to know whether Kant himself could ultimately prevent the problem of subjective locality, as I am calling it, from causing a drift back all the way back to a Cartesian conception of the self. Whatever Kant himself thought, I think that we can avoid the Cartesian drift embracing certain further distinctions that are nowhere evident in the strict letter of Kant’s work. Armed with these distinctions, we can, I think, develop a conception of our selfhood that still enshrines Kant’s genuine and deep insight that the unity of the self is a representational rather than a substantial unity, while replacing Kant’s transcendental subjectivity with a subjectivity that is intrinsically and immediately localized. The key is to is to go beyond the merely “formal” subjectivity that is contained in Kant’s transcendental ‘I think’ -- from which all particularity or locality has been bleach out – to what I call the many localized ‘I think’s. By the many localized ‘I think’s I mean to signify something like multiple immanent and psychologized versions of the Kantian generalized ‘I think’. We may think of the many localized ‘I think’s as what the Kantian transcendental ego becomes when it is made immanent, by being made to reside within a particular body and brain. I sometimes speak of the ‘I think’ *in* this or that body to emphasize that our local subjectivity must surely have a bodily basis.

I hope it is clear that the local ‘I think’ in this or that body is not anything like an inner Cartesian ego or substance. It may help to analogize the relation between Kant’s bare ‘I think’ and the many localized ‘I think’s to something like the relation between an abstract type and the many concrete tokens of that type. Thus construed, Kant’s claims about the character of the bare ‘I think’ amount to a way of abstractly characterizing something like a functional role. We may think of the many localized and immanent ‘I think’s within various bodies as the concrete realizers of this abstractly characterized functional role. The thought here is that while mere embodiment on its own may give us a kind of locality -- since mere embodiment already provides us with spatial locality at a minimum -- mere spatial locality does not as yet suffice for subjective locality. Rather, embodiment will suffice for subjective locality to the extent and only to the extent that it serves as the concrete realizer of a certain abstract functional role – the role indexed by the Kantian bare ‘I think’.

Though this way of thinking about the relationship between the bare ‘I think’ and the many local centers of subjectivity is helpful enough, it does not go quite far enough for our purposes. For all I have said so far, each localized ‘I think’ might be construed as itself a sort of concrete particular. And as such it might be thought to either stand in for an underlying ego -- if we think of it as a concrete token representation – or perhaps just to be or amount to the concrete ego. Now I do mean to construe the local ‘I think’ as itself a mere representation, but not quite as a token representation. A token representation is a dateable, locatable particular. As such, though a token may endure, it can never recur. Tokens are here today and gone tomorrow. Though the local ‘I think’ is a mere representation, it is neither fleeting nor enduring in the way that tokens are. It is what we might call a recurrent representation. As such, it is closer to a type than to a token. But this approach raises a puzzle. When I self-represent myself as having the that thought p and subsequently represent my self has having the thought that q, two distinct and separate occurrences of the localized ‘I think’ have been produced in me. But if we think of each occurrences of a local ‘I think’ as a token onto itself, the question naturally arises just what it is that guarantees that two tokens of, as it were, the same local ‘I think’ have reference to the very same selfhood?

Kant worried deeply about something akin to the question we are now addressing. At least for a given subject -- whatever a given subject turns out to be -- he seemed to hold that there must be something that guarantees that the multiple occurrences of the ‘I think’ have reference to a single subject. Otherwise, he claims, the subject would have “as multicolored and diverse of self” as the diversity of representations of which she is made conscious through the various apparently independent applications of the ‘I think’. I admit, though, to not finding Kant’s own treatment of this issue to be entirely transparent. He *seems* to want to say that the identity of the ‘I think’ throughout various episodes of self-representation is somehow vouchsafed for by the subject’s own reflective awareness of the synthesizing activity whereby she brings the diverse and disunited contents of intuition under the categories. But this way of looking at things at least threatens to reintroduce something rather like a Cartesian self-knowing subject to be, as it were, the source of that activity. As such, this approach has a similar feel, at least to my own ear, to the Cartesian view that thought knows itself by thinking itself. This way of putting matters naturally raises the question of exactly what it is that is that is doing the thinking of itself. Whatever Kant intended, the Cartesian answer may seem near to hand again. For it may now appear that Kant’s ultimate view is not that the bare ‘I think’ itself does the work of unifying the subject, but, rather, the subject’s “transparent“ synthetic activity of producing and reproducing the ‘I think’. And this, one might plausibly think, is just tantamount to locating the unity of the subject in the self-reflective activity of a Cartesian thinking subject.

**5. The Unity of the Self as Representational Unity**

I don't at all pretend to know how Kant would or should respond to this line of reasoning. Since Kant scholarship and exegesis are not my aim, I can afford to be steadfastly neutral on the issue of whether Kant has philosophical resources sufficient to overcome this difficulty. He may well have; he may well not have. It is a matter of indifference for my own approach. My own approach turns on a more direct implementation of the insight that the unity that is constitutive of our selfhood is not a form of substantial unity but merely a form of representational unity. The key is to focus more sharply on the distinctive character of self-representation, with particular attention to what I have elsewhere called the co-reference profiles of such representations and their roles as devices for thinking with what I call sameness purport. (Taylor, 2003, 2010, 2013, 2014)

I will need to digress briefly to explain what I have in mind. Elsewhere (Taylor, 2014), I have distinguished two broad classes of singular terms with respect to their co-reference profiles. Some singular terms function in both talk and in thought as devices of what I call *de novo reference,* while other singular terms function as devices of what I call *explicit co-reference*. By a device of de novo reference, I mean a singular expression or representation such that tokens of that type are referentially independent of each other and thus have no *intrinsic* purport to co-refer, even when they are what I call coincidentally co-referential. A device of explicit co-reference, by contrast, is an expression such that all tokens of that type are and have an intrinsic purport to be co-referential. Both shared natural languages and the de facto private language of thought, I have argued, contain both devices of explicit co-reference and devices of de novo reference. (Taylor 2010, 2014).

I have developed this distinction between at considerable length elsewhere and will not attempt to repeat the full details of that development here. But some brief illustrative examples may help to fix ideas firmly enough for the purposes of the present argument. My aim here is to deploy this distinction here in service of the claim that the unity of the self is a form of what I have been calling representational unity, rather than a form of substantial unity. In particular, I shall argue that the unity of the self is nothing but the “unity” that is thought through a certain dedicated inner device of explicit co-reference -- in particular through the many localized “I think’.

I start with some examples that may at first glance appear to be remote from our ultimate destination. My aim in so doing is to show that there is independent motivation for the distinction between devices of explicit co-reference and devices of de novo reference. Consider briefly the difference in communicative significance between tokening the same name again – say, ‘Cicero’ – and tokening the same demonstrative again – say, ‘that man.’ When one tokens the same name again there is a sense in which one has tokened something brand new under the sun. Until its coming, that very token has never existed before. After it passes, it will never rise again. In this sense, of course, every token of any word whatsoever is something new under the sun. What is more striking about names is that when we re-token a name, there is also a sense in which we are doing the same thing again that others have done before. In particular, except when does so in the context of an original dubbing, one typically is not performing what I call a de novo act of reference by tokening a name. One is not, that is, performing an act of reference that is entirely independent of any previous acts of reference by oneself or others. That is because token occurrences of the same name again simply do not purport to achieve reference from scratch. Once an object as been dubbed with a name, re-uses of that same name again, in effect, automatically inherit reference from earlier occurrences of that name. So in re-tokening the name ‘Cicero’ again, for example, I do not (purport to) refer de novo to Cicero, as if in some initial baptism. By previous uses of (tokens of) that very name, others have referred to Cicero before. In using that very name again, I am but performing one further episode of referring in a continuing history of such episodes. Moreover, I typically intend that it be manifest to others who also use that very name that they and I co-refer in using the same name again. That is one reason why, if I were to be asked, “Was Cicero a great Roman orator?” I would not answer by responding, “Yes, Tully was a great Roman Orator,” even though I know that Cicero is none other than Tully. I would not do so, at any rate, unless the identity of Cicero and Tully was already part of the shared common ground between my interlocutor and me. Mere use of the same name again, already on its own, does the work of making the preservation of subject matter explicit, without the need to deploy identity statements.

I take these facts about the behavior of names to be consequence of certain fundamental facts about their broadly syntactic characters as linguistic expressions.[[14]](#endnote-15) In particular, I claim that to be a name is, in part, to be an expression type such that tokens of that type are explicitly co-referential with one another and referentially independent of the tokens of any distinct name type. Two expressions *m* and *n* are referentially independent when nothing in their intrinsic linguistic characters guarantees that if *m* refers to *o* then *n* refers to *o* as well. To say that any distinct names are, as a matter of their linguistic character, referentially independent, is not to say that they must ipso facto fail to co-refer. It is just that when two distinct and thus referentially independent names do co-refer, their co-reference will be what I call a mere *coincidence of usage*. We can directly *display* the fact that two names are co-referential via true (and informative) identity statements. When referentially independent names *m* and *n* are linked by an identity sign, then the identity statement ⎡m = n⎤ puts on display the fact that *m*  and *n* are coincidentally or, as we might also say, extrinsically co-referential. Taken together, the correlative relations of referential independence and explicit co-reference partially characterize the broadly syntactic or structural role occupied by members of the linguistic category NAME.[[15]](#endnote-16)

While proper names are dedicated devices of explicit co-reference, demonstratives are dedicated devices of de novo reference. The reference of token demonstrative is not and cannot be inherited from the type of which that token is a token -- at least not without a lot of non-linguistic help from extra-linguistic context. For a demonstrative, reference must be established anew for each independent token. Thus each token demonstrative is referentially independent of any token demonstrative of the same type – even when two token demonstratives are coincidentally co-referential. Suppose, for example, that I point to Troy and say, “That man is a fine young philosopher” and then point to Troy again and say “That man is a fine young poet.” I have pointed to the same man twice. I have also referred to the man twice. I have even done so by using tokens of the same linguistic device again. Nonetheless, my two acts of reference are, in a clear intuitive sense, entirely referentially independent of each other. Though context may somehow make it clear that the same person is being referred to twice. Nothing in the narrowly linguistic relations or characters of the demonstrative expression ‘that man’ suffices to make this fact evident or to guarantee that my two tokens co-refer. It is because of the referential independence of each use of a demonstrative from every other used of that same demonstrative that I call demonstratives devices of de novo reference.

Devices of de novo reference are handy things for languages to include. Among other things, they make possible acts of reference that independently exploit the vagaries of local context and situation. But there could not be a language such that all of its referential devices function as devices of de novo reference. Any such language would, at a minimum, be significantly expressively diminished. In such a language, though speakers could still refer twice to the same object again, it would be significantly more cumbersome to make it manifest and explicit that the same object was being referred to again. If a language is to have the resources to make co-reference not only possible, but also explicit and manifest, then it needs more than devices of de novo reference. It also needs devices to make the preservation of reference explicit and manifest.

What goes for communication goes also in the de facto private language of thought. And seeing that this is so is, I want to suggest, the key to addressing a deep and important worry of Kant’s. Kant worried that unless we could find some principled basis for concluding that the “I think’ is one throughout all episodes of self-cognition the self would be multi-colored and devoid of unity. But it is precisely because the local ‘I think’ functions as a private device of what we might call de se explicit co-reference that the self is not present to us as a diverse and many colored disunity. Or so I shall argue.

But I begin by noting that Kant’s problem is in fact quite general, and not really simply a worry about self-cognition. All cognition whether of the self or of objects distinct from the self require that we be able to think with what I call intrinsic purport of sameness. For example, I can think of my son Kiyoshi today and think of him again tomorrow with a kind of inner assurance that I at least purport to be thinking of the same person twice. No I claim that I do so merely by deploying a mental name \*Kiyoshi\* across distinct thought episodes. If one deployed no such devices in thought, it would always be an open question whether, in purporting to think now of a particular *o* and now of a particular *o*’, one has thought of two distinct objects or has thought of the same object twice. It may sometimes, perhaps even often, be an open question for a cognizer whether two of her thought episodes share a (putative) subject matter, but it is surely not always so. Thinking with an intrinsic inner purport of sameness is possible just because the mind contains inner representations of various sorts that function, at the level of thought, as devices of explicit co-reference. For such devices, to think *with* or *via* them again is ipso facto to purport to think *of* the same thing again.[[16]](#endnote-17)

Mental names are what I call *de re devices of explicit co-reference.* They are dedicated devices for thinking with intrinsic purport of sameness of objects that are cognized as possibly distinct from self. But to say that names are de re devices rather than de se devices is not to deny that one can cognize oneself as the same again via the deployment in thought of a mental name for oneself. The point is rather that a name – even a name for oneself – does not have the form and function of a self-representation. A name is not, as Locke might have put it, a representational device through the deployment of which one cognizes the self *as the self*.

By contrast the local ‘I think’ functions as what I call a de se device of explicit co-reference. A de se device of explicit co-reference is a dedicate device for thinking with intrinsic purport of sameness not of objects cognize as distinct from the self, but of objects cognized as identical with the self. Though there is, I think, a sharp distinction between de se devices of explicit co-reference and de se devices, there is also a certain deep affinity between them. That is, there is a deep affinity between the way mental names function in thought and the way a local ‘I think’ functions in thought. Like a mental name, each local ‘I think’ functions, within each subject’s mental economy, as a device of explicit co-reference, rather than a device of de novo reference. As such, each token of the local ‘I think’ in a given agent enjoys an intrinsic purport to co-refer with every other token of that local ‘I think’. It is this and this alone, I think, which prevents the self from being multi-colored and disunified. But as a de se, rather than a de re device of explicit co-reference, the local ‘I think’ grounds our capacity to think with purport of sameness of the self *as the self*. That is, it is through the deployment and redeployment of an inner de se device of explicit co-reference that an agent is able to cognize herself as herself and to do so with intrinsic purport of sameness.

We are now in a position to appreciate in just what sense the unity of the self can be said to be a form of representational rather than substantial unity. First, note that a localized ‘I think’ is not at all the sort of thing for which Hume professed to search in vain – a single enduring representation, persisting through all fleeting thought episodes. Nor does the local ‘I think’ stand for or refer to some mysterious inner substance that somehow endures beneath our fleeting episodes of cognition, conation, and perception. But neither is a localized ‘I think’ a merely formal and empty representation of the sort that Kant claimed the generalized ‘I think’ to be. Each localized ‘I think’ is a de se device of explicit co-reference that grounds a capacity for self-cognition with intrinsic purport of sameness.

But to what does this inner device purport to refer, one may ask? And how does it manage to achieve that reference? I answer that each local ‘I think’ both refers and purports to refer to nothing but the living human being, animal or artifact in whose body is contained the ground for that very localized ‘I think’. A human (or other) body grounds a localized ‘I think’ by being psychically organized in ways that I have been at pains to spell out throughout this essay. In a sense, the reference of a localized ‘I think’ is automatic. As soon as a body is organized in a way that grounds the capacity for self-representation, those self-representations are thereby made to refer. You could say, in the jargon of the trade, that a de se device of explicit co-reference refers *directly* to the human being (or animal or artifact) whose de se device it is. It does not, that is, refer to that body via the mediation of Fregean sense or self-notions or mental files or any other such thing. Rather the very possibility of our having more fully developed self-notions, and the like, asymmetrically depends on the antecedent referential success of the localized ‘I think’.

So I claim that being a center of selfhood, at least a center of raw selfhood, is nothing but being a creature that deploys in thought a de se device of explicit co-reference. If one simply cannot resist the temptation to reify selfhood into a thing, my approach would permit you to do so, but without forcing on you an appeal to any metaphysically mysterious substance. A self, it turns out, is nothing but that which is cognized through a de se device of explicit co-reference. The “unity” of such a self amounts, not to the unity of some antecedent, free standing substance but to nothing but the intrinsic purport of co-reference that is thought through localized and private de se device of explicit co-reference. In the case of the self, representation precedes and makes possible its genuine existence.

Just as the unity of the self is grounded in nothing but representational unity, so the diversity of selves is grounded in nothing but representational diversity. Recall that I argued above that each localized ‘I think’ functions as a private de se device of explicit co-reference. But now consider not the co-reference profiles of each local ‘I think, ’ taken one by one, but the co-reference profile of the generalized ‘I think,’ considered as an abstract type. Considered as an abstract type, the generalized ‘I think’ is not name-like. It is more like a demonstrative or indexical. For considered as an abstract type, the generalized ‘I think’ is a device of de novo reference rather than a device of explicit co-reference That is because each localization of the generalized ‘I think’ is referentially independent, in the sense outlined above, of every other localization of the generalized ‘I think’. The selfhood that is cognized and recognized with intrinsic purport of sameness through the deployment and redeployment of one localized ‘I think’ cannot be cognized as the same again through the deployment of any numerically distinct localized ‘I think’. And it is just the referential independence of each localized ‘I think’ from every other localized ‘I think’ that grounds the boundaries of selfhood. The boundaries of our selfhood are, in a sense, boundaries in thought and representation. I cognize myself through a localized ‘I think’ that functions in me as a de se device of explicit co-reference. You cognize your selfhood through the deployment such a device as well. I cannot deploy the device in me to cognize the selfhood in you. You cannot deploy the device in your to cognize the selfhood in me. At the same time, the selfhood in you is thought through representations of the same general sort as the representations through I think the selfhood in me. That is, we each cognize our selfhood through the deployment of a concrete realization of the generalized ‘I think’. Each concrete localization of the generalized ‘I think’ refers, as it were, de novo to a center of selfhood. There is no de se referential device via which we may cognize two independently given centers of selfhood as the very same center of selfhood again. If that is right, then just as the unity of the self is nothing but the unity that is thought through the deployment of a de se device of explicit co-reference, so to the diversity of selves is nothing but the diversity that comes with the referential independence of distinct devices of de se explicit co-reference.

Now in the general case, it is possible for referentially independent representations to be coincidentally co-referential. Informative statements of identity typically involve just such pairs of representations. So the question naturally arises whether it is possible for two referentially independent de se devices of explicit co-reference to be coincidentally co-referential in this same way. I want to offer a qualified “yes” in answer to this question. First recall that considered as an entity, the self is nothing but a living human, animal or artifact that enjoys the property of selfhood and that a creature enjoys selfhood solely in virtue of having and exercising the capacity for self-representations. Moreover, I have claimed that self-representations refer not to some inner Cartesian entity, but to the living human being whose self-representations they are. Now in a normal, intact human brain, self-representations enjoy, I suspect, a pretty thoroughgoing unity and connectedness. In particular, the totality of self-representations tokened within a single brain, typically constitute what I call a chain of explicit co-reference. A collection of token representations forms a chain of explicit co-reference, roughly, if the tokens in that collection enjoy intrinsic purport to co-refer with one another. But it seems possible that the sort of thoroughgoing unity and connectedness that we find in a normally functioning brain may be fragmented or disrupted. And when this happens we may have something like two independent centers of selfhood separately grounded in one and the same human body and brain. A brain somehow managed to ground two independent centers of selfhood would, in effect, contain two independently functioning de se devices of explicit co-reference. Two independently functioning devices of de se explicit co-reference might produce two disjoint chains of self-representation. We could think of those two disjoint chains as sharing a reference – a reference to the same living human being. But if so, the two chains would be only co-incidentally co-referential, rather than intrinsically co-referential.

It is admittedly hard to fully imagine what it would take to get a single brain to independently ground two wholly functional, wholly independent centers of selfhood.[[17]](#endnote-18) But if a single brain could manage that feat, these two independent centers of selfhood would, in effect, be separate and distinct selves of one and the same living human being. Such a human being would, in effect, have two uncoordinated sets of self-representations, with no member of the one set enjoying an intrinsic purport to co-refer with any member of the other set. They would not therefore be devices by which she could think the same selfhood again as the same selfhood again. Nonetheless, there is a sense in which the devices would co-refer. For each would refer to one and the same living human being. Their co-reference would, however, be of the mere “extrinsic” or coincidental variety. If that is how one’s brain somehow came to be organized one would really have what Kant worried about – a multi-colored, fragmented, self – expect that it would not in any sense be a “single” self. But perhaps we are long past the day, thanks to Parfit and others, when we expected a single body to automatically ground a single self.

**6 Conclusion: Beyond Raw Selfhood**

I have primarily focused in this essay on what I call raw or basic selfhood. I close with an all too brief consideration of the relations among raw selfhood, determinable selfhood, and determinate selfhood. By raw or basic selfhood, recall, I mean nothing but the capacity to generate and deploy self-representations. But that is just the first layer of our full human selfhood. For we humans also enjoy what I call determinable selfhood. To enjoy determinable selfhood is to self-represent oneself as a determinable being. It is to have the capacity to ask oneself what I call “Who am I” questions -- “Who am I to be?” “What am I to do?” “What am I to believe?” But I also claimed that in us humans, at least, determinable selfhood is best understood as a dialectical moment on the journey toward more determinate selfhood. To achieve determinate selfhood is to become, partly through the efficacy of one’s own self-representations, the answer to the “who am I?” questions that are the hallmarks of determinable selfhood. In achieving determinate selfhood, we self-represent ourselves as beings of a certain kind and in the self-representing we make an often indispensible contribution to constituting ourselves as beings of the relevant kind.

It is perhaps logically possible and maybe even really possible for there to be creatures with the bare capacity for self-representation who, nonetheless, are not present to themselves as (self) determinable. Such a creature might self-represent its ever unfolding life, but it would not self-represent that unfolding life as in any sense the result of her own self-determination. Her self-representations would amount to little more than running egocentric commentaries on the passing inner and outer show. I would not want to deny that such creatures are really possible or to deny that they enjoy a kind of selfhood. Indeed, I suspect that some of nature’s creatures may in fact enjoy no greater degree of selfhood than this. Though this minimal level of selfhood falls far short of the full spectrum of our distinctively human selfhood, I conjecture that the step from such minimal or raw selfhood to determinable selfhood is ontologically minimal in the sense that it involves nothing more than a certain kind of refinement in our powers of self-representation. The step from determinable to determinate selfhood is likely more complicated. But it too, I suspect, is likely to be still ontologically minimal. It is no doubt tempting to think that achieving determinate selfhood requires a degree of radical autonomy or freedom of the sort that is not plausibly present in a merely natural mechanism. But I conjecture that in reality determinate selfhood requires no such thing. I conjecture, in fact, that it requires nothing more than an architectural reorganization of the sort of self-representational powers that are already present in less developed and layered forms of selfhood. If that is right, then all three tiers or facets of our full human selfhood turn out to involve nothing but variously structured exercises of our capacity for self-representation -- a capacity ultimately grounded in nothing but the psychic organization of normally functioning brains. Decisively establishing this conclusion, however, is a task that must wait for another occasion.

ENDNOTES

**Works Cited**

Barnett, D. 2010. “You are Simple.” In *The Waning of Materialism*, edited by George Bealer and Robert Koons, Oxford: Oxford University Press (2010)

Bayne T. 2008. “The unity of consciousness and the split brain syndrome.” *Journal of Philosophy* , 105(6): 277–300.

Bayne. T. 2010. *The Unity of Consciousness.*  Oxford: Oxford University Press.

Bayne, T and Chalmers, D. 2003. “What is the Unity of Consciousness?” In *The Unity of Consciousness: Binding, Integration and Dissociation*. Edited by A. Cleeremans. Oxford: Oxford University Press.

Brook, Andrew, "Kant's View of the Mind and Consciousness of Self." *The Stanford Encyclopedia of Philosophy* (Fall 2013 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2013/entries/kant-mind/>.

Block, N. 1995. “On a confusion about the function of consciousness.” Behavioral and Brain Sciences, 18: 227–47.

Campbell, J. 1999. Schizophrenia, the space of reasons, and thinking as a motor process. Monist, 82:609-625.

Davidson. D 1970b, ‘Mental Events’, in Lawrence Foster and J. W. Swanson (eds.), Experience and Theory, London: Duckworth, reprinted in Davidson 2001a.

Dennett, D. 1991 Consciousness Explained, Boston: Little Brown and Company

Dennett, D. 1992. “The Self as Narrative Center of Gravity.”

Dennett, D. C. 1992. “The self as the center of narrative gravity”. In F. Kessel, P. Cole, and D. L. Johnson, eds. Self and Consciousness: Multiple Perspectives. Hillsdale, NJ: Lawrence Erlbaum.

Hume, D. 1739/1968. A Treatise of Human Nature. ed. L Selby-Bigge. Oxford: Oxford University Press.

Ismael, 2007. The Situated Self. Oxford: Oxford University Press.

Korsgaard, C. 1996. *Sources of Normativity*. Cambridge: Cambridge University Press.

Locke, J. 1688/1959. An Essay on Human Understanding. New York: Dover.

Nagel, T. 1971. “Brain Bisection and the Unity of Consciousness,” Synthese, 22: 396-413.

Nagel, T. 1983. The View from Nowhere. Oxford: Oxford University Press.

Perry, J. 2002a. “The Sense of Identity,” in Perry 2002b.

Perry, 2002b. Identity, Personal Identity, and the Self. Hackett.

Perry, J. 2011. In “Knowing the Self” The Oxford Handbook of the Self

Edited by Shaun Gallagher. Oxford: Oxford University Press.

Sass, [L. A.](http://schizophreniabulletin.oxfordjournals.org/search?author1=Louis+A.+Sass&sortspec=date&submit=Submit) and Parnas, J. 2003. “Schizophrenia, Consciousness, and the Self.” Schizophrenia Bulletin, 29(3), pp. 427 – 444.

Sass, L. A. and Parnas J. 2011. “The Structure of Self Consciousness in Schizophrenia.” In Oxford Handbook of the Self. Edited by Shaun Gallagher. Oxford: Oxford University Press.

Taylor, K. 2003. *Reference and The Rational Mind*. Palo Alto, CA: CSLI Publications.

Taylor, K. 2010. “On Singularity.” In Robin Jeshion (ed). *New Essays on Singular Thought.* Oxford: Oxford University Press.

Taylor, K. 2014a. “The Things We Do With Empty Names: Objectual Representations, Non-Veridical Language Games, and Truth Similitude” in Manuel García-Carpintero & Genoveva Martí (eds.) *Thinking and Talking about Nothing.* Oxford: Oxford University Press.

Taylor, K. 2014b. “Names as Devices of Explicit Co-reference, *Erkenntnis*,

Wearing, D. 2005. *Forever Today*: *A True Story of Lost Memory and Never Ending Love.*  Corgi Books

1. John Perry has defended a view similar in spirit over in a number of different papers. See the various essays collected in Perry (2002b), especially Perry(2002a) and also Perry (2011). Though I haven’t worked through all the details, my view also seems to have deep affinities to the view elaborated in Ismael 2007. I am not sure, but one important difference between my own view and the Perry-Ismael view, if I can call it that – is that my view seems, at least on the surface – much more representationalist than their view. I am deeply committed to the view that self-representations are a special kind of representation, with a special kind of role, in the language of thought. I’m not sure whether Perry and/or Ismael fully share that commitment. In particular, Ismael and Perry both appeal to unarticulated constituents. But I have no place for such things. I regard theories that appeal to unarticulated constituents as less than full-throated in their commitment to a language of thought. But this may amount to a merely verbal difference in the end. Whether that is ultimately so, I am not at present prepared to say. But I should say that the main burden of this essay is to spell out the special character of self-representations, considered as items in the language of thought. I claim that in each of us, taken one by one, our selfhood is grounded in our capacity to deploy what I call a de se device of explicit co-reference. But more on this matter in due course. [↑](#endnote-ref-2)
2. I should make it clear that my concern is primarily with certain kind of mental representation and not primarily with expressions in public natural languages. Though certain elements of my view may have implications for the natural language semantics of ‘I’ this essay is not primarily an exercise in the philosophy of language. It is primarily an exercise in the philosophy of mind. At a certain point, though, I shall have occasion to exploit a certain parallelism between the way certain sorts of expressions function in shared natural languages and the way certain sorts of mental representations function in thought with respect to what I call their co-reference profiles. [↑](#endnote-ref-3)
3. See Dennett (1991), (1992), Hume (1968) [↑](#endnote-ref-4)
4. Though she doesn’t use the vocabulary of selfhood and embeds her theory of the self in the context of a model of mind about which I remain neutral in this essay, Ismael (2007) anticipates something rather like the view on offer her. She and I both hold, I think, that being a self is being a creature psychically organized in a certain way. But I admit that I haven't stopped to try to sort out all the agreements and disagreements between her view and mine, however, or to try to figure out whether our differences are merely notional or substantive. [↑](#endnote-ref-5)
5. One person’s ponens is, of course, another person’s tollens. Cartesians will of course object that since the human being is not simple, the self clearly cannot be identical with the human being suitably organized. See Barnett, (2010) for an argument from the supposed simplicity of the self to dualism. [↑](#endnote-ref-6)
6. Perhaps this isn’t quite right as characterization of any actual thought of Locke’s. Though Locke certainly got us all thinking about the distinction between bodily continuity and continuity of the self or the person, he had almost nothing to say about, as it were, the generation of our selfhood. It’s not even clear that he thought of selfhood as something generated, let alone as something generated by the body. For all he has to say, selfhood could be something simply super-added to this or that body. That’s not to suggest that Locke was any sort of Cartesian. He did, after all, take pains to deny that a person or self is a substance of any sort – either material or immaterial. But it’s not really until you begin to see the self as something generated, I think, that you can begin to be adequately puzzled by myriad ways in which our selfhood can be interrupted, diminished, of fragmented. For some examples of such puzzlement, see Nagel () on the effects of commissurotomy on the supposed “unity” of the self. See also Campbell on schizophrenia and thought insertion, or Sass and Parnas, (2003), (2011) on schizophrenia as a pathology of selfhood. Examples can of course be multiplied. One ultimate ambition of the theory I begin to develop in this essay is to provide us with a set of philosophical tools that better enable us to theorize the relationship between the self and the bodily processes that generate our selfhood. A good theory in this domain ought, ultimately, be able to explain both the fully intact fully function self and also fragmentations and diminishments of selfhood. [↑](#endnote-ref-7)
7. For strong evidence that this is indeed what the fictionalist does believe it would take to vouchsafe for the robust reality of the self see the discussion in Dennett (1991) of the distinction between what he calls the Cartesian theater and the multiple drafts model of the self. See also Dennett (1992). On my view, Dennett overestimate what it would take to establish the real existence of selves. For a diagnosis similar in spirit to mine of where Dennett has gone wrong, see Ismael, 2007, especially chapter 12. At the same time, Ismael seems to think that we can rescue something of notion of the self, conceived of as something like a central executive – as long as we let the central executive be a sort of emergent entity. Perhaps -- but also perhaps not. Here is a place where I am just unsure of the extent to which Ismael and I are on the same page. It seems to me that once you have given the metaphysical priority to the property of selfhood and have identified selves with nothing but that which enjoys selfhood, there is no longer any precious to accommodate residual Cartesian intuitions. [↑](#endnote-ref-8)
8. Barnett (2010) f argument for dualism that has just this character and makes just this mistake. See also Bayne (2010) who seems to take what he calls the phenomenal unity of our self consciousness as a sort of basic non-generated fact about consciousness that is directly certified via introspection. [↑](#endnote-ref-9)
9. Something like this temptation clearly drives Nagel (1983) and elsewhere. It also seems to be at the center of, for example, Koorsgaard’s (1996) thinking about normativity. See especially her diagnosis of what she sees as the failures of naturalism. [↑](#endnote-ref-10)
10. In creatures like us, those interconnected roles typically come as a single package. But it possible for there to be creatures in which something like self-representations occupy some but not all of the roles occupied by self-representations in us. Consider, for example, the oft-cited case of Clive Wearing who suffers from so-called anterograde amnesia. Wearing cannot consolidate short-term memories into long-term memories. As a consequence, the horizons of his self-consciousness is confined to the present moment. He does not experience his own consciousness as continuously unfolding, but as always beginning a new, as it were. For the poignant details of his life, see Wearing (2005). Though Wearing still has the power of self-representation, his self-representations do not play one of the roles commonly played by such representations in intact, normally functioning human brains. Could there be creatures that were naturally and normally like is for Wearing? It seems at least logically possible and maybe even really possible. [↑](#endnote-ref-11)
11. One may think here of the distinction between conscious and unconscious thoughts. But thoughts that are conscious in the sense I intend here need not be thought of as phenomenally conscious. They are closer to what Block (1995) has called access conscious. There are other things that philosophers have meant by the “unity of consciousness” than the sort of meta-representational notion of unity, as it might best be called, that I have in mind throughout this essay. For a discussion of various possible versions of the unity thesis, see Bayne and Chalmers, (2003).

    [↑](#endnote-ref-12)
12. Though I do not need to deny that when A believes that p and B believes that not p, then A and B may exert rational pressure on each other. That is partly because at least in a rational mind beliefs, in some sense, “aim” to fit the way the world is. And that means that a roll call of beliefs is not a mere roll call of opinions, which may persists together happily even in disagreement. But the deeper point is that the very possibility of a belief in one mind exerting rational pressure on a belief in another requires that both minds be self-representing (rational) agents, each of whom views him or herself as one such agent among others occupying a common world. But elaboration of this point would carry us very far afield. [↑](#endnote-ref-13)
13. See Brook (2013) for an extended discussion of the extent to which Kant’s theory of mind can be considered functionalist. [↑](#endnote-ref-14)
14. See Taylor (2014) for my most extended argument to this effect. [↑](#endnote-ref-15)
15. Again, I have elaborated this view at considerable length elsewhere. See especially Taylor (2003), (2010), (2014a), (2014b) [↑](#endnote-ref-16)
16. For a further defense of this claim see Taylor (2010) and Taylor (2014). [↑](#endnote-ref-17)
17. But it is far less difficult to imagine the ways in which the unity and connectedness of our self-representations might be disrupted or fragmented. See, for example, Campbell (1999) as well as Sass and Parnas (2003), (2011). Nagel (1971) argues from what happens in the case of commissurotomy that our consciousness is, in fact, *never* really unified. It seems to me that Nagel vastly oversells his case – though I wouldn’t go so far as, say, Bayne (2008) does in criticizing his arguments. What I think Nagel shows is that such unity as our consciousness does achieve is contingent and fragile thing. To be sure, philosophers often seem to be talking at cross-purposes with one another when they talk about the unity of consciousness. See for example, Chalmers and Bayne (2003) and also Bayne (2008) (2010) for defenses of what they call the l “phenomenal unity” thesis. I have already said that I do not mean anything phenomenological when I talk of the unity of the self. Strikingly, armed with a phenomenological conception of unity, Bayne argues in a number of places – e.g. Bayne (2010) and Bayne (2008) -- that none of the phenomena widely taken to show that our consciousness is at least sometimes disunified or perhaps never really unified, suffice to show any such thing. I don’t want to wade into these debates here except to say that I doubt that phenomenal unity is the most fundamental kind of unity or that it is even possible without the kind of representational unity that I have been discussing in this essay. Indeed, Bayne’s view, as expressed in Bayne (2010) that the self is what he calls a “virtual” or “merely intentional” entity seems to require for its intelligibility just the kind of representationalist story that I have been telling in this essay. Strikingly, though, he he rejects representationalism, functionalism, and what am animalistic view of the self. But it is just this package of views about our selfhood that I have attempted to defend in this essay. [↑](#endnote-ref-18)