Normal Service Will Resume Shortly

Content Notes: Mental Health, Neurodiversity, Bipolar Disorder, Plurality/Multiplicity, TERFism, Personal Identity, Posthumanism. Length (~18K). PDF.

0. Vicious Cycles

Another year, another extended absence. What a year though, right? Given how 2020 has demolished any claim 2016 had to the title of 'worst year in living memory', and that 2008 and 2012 weren't exactly peachy, I'm really not looking forward to seeing what 2024 will bring.

For me, last year saw another entry added to the list of ways in which my body is trying to sabotage me, and it wasn't even COVID-19! I can now add mysterious metabolic problems that translate carbohydrates into crippling fatigue to a list that already includes chronic cervicogenic headaches and poorly managed bipolar disorder. Trying to sift discrete symptoms from the cacophony of miserable noise has been pretty difficult, and it's taken a long time not just to glean what was going on but to find a dietary regime that leaves me cogent and capable most of the time.

Worse, this all started after a medication (baclofen) that *really* helped with the above mentioned headaches induced an extended period of hypomania which resulted in a significantly worse depressive crash than usual. (Score one for the hypothesis that mania causes depression.) Add in the nightmare that is caffeine withdrawal, and January-March 2020 was *extremely unpleasant*, even before the pandemic hit and our collective perception of time coiled in upon itself, turning each day into an exercise in coping with indefinite isolation. In particular, watching myself try and fail to deliver comprehensible lectures on Aristotle to first year undergraduates as, unbeknowst to me, my morning croissant slowly sent me into a stupor, felt like some special Sisyphean punishment for my hubris in thinking I could ever be a university lecturer.

For much of last year I felt like an away message in human form: "I'm afraid Pete isn't here right now, but he will be sure to get back to you when he is able. *Normal service will resume shortly*."

1. Absent Selves

The dysphoria associated with catastrophic dysfunction of those capacities with which one most identifies is hard to describe, let alone define. It's a curious feeling of absence. As if one isn't really present, even when one is. Or maybe it's something like a generalised imposter syndrome, in which one is put in the unusual position of impersonating oneself. Everybody has bad days, bad weeks, and sometimes longer slumps, but they don't always produce this type of dysphoric interruption in the circuit of self-recognition. This might be because they don't always come with the fear that the slump won't end, or that the missing capacities won't come back. The fear that, contrary to the message you're repeating to yourself and those around you, *normal service will never resume*.

Alas, this fear can grow and reinforce itself over time, as episodes come and go. It's hard to hang on to inductive evidence that the sun will in fact rise again when the darkness comes, especially when each night seems just a little darker, and each dawn just a little dimmer. Anticipation of the new day gradually sours into mourning for the days gone by. This makes it quite easy for other, less cyclical disruptions to trigger the same sorts of anxiety and dysphoria as the more regular ones. And so, the familiar coping mechanisms kick in: One waits. One defers. One apologises for any inconvenience caused by *the disruption to normal services*. And one promises to get back to everyone once one has gotten back to oneself.

I think that's probably enough introspection for now. The ostensible purpose of this post is to once more let people know why the blog has been silent, and to promise that, in fact, *normal service* will be resuming in some form, at least until the next time it isn't. Last year I'd gotten into the habit of using the blog to collate and comment on some of the things I'd written elsewhere on social media, and though I've been characteristically silent for quite some time, I have a stash of half-finished posts from back then that I intend to revise and release at a pace compatible with my other activities. There's also plenty of new content from Twitter to collate and curate for consumption here.

However, I'm of the belief that, much as we learn more about neurological function by examining cases of neurological dysfunction, we can learn a lot about the nature of self-identity by examining cases in which it breaks down in various ways. It's much easier to treat the *self* as a simple, indivisible substrate of experience when one has not seen how the sausage of selfhood is made, and it's much easier to treat *personal autonomy* as a given when one has not experienced the ways in which its supporting machinery malfunctions. My hope is that those of us who have some first hand experience of the sometimes dysphoric fact that selves are constructed, might contribute to that most (in)humanist of projects — *the task of building better selves*.

2. Abject Bodies

Many such contributions already exist, and they take many different forms. But right now, allow me to exact some small revenge upon my body, wretched meatsack that it is. I have grown more and more critical of the discourse of 'embodiment' over the years, even as it has proliferated across the academic landscape. There are various reasons for this, some of which I've outlined in a recent interview with Anthony Morgan for *The Philosopher*'s issue on *Bodies*. However, there's an aspect of my position that I don't present there, not least because it's more controversial: my growing suspicion that enthusiasm for 'embodied' takes on everything from cognition to political praxis, and the historical narratives that organise them encourage an uncritical valorisation of the body that blends seamlessly into normative naturalism. This is to say that an intellectual tendency which prides itself on the *radicalism* it displays in breaking with tradition can easily be used to support certain forms of *conservatism* we should find deeply worrying.

For instance, I think most of us can get behind the idea of *enabling* people to love their bodies, by working to dismantle beauty standards that not only socially disadvantage those who fail to meet them, but can psychologically harm those who internalise them. But I think we also see that there's a fine line here

between enabling and *demanding*, especially when the actions in question are aimed at changing systemic features of the wider cultural, political, and economic context. We shouldn't demand that everyone love their body, especially not those who are *disabled* by it in some way. If they can love it, and want to love it, we should both allow and support them to do so, but we shouldn't begrudge anyone their ambivalence, uneasiness, or even their hate. This is my body, there are many like it but this one is *mine*, and my hate for it is mine to dispose of as I wish.

However, critiques of elective cosmetic surgery can easily slip over this line and into a more or less explicit *aesthetic naturalism*. In fact, the murky boundary between what counts as elective cosmetic surgery and what counts as expected reconstructive surgery, both philosophically and clinically, demonstrates the extent to which such naturalism already lurks in the background when it comes to the provision and justification of choices made about our bodies. This applies to questions about women's reproductive choices as much as it does to their cosmetic ones, and clearly extends into questions about treatments and surgeries used by trans people to bring their biology in line with their gender identity. There are plenty of bioconservatives agitating along both of these lines, and there are new and quite troubling alliances across them.

Given that philosophical arguments are already being brought to bear in these ongoing political conflicts, we must be scrupulous in denying bioconservatives the ideological resources of normative naturalism on which they subsist. In my view, bodily autonomy (or better: morphological freedom), is the limit-case of personal autonomy as such. It must be defended staunchly wherever necessary and ratcheted aggressively wherever possible. No doubt some will find my worries here to be spurious, as there seems to be a wide gulf between work on enactive cognition, embodied phenomenology, or new materialist feminism and the pet pedants of the transatlantic TERF set. Well, consider the following question: does the demand to *identify* with my body entail a corresponding demand to *love* it?

One might reply that I'm as entitled to hate myself as I'm entitled to hate my body. True enough, but the *rational basis* of this hate is different in each case. There may be good reasons to hate oneself, and good reasons to hate one's body, but are good reasons to hate one's body always good reasons to hate oneself? Of course, there are *irrational* hates, maybe even *delusional* hates, and we might in some sense be entitled to them too, but it's not the type of entitlement that comes from justification. My question is: Can one be justified in hating one's body without hating one's self, or can this only be characterised as a *delusion* induced by the ideology of disembodiment targeted by the radical critiques mentioned above? Is it a symptom of Plato's insidious influence, perhaps proceeding through Augustine's denigration of the flesh? Or is it a symptom of Descartes' disastrous dualism, or the myriad mistaken metaphors for mind it begat in subsequent centuries?

Okay, that's too many rhetorical questions and *way* too much alliteration. Let me simplify the question then: How much of one's body can one rationally hate without the belief that one does not hate oneself becoming delusional? This question is a serious test of the constitutive claims of the embodiment paradigm, because

it threatens every foothold established in arguing for the importance of the body to understanding the mind: I don't hate myself, I hate my tools (extended mind?); I don't hate myself, I hate my environment (embedded mind?); I don't hate myself, I hate my peers (situated cognition?); I don't hate myself, I hate my musculo-skeletal system and the cut-rate motor cortex that steers it (enactivism?); I don't hate myself, I hate my gut, heart, brainstem, and even my sodding amygdala (affect theory?); I don't hate myself, I hate my scarred hippocampus (???); I don't hate myself, I hate my prefrontal cortex and its lousy excuse for executive function (???); I don't hate myself, I hate the stupid variant of the CACNA1c gene lurking in each and every one of my cells (???); fuck it, I hate my whole fucking brain (materialism?). How far down this hateful path is *too* far?

3. Abstract Brains

And so we find ourselves knee deep in the mereology of hatred, trying to work out the sense in which hating a *part* implies hating a *whole*. This is made particularly difficult by the fact that we're considering relations of *functional composition* that aren't straightforwardly *spatial*: my immune system is a part of my body *qua* organism (a subsystem), but it isn't localised in the way that a limb or an organ is (a continuous region), and my genes are a part of my genome (a systemic feature), but they lack even the residual spatiality of subsystems (their instances, though everywhere, are completely discontinuous). To frame this in different terms, we're actually grappling with the conditions governing *identity over time* (in contrast to *continuity in space*), and the extent to which identity of wholes over time (e.g., remaining the same organism) is determined by identity of parts over time (e.g., retaining the same brain).

The most famous problem in this area is the ship of Theseus paradox: wherein each wooden piece the ship is built from is slowly replaced over time, until none of the original pieces remain, before the original pieces are reassembled separately, leaving us with two distinct candidates for identity with the original ship, one that is *processually continuous*, and another which is *mereologically indistinguishable*. There's a variation on this paradox that gets applied to the problem of mind-brain identity, in which individual neurons are gradually replaced with artificial ones, until we've replaced them all. There's rarely any discussion of putting the original neurons back together afterwards, but the worry about processual continuity remains: Is the resulting artificial brain identical with the original organic one? Or perhaps even: Is the resulting artificial *mind* identical with the original organic one, even if they're no longer housed in the same *brain*? This would be to treat the mind as an *abstraction* that preserves the functional properties of the brain.

It's worth appreciating that this variant case implies a corresponding sorites paradox: Exactly how far can the process of replacement go before the mind/brain ceases to be identical with the original, given that it seems as if replacing a single neuron can never make the difference between identity and distinctness? The fact that we take ourselves to persist across strokes, head trauma, and even heavy nights of drinking would suggest that we could easily lose a single neuron without much fuss. However, there's a small disanalogy between the two cases here. The ship of Theseus is a sorites problem only if you don't *functionally differentiate* between the pieces, i.e., if you assume that they're basically fungible bits of wood

that can be treated in the same way. For instance, it's possible to maintain that replacing a plank of decking will never change the identity of the ship, but that replacing its main mast will.

It's only by making the problem recursive — allowing subdivision of the mast into component pieces — that it becomes an unavoidable sorites paradox. The reason this isn't a problem for the neuronal variant is that it has reached a threshold at which further subdivision makes no difference: we see neurons as functionally indistinguishable *atoms* out of which cognitive systems may be built. In essence, we only have a true a sorites paradox when we're dealing with a heap of qualitatively homogeneous *matter*.

This may already be too much mereology for some readers, but it's really only the beginning. Both these cases only consider changes that preserve functional properties of the system as a whole, but some of the changes that organisms undergo don't work this way, most obviously *growth* and *decay*, which involve functional changes that develop or diminish the system's capacities. If we want to talk about minds and brains, then we're also going to have to talk about *learning* and *forgetting* amongst other things. Even if the mind is a functional abstraction, it's one that necessarily accrues significant functional modifications over time: expanding, revising, compressing, and sometimes even shrinking its capacity to recall and process information as it does.

Moreover, there's no reason we can't extend this abstraction beyond the brain narrowly defined to incorporate those functional features of the nervous system, respiratory system, digestive system, musculoskeletal system, local environment, social context, and available equipment which the different strands of the embodiment paradigm treat as *constitutive* features of our cognitive architecture. This means that there's a cybernetic variant of the ship of Theseus for every supposedly indispensable component of the mind, in which we substitute it for a functionally indistinguishable one, creating *concrete cyborgs* with the same *abstract minds*: humans with artificial organs, limbs, lymphocytes, and organelles; people with virtual workspaces, toolkits, colleagues, friends, and maybe even family. Worse, if one accepts that, for the most part, increases in capacity do not effect identity over time (i.e., acquiring a new skill does not a new mind make), then these concrete cyborgs can, in some cases, be genuine *upgrades* of their organic counterparts. To put it in slightly more technical terms, cybernetic subsystems must be able to simulate their organic counterparts, but not necessarily vice versa.

At the bottom of this cybernetic slippery slope lies the possibility of total simulation, or to frame it in more practical terms: mind uploading. For those who refuse to countenance this possibility (and there are many), the only way out is to resist the idea of functional abstraction on which it rests. This means that they think it's *necessary* that they are composed of the same matter across time, or a vague amount of the same matter, at least. There's a rather entertaining argument along these lines made by one of the early church fathers to the effect that lions cannot digest and incorporate human flesh, because God must be able to find and recombine the bodies of all those poor Christian martyrs come the day of judgment. There might even be some who think this *sufficient* for identity across time, and so get quite particular about the disposition of their mortal remains, regardless of their functional composition.

However, it's important to see that the cybernetic substitutions just considered collapse into cybernetic sorites as soon as one suggests that underlying matter makes a contribution to a system's identity conditions that's *independent* of its functional contribution to the system as a whole. And this is where vagueness becomes problematic, because it doesn't seem like there could be any principled way of delimiting precisely which and/or how much matter is necessary to sustain identity. As such, if one is to avoid sliding down the cybernetic slippery slope, one must bend one's *materialism* into a some quite peculiar shapes. As far as I can see, there are really only two principled strategies available to those who insist on the primacy of matter, which I'm going to call the gambit of *indivisible substance*, and the appeal to *heterogeneous matter*.

The first option is, as it were, to find a 'main mast' on which to pin the mind: identify a core component of the organism that cannot be subject to *any* material substitution without breaking the continuity of the mind from one moment to the next (e.g., a crucial region of the brain). However, there are two obvious problems with this. On the one hand, if the reasons for singling out this component concern the functional role it plays in the system as a whole, then one invites questions about *how* it plays this role, which in turn invites questions about *why* its matter couldn't be changed without preserving it. There really aren't any good responses to these questions, beyond the bald faced Searlean tactic of insisting there must be some *special property* of grey matter we've yet to even comprehend the possibility of, let alone actually understand. On the other, the resulting position is weirdly homologous to those theories of mental substance to which materialism is nominally opposed. The only thing that differentiates it from postulating an immortal soul is the admission of mortality. Yet even then it's a very peculiar sort of mortality that's more precarious than regular death, insofar as the slightest material change in the substance of your indivisible soul is enough to end you, even if the new soul that results is blissfully ignorant of their birth. It's as if one discarded Descartes' theory of thinking substance, but retained his speculations about the role of the pineal gland as the seat of consciousness.

The second option is much more subtle. It consists in maintaining that the very idea of 'homogeneous matter' on which the sorites paradox depends is untenable. This is the favoured choice of Deleuzians, who are ever eager to demonstrate that difference surges beneath every seeming sameness. To give them their due, there is some logic to denying the Aristotelean proposition that matter is essentially *passive potential* waiting to have form *actively imposed* on it from above, if only because the thermodynamic miracles from which the self-organisation, ramifying mutation, and continuing evolution of life spring seem to bubble up from below. The material strata of complex behaviours that have assembled themselves in the billions of years since the quark-gluon plasma cooled down enough to form familiar configurations of particles seem to suggest patterns of emergent causation in which *molar structures* more or less robust in relation to random fluctuations in their *molecular substrate* can nevertheless feed on these fluctuations in a way that leads to substantive novelty. On this basis, the proponent of heterogeneous matter can claim that it's not just the structure of cybernetic signals that's important to the identity of a system over time, but the texture of the noise through which they surf.

I've seen this argument made several times to deny the possibility of transferring cognition from one material substrate to another. Those making the argument are usually content to insist that the certain existence of behavioural differences between these substrates, no matter how small, are sufficient to produce significant and unpredictable divergences in overall behaviour that undermine any claims to persistence across the transfer based on functional indistinguishability. To frame the argument in more plain language: subtle differences in the way a cyborg body responds to its environment that seem irrelevant to its overall functioning will inevitably produce differences in long term behaviour compared to the original organic body. The feel of my original flesh is like the warm crackle on a vinyl record, a seemingly unquantifiable uniqueness that somehow guarantees the *authenticity* of the experience.

I'm unsurprisingly unsympathetic to such investment in somatic authenticity. If nothing else, these myriad differences lurking beneath the functional level are always smaller than functional differences that supposedly make no difference: if I can have my arm blown off, my liver transplanted, or even suffer significant brain damage without a consequent discontinuity of identity then borderline infinitesimal changes in my material substrate aren't going to make any significant difference, no matter how one plays up their absolute singularity or holistic character. There's a vast range of things that can happen to me which will drastically change my long term behaviour more than the statistical patina that functional structure abstracts away from. Furthermore, Deleuzeans should know better than to use his metaphysics to defend *any* claim to identity over time, as if it could be more than the effect of some transcendental illusion disguising traces of an underlying dynamic of repetition.

There remains a third, much weaker position that's still on the table, but I don't think it's very satisfying. It's what we might call the criterion of *minimal spatio-temporal continuity*. It's always possible to maintain that there can be no identity over time unless there's some spatio-temporal overlap that proceeds by sharing of components that are themselves identical over time. An infinite regress threatens here (i.e., components of components of components of...), but it's not an intolerable one. Anyone this committed to preserving their intuitions about the necessity of material continuity will happily accept this notion as an explanatory primitive. Another way of framing this principle is to say that identity *between* distinct times requires identity *over* the intervening times. No instantaneous teleportation or mind-uploading is allowed, because it would create an unacceptable *discontinuity*. Nevertheless, there are still plenty of weird edge cases that will violate the intuitions they want to preserve, including fissions and fusions of every imaginable shape.

The major upshot of this position is that it permits a more elaborate ranking of candidates for identity by the extent of material continuity. The clone fortunate enough to retain a single extra cell from my original body gets to lay claim to the title, like an eldest son born seconds before his twin (or *n*-plets). This should be cold comfort to our material primacists (or is it 'material supremacists'?). It's more of a procedural hack designed to make the social bureaucracy of tracking who's who run more smoothly than a principled decision about the nature of personal identity. A necessary but minor choice made in the assignment of variables (x_1, x_2, x_3... x_n).

4. Abased Spirits

Wait a minute... are we talking about the identity of *minds* or the identity of *selves*? When exactly did we return to the topic of personal identity? We started with questions about whether parts of the body are parts of the self, which lead us to questions about the persistence of the body and its parts over time, from which, by a process of functional abstraction, we eventually arrived at questions about the persistence of the mind over time. But somehow, in the process of rebutting objections to such cybernetic functionalism, we slipped from the language of 'minds' back into the language of 'selves', as if there were no real distinction between the two. The question is, are they really the same thing? Could there be minds without selves, or selves without minds? Could there be minds with multiple selves, or selves with multiple minds? What are the parameters of the relation between minds and selves, if they aren't simply identical?

Of course, this terminological slippage was deliberate. But the rhetorical questions it invoked make a serious point about the ways in which certain debates collapse into one another when one hasn't discerned the underlying motivations of one's intellectual opponents. All the points about the mereology of minds in the last section are sound, but the objections they anticipate aren't really coming from opponents who care about the identity conditions of minds *independently* of the identity conditions of selves; they come from opponents who care about the identity conditions of minds only insofar as they *determine* the identity conditions of selves.

Our material primacists don't care about whether or not a cybernetically enhanced/computationally simulated brain might house the same mind, but only whether or not someone who has undergone the relevant procedure is still the same person, and they usually only care about this because they have strong introspective intuitions about the *continuity of consciousness*. They're generally not worried about any discontinuities produced by natural sleep, or even those induced by general anaesthesia, but they're terrified either of going under the cybernetician's knife or undergoing a destructive brain upload, because they fear that whatever regains consciousness on the other side will not be *them*. This leaves us with yet another term to distinguish: Is a consciousness the same thing as a brain, mind, or self, or is it something else all together? What if our introspective intuitions about 'consciousness' are a confused and largely unhelpful addition to these debates?

At this point it's worth restating some of my own worries about philosophical debates on the topic of personal identity. These debates take various forms, but more often than not they divide the issue in two: first, there are *metaphysical* questions about what kind of thing a person is, and under what conditions they persist across change, and second, there are *normative* questions about whether or not a person's rights and responsibilities persist across such changes. Moreover, the normative questions are seen as essentially downstream from the metaphysical ones: work out what kind of thing a person is (e.g., whether they must exhibit continuity of memory) and this will give you the resources to answer the relevant questions about rights and responsibilities (e.g., whether someone with permanent retrograde amnesia can be held responsible for acts they can't remember committing). My view is that this way of framing the issues

hypostasises selfhood in manner that disconnects it from the normative questions which properly define it, and in so doing forces us to fall back on vague intuitions about things like continuity of consciousness, before returning to the normative questions with theories reverse engineered to meet these intuitions.

This isn't to say that such theories can't produce counter-intuitive conclusions, only that more often than not these are trade-offs between intuitive priorities forced by the conceptual constraints they're operating under. For instance, there's a possible trade-off between *continuity* and *uniqueness*. On the one hand, one can make minimal material continuity more satisfying if one abandons uniqueness: this means that *multiple* persons can be continuous with a past person, perhaps through some process analogous to asexual reproduction. None of the resulting clones is strictly *more identical* with the original than the rest, even if they might be *more similar* in certain respects. This makes continuity with our past selves less a matter of *identity* than one of *descent*. On the other hand, one can make the appeal to heterogeneous matter more satisfying if one abandons continuity: this makes each person a Heraclitean river, changing from moment to moment, but whose contingent enmeshment in their environment makes them *absolutely singular*. No attempted copy can ever be the *same*, no matter how *similar*. This makes difference from one another less a matter of *quality* than one of *context*. One can even recombine these positions on the fly, as some Deleuzo-Guattarians are wont to do, by insisting that we are a veritable swarm of selves: there's a thread of continuity whenever we want one, but no way to isolate and copy it whenever we don't. We can all have our cake and eat it too, insofar as, given that we contain multitudes, the one having the cake is never the one doing the eating.

It's important to see that, in each of these cases, implicit motivations have been laundered into explicit metaphysics. This desire for a metaphysical guarantor of the intuitive features of selfhood is precisely what motivated our ancestors to postulate the *soul*, even if this resulted in a variety of models of its composition beyond the simple, singular, indestructible version on which the Christian tradition ultimately settled. Indeed, what's really interesting about these more complex models is that they're more obviously folk-psychologies, in which the different component parts of the mind are intended to *explain* different aspects of human behaviour. This means that there can be separate parts of the mind associated with outward bodily behaviour (e.g., reflexes and urges) and inward mental behaviour (e.g., contemplation and volition). This seems to be precisely what happens in the Chinese tradition, where the *po* soul articulates the passive aspect of the relationship between mind and body (*yin*), while the *hun* soul articulates the active aspect (*yang*). When the question of immortality gets posed within this framework, it's subordinated to these explanatory concerns. It's not exactly surprising that the contemplative *hun* is then deemed able to persist and migrate after bodily death, leaving the *po* behind, if only because the same idea emerges in the Platonic tradition, and ultimately feeds into Christian theology.

We can now see a more general tension at work in the evolution of theories of mind, be they folk psychological, theological, or more thoroughly philosophical: a conflict between the need to *functionally decompose* the mind into its component subsystems in order to *explain* our behaviour, and a need to *impose limits* on this process of decomposition in order to *preserve* our intuitions about our persistence as

unique persons over time. As such, there must always be some primitive, indivisible part of the mind that guarantees unique persistence — some inexorably tangled knot of selfhood — but the more psychological machinery one pulls out of it in one's quest for psychological explanation the less one can claim belongs to its ravelled essence. To put this in different terms, there's a trade-off between *psychological explicability* and *quintessential personality*. The more one wishes merely to be who one is, the more must be locked in the black box marked 'self'; and it doesn't matter whether one writes 'self' instead of 'immortal soul', because mortality is a secondary issue. This means that the attempt to accuse transhumanists and fellow travellers of a crypto-theological belief in an immortal soul, while not entirely off the mark, often disguises its own theological impulses. The rush to metaphysics, materialist or otherwise, is not so much a way of exploring the problems of personal identity as it is a means for making those problems go away.

So, what get's lost when we hypostatise selfhood in this fashion? Well... this returns us to those questions about the relations between minds and selves with which this section opened, and the even more general questions about their relations to bodies with which we are concerned. My suspicion is that, in bypassing the requisite *functional abstractions*, the slippage between 'mind' and 'self' has ignored the *concrete possibility* of alternatives to the types of bodies, minds, and selves with which we are familiar, and perhaps even rendered the *matter* they're composed from into some homogeneous metaphysical soulstuff — a brute material guarantee of those introspective intuitions to which many are so attached. That these views often go hand in hand with some form of vitalism/panpsychism is thus not entirely surprising. It's simply a further elaboration of the metaphysical limit implicit in the desire to preserve an intuitive connection between personal uniqueness and continuity of organism/consciousness. The rising popularity of animism in certain circles, the justification for which is often little more than a lazy association made with some supposedly generic subaltern (or 'indigenous') worldview, demonstrates the perennial appeal of such *metaphysical self-deception*. In essence, the commitment to material primacy is often a disguise worn by latter day spiritualists less invested

5. Aberrant Minds

So, let's return to the concrete possibilities: Could there be one mind with multiple selves sharing cognitive subsystems? *Prima facie*, this would seem to be what's going on in cases of dissociative identity disorder. Two or more distinct personalities that share a certain base set of cognitive capacities, but with episodic memories, personality traits, and sometimes wider skillsets that diverge after some initial (often traumatic) schism. However, its status as a *pathology* invites interpretations that position it as a sort of *delusion*: one person under the mistaken impression that they are two, or even more. It doesn't matter how elaborate this delusion is, how extensive the information partitioned between personalities, or how radical the divergences between their behaviour; it can still be presented as a *dysfunctional self*, rather than *several* functional ones. Nevertheless, there is an alternative to this pathological model: there are persons, or systems of persons, who *identify* as several distinct selves consensually sharing a single body. These persons represent a small and less well known segment of the neurodiversity movement that grew out of people diagnosed with autism resisting its pathologisation. Over the past few decades the internet has facilitated

the growth of a thriving online ecosystem of self-identifying neurodiverse individuals with a continually evolving taxonomy of phenotypic variations; borrowing terms, ideas, and strategies from more established minority communities in the fight for recognition and accommodation. What are we to make of these contrasting attitudes?

It's no secret that I'm a big supporter of the neurodiversity movement. I have a lot of friends out on different fringes of the neurological map, from 'high functioning' autism, through late life ADHD diagnoses, to schizophrenia, schizo-affective, borderline personality, aphantasia and *yes*, even plurality. Some of these friends consider themselves to be neurodiverse, and some of them consider themselves to have mental health problems. Not all, but *most* of them, would put themselves into both categories, and would *not* consider their mental health problems to be entirely separable from their neurodivergence. And what about me? Does my capacity to temporarily overclock my cognitive abilities at the expense of affective dysregulation, executive disinhibition, and subsequent cognitive crashes qualify as neurodivergence, mental illness, or both? Does this not further bias my attempt to adjudicate the border between pathological dissociation and divergent plurality? I have extensive and somewhat complicated views on these topics, but this post is already long enough without expanding its scope any further.

For now, I propose a methodological compromise. Let's consider the opposing extremes in the range of positions that can be taken on the reality of plurality (also termed multiplicity). At one extreme, one can maintain that no matter how seemingly functional or insistently identifying a system is, there is strictly speaking only one person at issue. At the other, one can maintain that no matter how seemingly dysfunctional or inconsistently identifying a system is, there are precisely as many persons at issue as they present themselves as being. It should be no surprise that this opposition parallels the extremes adopted in those debates about the validity of trans identity that we touched on on earlier, and I suspect people's opinions on the one will correlate with their opinions on the other. However, I think that we're now touching on something deeper, insofar as we're not simply addressing the connection between the type of person one *is* and the type of person one *claims to be*, regardless of what we think about such types (e.g., sex, gender, race, class, etc.). One can't treat systems as a type of person without thereby deciding the question against them. They aren't a type of *person*, they're a type of *mind* distinguished by the fact that it contains multiple persons (of potentially differing types!).

Given this framing, my aim is to rule out the first extreme without endorsing the second. This is to insist that there are *true* multiples in three distinct senses: i) that it's true that *some* minds contain multiple selves, ii) that we can be mistaken about *which* minds contain multiple selves, and iii) that this might entail the existence of uncomfortable boundary cases in which there's a conflict between the way a mind is *sincerely* presented (as containing multiple distinct selves) and the way it *really* is (not fully multiple). This last point is the most tricky, because it risks invalidating the (multiple) identities of self-identifying systems. This is where the *rhetorical* parallel with debates about trans identity is unavoidable, even if we recognise that the *logical* parallel is less strict.

Nevertheless, no matter how tricky they are, it's important not to avoid these issues by adopting the second extreme, for to do so is to reduce the relevant truths to more or less meaningless trivialities. It's precisely this mistake which gets rhetorically exploited by concern trolling 'gender critical' feminists when they frame their position as a form of gender abolitionism. In essence, they claim that persons should be allowed to identify as whatever they wish, but only on the condition that the terms used to articulate these identities (i.e., 'man' and 'woman') have been emptied of *any* meaning that might have practical ramifications, which, in the limit, is *all* meaning — the complete abolition of gender is always kept conveniently *just* beyond the horizon of present political concern, while the practical work of deconstructing it is (temporarily) *indexed* to the traditional meaning of terms it will (eventually) abolish.

In order to defend against this rhetorical strategy, and to secure the meaningfulness of claims that enunciate parameters of selfhood, we must insist on the logical point that such claims can be made in error. This has deeper consequences for the pragmatics, epistemology, and semantics of these claims, which I'll return to. But for now, my methodological compromise is to avoid the boundary cases where such errors must be adjudicated as much as is possible. Even so, plurality confronts us with real questions about the relationship between minds and selves that can only be addressed by abandoning our metaphysical biases and considering what can and can't *work* in principle, i.e., by articulating these relationships in *functional terms*. What's so significant about plurality is precisely that it really does seem to work in many cases, even though we haven't yet figured out *what it means* for some configuration of minds, bodies, and selves to 'work' in the first place.

Of course, one could always respond that *the very distinction* between working and non-working mental configurations is illicit. This is a very popular position in circles that draw inspiration from 'post-structuralism', which in practice usually means some mash-up of Foucault, Deleuze & Guattari, and maybe even Lacan, amongst others, filtered through a frame articulated by Derrida. The results can be more or less nuanced, but they often involve aligning and combining a range of simple symbolic oppositions (e.g., speech/writing, mind/body, male/female, human/animal, etc.) into a single overarching 'dualist' worldview that privileges one side over the other (what we might call 'deconstruction by numbers'), while assimilating a variety of complex normative distinctions (e.g., normal/pathological, sane/mad, legal/criminal, straight/queer, etc.) to the same model in such a way that the imperative to reject dualism calls into question the very possibility of making valid normative distinctions (we might call this 'transgressive genealogy'). The paradoxical character of this blanket normative judgement about normative judgements (i.e., some variant of 'norms are bad') is not just acceptable, but sometimes even attractive in these circles, insofar as it motivates an *immanent ethics* of transgression, in which each *substantive* value is overridden by a *formal* imperative to subvert every norm.

This might seem like a hasty caricature, but I've encountered this view in person time and time again, though the references vary from exponent to exponent. It's a perennial fixture of art circles, precisely insofar as it generalises the *immanent aesthetics* of transgression that remains when one subtracts any specific commitments to compositional mediums, guiding concepts, or practical projects from the historical

trajectory of contemporary art. I have a particularly vivid memory of trying to defend the seemingly innocuous proposition that we might *need* distinctions between good/bad, true/false, and just/unjust simply to practically orient ourselves against a hostile room full of artists and theorists during a private seminar several years ago. Still, I don't want to paint this position a self-evidently ludicrous, even if I do think it's pernicious. It's important to see that its appeal lies in an abstract demand for *radical autonomy*, which can only be concretely expressed in acts of *performative transgression*: true Art consists in nothing but violating the limits implicit in every extant aesthetic configuration; and true Ethics consists in nothing but refusing every externally imposed constraint. These ritualised invocations of negative freedom are in some ways the purest expression of the residual ideals of the liberal political order. However, there's another sense in which they reject this order, in principle, if not in practice.

As I've explained elsewhere, the essence of liberalism is a demand for freedom that refuses to articulate *what freedom is*. It may be secular in the sense that it's *agnostic* about whether or not our capacity to choose for ourselves is a gift of divine provenance, but it remains thoroughly *gnostic* about the nature of this capacity and its enabling conditions. Liberalism remains bound by those implicit introspective intuitions about the nature of personal identity which we've been exploring, but it codifies them in legal systems and jurisprudential frameworks rather than transposing them into metaphysics (though economics has historically acted as a bridge between the two). It's in the hastily patched edge cases where one sees the liberal conception of personhood begin to fray: in the way it conceives children whose autonomy is still in development, in the way it treats the elderly whose autonomy is now in decline, and in the messy patchwork of measures designed to manage those instances of neurological divergence and dysfunction with which we're concerned. If one wants to understand the reason that every liberal violation of liberal principles has been framed in the language of paternalism, one simply needs to note that the relation between parent and child is the only model it inherited for dealing with freedom as a conditioned, created, and cultivated object; a model that is neither absolutely immutable nor even relatively fixed during the relevant historical period.

By contrast, the 'radicalness' of these poststructuralist tendencies consists in their willingness to push the demand for autonomy beyond the limits of these liberal frameworks, and indeed, any such framework. There's no single way of working out what this entails. There's a range of options to choose from, running from those essentially indistinguishable from liberalism, through those articulating types of minoritarian praxis, to those suggesting full blown alternatives to liberalism. However, this detour into political philosophy is long enough already. What we're really interested in are the ways in which they transform the liberal subject (or soul), which are all to some extent ways of destabilising, fragmenting, or otherwise calling it into question. To be more precise, when they are confronted with the edge cases that liberalism attempts to *legislate* in a manner consistent with its 'common sense' intuitions about personhood, their response is to treat such legislation as in principle illegitimate and to *subtract* the contested features from the underlying common sense. They whittle away every supposedly essential element of selfhood (e.g., sex, sexuality, rationality, etc.), and compensate by multiplying the 'positions' that this *streamlined subject* can occupy (e.g., gender, orientation, discursive situation, etc.). If the liberal subject is the culmination of the

characteristically *modern* concern with the personal autonomy that began with Renaissance humanism, then this streamlined subject might properly be called *postmodern*.

The problem with this (postmodern) position is that it leaves no distinction between *conditions* of identity and *co-ordinates* of identification. While liberalism struggles to get any purchase on the functional prerequisites of personhood distinct from its image of *normality* (e.g., to distinguish rationality from neurotypicality), its radical successors insist that every posited prerequisite is nothing but prejudice against *abnormality* (i.e., that rationality is an ideal on par with toxic beauty standards). The concept of agency is emptied of all content in the name of autonomy: any attempt to articulate the essence of *freedom* is itself interpreted as a gesture of *oppression*, not least because the very notion of essence has been deemed irredeemably corrupt. This is to say that the gnosticism implicit in the liberal soul has gradually made itself explicit, as its inherent nothingness is slowly revealed, and we are driven closer and closer to an apophatic conception of self-understanding. Though not necessarily metaphysical, this is eminently compatible with the weirder materialist currents discussed in the previous section. The latter provide a *theoretical* justification for apophasis, which is an essentially *practical* orientation. As such, it doesn't much matter which flavour of materialism is chosen, insofar as the resulting mysticism cares less about the location of the spiritual than its (essential) ineffability.

What of plurality then? Well, the radical/postmodern/apophatic tendency pushes towards the second extreme. Of course, it admits that our self-understanding might be wrong, and thus that enunciations of identity/identification (e.g., "I'm the same person I was when I married you."/"But I'm (now) a woman.") might err — perhaps even *must* err — but it removes any possible basis for articulating and addressing such errors beyond further enunciations (e.g., "I was wrong, I'm not the same person anymore."). It avoids treating plurality as a variant type of selfhood only insofar as it subtracts *unity* from selfhood *as such*. We cannot speak of the self as *either* one or many, but only as *both* not-many and not-one. This avoids disenfranchising the multiple by recognising and even valorising the ways in which we're all already fragmented. Yet this threatens to do what we cautioned against earlier, namely, render claims about 'plurality' effectively meaningless. Imagine the following exchange taking place at a party, between an enthusiastic grad student who's read a lot of D&G and a system of persons who couldn't care less about such things:

> A: "You're multiple, that's so cool! Everything's a swarm... a flux... a shifting pattern of contextual relations. Down with substances, dualisms, hierarchies, norms... and all that shit! I'm a swarm too!"
>
> B1: "We're not a swarm, *thankyouverymuch*. We're a well defined and quite stable quartet. I'll have you know it takes some concerted effort to make this arrangement *work*."
>
> B2: "Yeah! It's not all taking psychedelics and staying up till four in the morning talking about potatoes or whatever. Get a life, or several, you fucking tourist!"
>
> B3: "Go easy on him 2, he's obviously got enough problems as it is..."
>
> B4: "Not gonna disagree with you 3, but.... Listen, A? *You do you*. If you want to be a swarm, go nuts. Just remember that we're talking about real lives here, and real choices, not some abstract template you can apply to

everyone regardless of how they live or what choices they make."

No doubt some readers will be thinking that I've just used imaginary lives to justify *my own* abstract template, but bear with me here. What's the obvious apophatic reaction to this exchange?

I think it would probably be to diffuse this disagreement by making the meaning of 'plurality' plural — allowing for multiple models of multiple-ness (swarming, non-swarming, and maybe more...) that can coexist without conflict. Each party is allowed to be *equally* right, precisely insofar as each is always *slightly* wrong. They are neither one nor many, but their non-unity may be further differentiated ('swarm'/'non-swarm') by a never-ending sequence of negations that points in the direction of some deeper meaning without ever enclosing it. But this does exactly what is being complained about, namely, impose an abstract template that ignores the concrete question at issue: Not simply, is there some vague, protean *many-ness* here? But rather, are there definitely many *persons* contained in this mind, or cohabiting in this body?

The (essential) vagueness of apophasis creates a fundamental asymmetry between the casual swarm and the dedicated system that can never resolve into the desired equality. If we *validate* the (precise) content of the system's multiple-identities, then we retain a distinction between a single person identifying as (vaguely) many (which is ultimately a type of person), and several persons identifying as distinct from one another. Yet if we reject the idea that identifying is something done by (unified) persons, in order to insist that the swarm and the system are doing the same thing, then we *invalidate* the (disparate) content of the system's professed identity. Even if we acknowledge that there are many different ways of being multiple, and even interstitial states that a mind passes through on its way from one to many selves, there's still no way to resolve this asymmetry. A swarm might be an incipient system, in passage from one to many, much as an infant is an incipient person, in passage from zero to one (or more). None of this makes any difference to the *precision* of these numerical distinctions once distinctness has been achieved. The question of *when* such passage is complete might be irreducibly vague, but the question of *what* has thereby been completed needn't be. Compare with some more familiar sorites cases: When does a fetus become a child? When does a child become an adult? When does a collection of cells become a person?

Of course, I may be reading too much into the poststructuralist positions I'm considering by insisting that they are strictly apophatic. However, there's a more general way of stating my objection that rests on what I've quite deliberately called their postmodern character, which is to say, the manner in which they attempt to radicalise liberalism by projecting its demand for personal autonomy beyond the reach of every possible liberal order. The point is this. **One cannot defend personal autonomy by dissolving the notion of personhood.** This is what it means to reject the claim that any definition of autonomous agency is inherently oppressive — the concept of person has an insolubly normative core. We can't unmoor metaphysical speculation about personal identity from normative considerations regarding transmission of rights and responsibilities between candidate persons, because it's the need to *maintain* these lines of transmission that gives us our basic functional purchase on the concept of personhood. The casual swarm may contain unstructured *fragments* of personality, but the dedicated system sustains structured *loci* of

integration — it incorporates a process that does the essential work of keeping track of *who* is responsible for *what*.

There's a fairly short argument for this idea, though it depends on the principle of ought-implies-can, which I've defended elsewhere. If every responsibility implies a corresponding capacity, such that one cannot be held responsible if one is not capable of carrying out the responsibility, then there is some minimal set of capacities underlying every responsibility, namely, those that enable us to *keep track* of what we are responsible for, and to *fulfil* these responsibilities when their occasions arise. It doesn't matter if the mechanisms underpinning these capacities are cognitive subsystems shared by multiple selves within the same mind. It doesn't even matter whether these cognitive subsystems are properly unconscious, or whether they involve some conscious effort on behalf of the selves in question. All that matters is that they enable the contributions of these selves to be *reliably differentiated* from one another. There are two questions that naturally follow from this.

Firstly, what other capacities are implied by these basic ones? The ability to keep track of our responsibilities entails a capacity not simply to maintain an *integrated representation* of the world as it is, but also of the world as it should be. This differential between the *real* and the *ideal* is the force that moves us to action. It doesn't matter whether we're talking about simple *sensorimotor expectations* (e.g., the bitter tang of coffee) trickling down hierarchies of control systems (e.g., the regions of the motor cortex controlling the various muscle groups in my arm) until our perceptual input matches its reference signal (e.g., I've put the cup to my lips and taken a sip), combinations of persistent *drives* and transient *affects* (e.g., hunger and anxiety) emitting and modulating the relative intensities of these impulses (e.g., ratcheting thirst for caffeine but suppressing appetite), or complex *practical commitments* (e.g., democratic socialism and parenthood) demanding iterative elaboration of their causal consequences (e.g., canvassing, party meetings, economic policy) and navigation of their mutual incompatibilities (e.g., spending time with one's children, sending them to private school); it doesn't even matter how well these layers of motivational abstraction are integrated into a smooth picture of what we should do, or whether this involves flattening them into fungible quantities of subjective utility; they are united in bridging the gap between recognising the real (*truth-taking*) and realising the ideal (*truth-making*).

Secondly, how reliable must these capacities be in translating avowed commitments into successful actions? There are no doubt a wide variety of concrete dysfunctions in the above mentioned mechanisms that can result in a failure to realise ideals implicit in the responsibilities one acknowledges, but we can divide them into three basic types: failures of *ability*, failures of *understanding*, and failures of *volition*. In the first case, our actions are obstructed by some brute failure of the mechanisms that let us achieve certain goals, such as we might ascribe to inherent *talent* (e.g., physical dexterity) or acquired *skill* (e.g., playing the piano). In the second, they are obstructed by an incorrect, inconsistent, or merely incomplete grasp of the consequences of our commitments, which invalidates our *plans* for achieving them (e.g., not understanding that 'doing the washing' entails separating colours from whites beforehand, and so ruining them). In the third, they are obstructed by a weakness of the will, or a disconnect between an adequate

grasp of our responsibility (e.g., knowing one should practice piano/do the washing) and the mechanisms that translate this into the impulses driving our behaviour (e.g., being unable to drag oneself away from the TV).

The question is, *how much* of each type of failure can be permitted before it invalidates one's ability to be treated as responsible in the relevant ways? It seems that we've stumbled into a nest of sorites problems. However, insofar as we're not here to adjudicate the details of *specific* commitments (e.g., practicing piano/doing the washing), but trying to talk about commitment *in general*, we can narrow the scope of this question quite considerably. Specific commitments are more often matters of *practice* than matters of *principle*, to be left to experts in the relevant domains (e.g., music/laundry) rather than philosophers unfamiliar with them. So, though we can defer questions about the reliability of mechanisms that track and fulfil specific commitments, we have to say something about the reliability of those mechanisms involved in tracking and fulfilling commitments as such. It seems uncontroversial to say that we can't count someone who *never* acts upon an avowed commitment as responsible. The same might be said of anyone whose actions are at best *randomly correlated* with their stated intentions. Indeed, the point is that it'd be difficult to count them as 'someone' in the first place. These are cases of *absolute dysfunction*. The question is then, *relatively* speaking, just how unreliable do these general mechanisms have to be before they cease to support even a dysfunctional self? If we extend this to the case of plurality, it is rather, how much of these different types of failure can be attributed to dysfunctions in *cognitive overlap* (i.e., shared abilities/understanding) and *agential differentiation* (i.e., separation of distinct wills) before it no longer makes sense to talk of a dysfunctional system of selves, but only a dysfunctional mind without *any* persistent self?

Thankfully, as interesting as these questions are, they run up against the methodological compromise I proposed earlier. I don't want to have to take a position on these most delicate issues, any more than I want to decide the exact moment at which a fetus becomes a potential person. I aim only to elucidate the relations between these questions, and to describe the dialectical terrain in which they're situated. Nevertheless, one might say that these are also questions of practice, rather than principle, but that the practices in question aren't so much those involved in *training* and *calibrating* a person's capacities to perform specific types of task, as they are those involved in *rearing* and *educating* a person who could undertake to perform such tasks at all. This returns us to the great lacuna of liberalism: *childhood*. As things currently stand, there's essentially no principle to be found here, only best practice, and there's precious little of that as it is. This is another issue I'm going to have to put a pin in, but I'm far from the only one who thinks that childhood is a loose conceptual thread that unravels every ambient ideology if we pull on it hard enough. Far from being miraculous, the concrete genesis of freedom is the one issue that can undermine every ersatz spiritualism or emergent mysticism.

6. Alien Personae

What else is to be said about the relation between minds and selves? Asking whether one mind can have multiple selves has gotten us this far, but to go any further we must turn the question around: Can one self have multiple minds? This is a far more contentious question, because it remains stubbornly *hypothetical*. There's a wealth of speculation about this possibility in science fiction and elsewhere, but no extant exemplars whose demands for recognition would force us to settle the issue. However, the question gives us new purchase on the relation between selves and bodies from which we began, insofar as a single self with multiple minds is, *prima facie*, a self with multiple bodies. It seems plausible that there might be minds with multiple bodies, insofar as elements of these bodies taken together might constitute a single distributed mind *qua* concurrent communicating system. But it's not clear that this is the only way we can interpret cases in which there are many bodies that share a single self. Indeed, if we push this idea of a mind constituted by multiple subsystems communicating concurrently, we quickly get into situations in which it seems like these subsystems must be minds in their own right.

For example, say that I get up one morning, and realise that I have five different things I need to do today, that cannot be simultaneously achieved by a lone embodied human being. So I do what any reasonable person would do in the circumstances, and *fork* my consciousness into five qualitatively identical threads running in parallel on five qualitatively indistinguishable bodies. Of course, this is one of those controversial hypotheticals I mentioned above, and there'll no doubt be many people ready to give me hell about even contemplating using terminology from computer science to articulate its parameters. Some people won't be satisfied until *literally* five of me knock on their door at 2AM and start demanding they refer to me using the correct second person singular pronoun ('Why are *you* here Pete?' not 'Why are *yinz* here Petes?'), or at the *very* least the correct plural noun ('Why am I being accosted by a *flock* of Pete?'). There'll also no doubt be some people who think I've just made a joke at the expense of those who insist on correct pronoun usage ('How dare you do that Pete!), whereas I'm actually making a joke about how, if I could fork myself into five concurrent copies, I'm precisely the sort of (singular) person who would turn up at your door at 2AM and want to talk about correct pronoun usage when addressing *concurrent forks* ('You seriously *would* do that Pete, wouldn't you.'). Some people will never be satisfied by anything less than brute actuality. For the rest of you, read on!

Returning to the example, we might wonder what sort of concurrent interaction between these forks sustains their joint identity. Must there be some form of real time *communion* between brains in order for this to work (e.g., such that 'I' can simultaneously see out of each set of eyes)? Or is the capacity to *communicate* with one another in the same manner as fully separate persons sufficient:

1: "Hey Pete!"

2: "Back at you."

1: "I'm off to re-read Kant."

2: "I'll tackle the washing up."

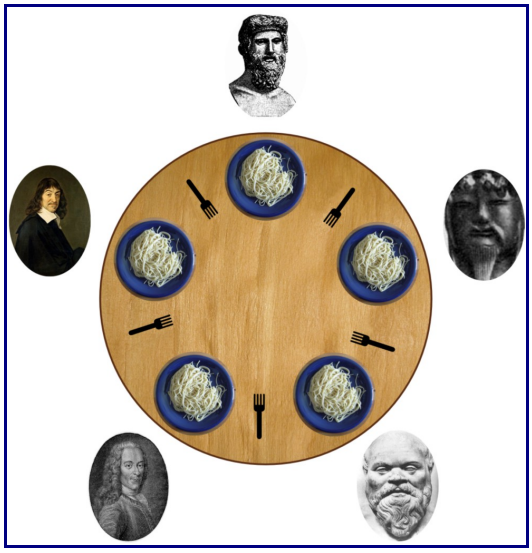
3: "That means I'm left doing the taxes... [you are/we are/I am] such [a] bastard[s]."

One way to think about this is to consider how much information about what you've done in any given circumstance is necessary to make your actions *consistent* with what you're doing in another circumstance. We all have memory lapses, and similar moments in which we can't quite recall every relevant detail of something we've done. What's actively available in short and medium term episodic memory, not to mention which cognitive subsystems dedicated to performing specific tasks are booted up and operating optimally, varies quite significantly over time. When I'm seriously hung over and can't quite remember the promises I've made the night before, the best strategy to ensure my present actions cohere with my past ones is to ask someone else and otherwise do as little as humanly possible. As we saw in the last section, consistent behaviour over time, at least as far as one's ongoing rights and responsibilities are concerned, is the normative substance from which selfhood is spun. What we're doing here is moving from thinking about how one tracks the transmission of *normative statuses* between candidate selves that are temporally distant (phases) to thinking about how this works between candidate selves that are spatially distant (forks). The question is thus, how much communication between concurrent forks is needed to update their *models* of their personal rights and responsibilities, if their overall behaviour is to remain consistent?

As we said in the last section, capacities to keep track of rights and responsibilities presuppose other representational capacities, and this means that updating models of normative statuses might require updating the associated representations. If one fork has spent the afternoon opening a bank account and securing a small business loan, they can communicate the new *entitlements* this grants to the fork out buying supplies and the fork visiting potential premises, but they'll probably have to include the details of the passwords or other systems of authentication by means of which these permissions can be used (i.e., so that money can be spent). Similarly, they each have to communicate to every other fork the financial *commitments* they've undertaken in order to prevent incompatibilities between the choices they've made in their respective situations (i.e., so that they don't spend more money than they can afford), but they'll also have to communicate what goods and services have thereby been purchased if they're to ensure they purchase every item they need *exactly* once (i.e., so they can realise their joint enterprise).

These kinds of co-ordination problems turn up even in much more intimate settings, such as when my five forks sit around a table in a restaurant trying to determine what to order, under certain constraints. *Obviously*, I wouldn't order the same thing for every fork to eat (with their respective forks). I'm *also* the kind of person who would split myself into five forks just to try every option I'd like to eat on a restaurant menu. There are so many places I want to eat, and so little time to systematically experience everything they have to offer. But in this case, there needs to be some *procedure* for determining both who orders what (does each pick one at random and then break ties with rock paper scissors?), and who makes the orders and in what order they do so (do we go by some numbering, 1-5, or does 1 order for everyone?), lest their actions conflict and I end up ordering and eating nothing. It's possible that each fork could pick the same menu item, and then turn to play rock paper scissors with the fork to their left, producing a five-fold symmetry in which it's unclear which game resolves first. If we try to resolve them all at once we can create cycles that

don't resolve: such as when three out of five forks are tied on an option and one picks rock, one picks paper, and one picks scissors. Even worse, if the choice between rock, paper, and scissors is pseudo-random, they could all throw the same sign over and over again, even if they changed sign from turn to turn. In short, we have something of a dining philosopher's problem on our hands.



Just in case it hasn't sunk in fully, I'm trying to emphasise the need to talk about computational concurrency when thinking about forking. Doing so reveals problems that are forced by distribution in space, but which can sometimes turn up even when forks are in close proximity to one another. The main difference between distance and proximity is how much auxiliary information about each fork's local environment needs to be transmitted to the others in order to make sense of the relevant normative status. Nevertheless, these problems give us some theoretical purchase on the nature of the consistency checks involved in *syncing* spatially distant forks that share a common self, even if questions about precisely how well these need to work in practice must be bracketed for the same reasons as those concerning temporally distant phases of the same person. One strategy for achieving such consistency, which I'll continue to call communion, is to try and maintain something resembling a *global state* of the system as a whole, by creating a *control centre* that monitors and modulates the actions of every other element. I've seen people suggest that this requires some sort of quantum entanglement between brains, but that's completely preposterous, insofar as the internal connections between cognitive subsystems within our brains need no such superfluous weirdness. In practice it just means having a primary fork (or *root*) acting as a control system for the rest.

However, this isn't the only way to do things, and might generally be an undesirable way to go about them. The whole point of studying concurrent computation is to design systems composed from a collection of processes with their own *local state*, communicating with one another asynchronously, which are nonetheless guaranteed to behave in certain well defined ways (e.g., avoiding deadlock). It's possible to build decentralised control systems out of such message passing (e.g., using coroutines). This means that it's quite possible to imagine a decentralised network of concurrent communicating forks displaying *unitary* executive function, i.e., having a single *will*. Moreover, it's possible to imagine them displaying various forms of executive dysfunction without thereby ceasing to be a single person. A singular person can easily

be in two minds about a given choice, and if those two minds just happen to be in separate bodies it doesn't necessarily mean they're now two selves. All that's required is some strategy for resolving such internal conflicts, and returning the overall pattern of action to something resembling consistency. Just because one of my forks takes on the role of devil's advocate in relation to the plan of action proposed by another, doesn't mean their interaction can't be guaranteed to resolve in a non-catastrophic manner. Most of the arguments I usually get into with myself do.

That being said, it's worth considering some hypothetical catastrophes: What happens if four out of five of my forks are suddenly destroyed? If they're properly decentralised it would seem that any one of them is capable of continuing on its own. I would lose whatever memories (local state) the others had not yet synced up, but this doesn't seem all that different from getting black out drunk and forgetting what I did last night. I've simply lost some versions of myself running in *parallel*, rather than in *sequence*. Nevertheless, this does raise the question of whether and how my five forks could be merged back into one. That's a thorny problem for neurocomputational version control. I obviously can't iron out the details here, but I think it's fair to say that, if we recognise that a whole fork can be accidentally lost without much consequence, it's fine for some aspects of each fork to be deliberately cut in the name of preserving others. In the limit, I might simply choose not to retain any aspect of the fork doing the washing up, as there's nothing to be gained from such mundane experiences. The important thing to remember is that it doesn't have to be any one fork making these *editorial choices*. They can be made by the system considered as a whole. Forks could bid on which bits of their connectome get priority when resolving incompatibilities, they could rely on an algorithm that makes the choice for them, or something in between. It really doesn't matter. They're all me.

The capacity of any one fork to survive independently seems sufficient to justify the claim that we're dealing with multiple minds sharing a single self, but it's worth entertaining another catastrophe: What if a fork gets isolated from the others for an extended period of time? What if it only thinks the other four have been destroyed, and so evolves independently of them for an extended period of time? What if when they finally find one another, years later, there's too much divergence to integrate? There are really two different issues here, but differentiating them is tricky. On the one hand, the divergence between the wayward fork and the other branch might make integration technically *impossible*. As far as we can tell, Human memory is less about isolated episodic storage than holistic network topology, and the neurological changes generated by divergent experiences might be essentially incompatible. On the other hand, the divergence might make integration existentially *undesirable*. One branch or the other might object to the way in which its counterpart has evolved, in such a way that it considers itself a different person not simply in *practice*, but in *principle*.

Allow me to clarify the trickiness here. So far, I've argued that there can be multiple minds sharing a single self, yet precisely what makes these minds multiple is their capacity to operate independently of one another. One could object to this position that *de facto* independence of minds just is *de jure* independence of selves, and that the forks I've been discussing *really are* five distinct persons who mistakenly believe

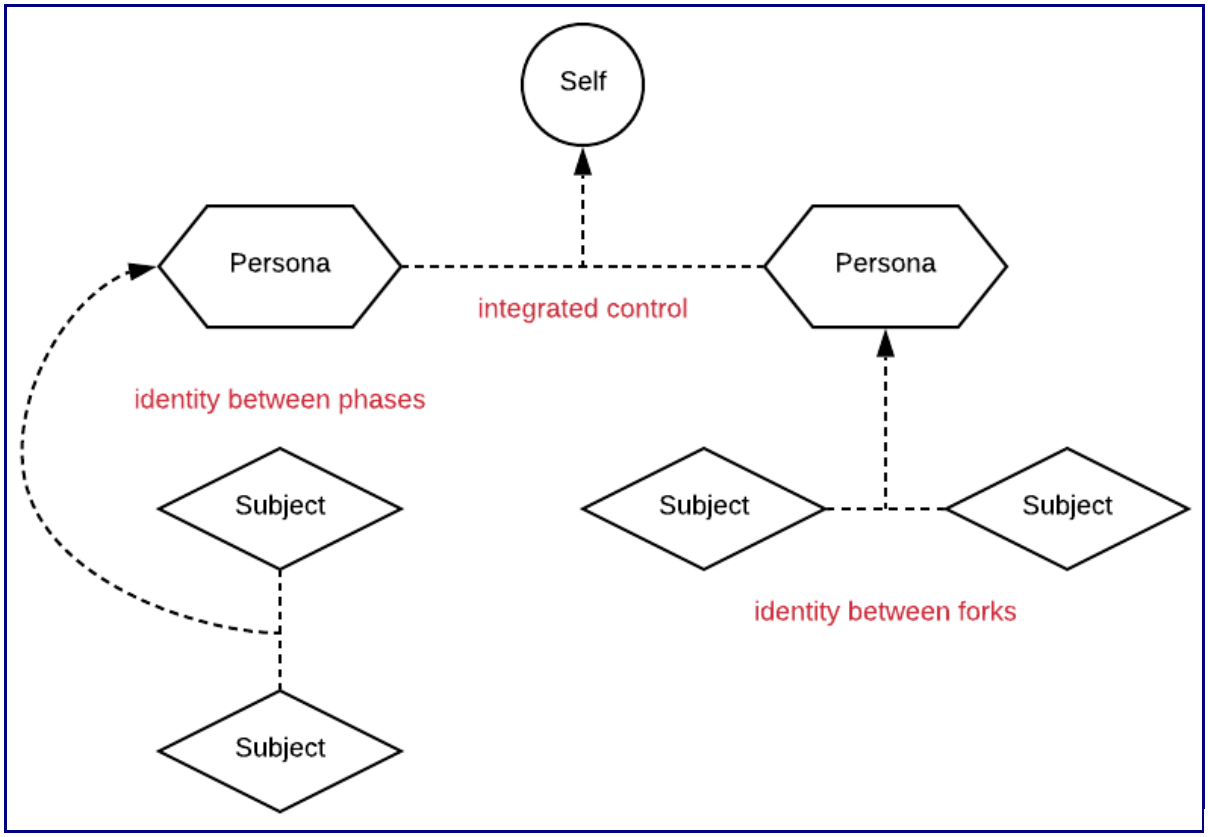
they're identical. In this case the wayward fork has simply discovered something that was true all along. Alternatively, one could claim that the only truth of the matter regarding whether the wayward fork really is distinct is what they say about the matter. If they stipulate that they're distinct, then they are, and that's all there is to it. The former tries to collapse the difference between self and mind in precisely the same way as those who wish to preserve their intuitions about continuity and uniqueness, while the latter obviates questions of identity in precisely the manner that leads to the postmodern dissolution of personhood. To be clear, as in most cases, I'm generally inclined to agree with the latter position in practice, but not in principle, because it makes its case by denying that there is any matter of principle here. This puts us back in the same position we were with plurality.

Consider one more variation then: What if the wayward fork is more or less non-functional? Say it's been badly damaged either by a sudden accident or by some gradual decline in isolation, in such a way that it simply refuses to believe that its fellows have found it. No matter what they say or do, it's convinced that they're imposters. It has somehow become unable to recognise them for *who* they are, which is to say, for who it *itself* is. This is similar to what happens in the Capgras delusion (in which one misrecognises others as imposters), which is itself linked to the Cotard delusion (in which one misrecognises oneself as an imposter: as already dead or otherwise unreal). In this case there's reason to think that the main branch should be allowed to forcibly resync and/or merge with the fork, especially if this misrecognition is merely a symptom of a wider dysfunction in the capacities that enable a mind to maintain a coherent self. This is an *incredibly* delicate case, not least because it has parallels with the forcible re-integration of personality fragments into a unitary self, which returns us to the murky ontogenetic interval between a singular person and an organised system of selves, only without the shared body and mind. Once more, I don't want to adjudicate specific cases, but to trace the principles governing them in outline. This means coming up with a better way of talking about candidate selves (phases/forks) and the way they (mis)recognise one another as identical or distinct, because the terms I've been using thus far beg the very questions we're interested in.

Let's call a 'subject' any operative cognitive process for which there's a meaningful question whether or not it's 'identical' with another process, in the sense of belonging to the same self. Our minds contain a lot of cognitive machinery that isn't always used at the same time, or for the same purpose. These components can be assembled on the fly into an active configuration capable of performing a task. A subject is just some such active configuration. The term is loose enough to allow recursive decomposition: subjects can contain subjects, just as tasks can contain subtasks. It applies both to processes located within the same mind-body and those that aren't, providing a single vocabulary for describing cases of plurality (several distinct subjects in the same mind-body) and cases of forking (several identical subjects in different minds-bodies), but it can equally be used to describe more familiar forms of personhood (several identical subjects in the same mind-body). Just as we can see multiply ensouled bodies (systems) as implementing forms of *collective agency* internal to a single mind (i.e., explicit co-operation), we can equally see singly ensouled bodies as implementing the same sorts of *individual agency* as those that are multiply embodied (i.e., concurrent executive function). What might this look like?

I think it's not uncommon to feel like a different person in different contexts, such as slipping into distinctive patterns of behaviour in different social groups, as if wearing a mask suited to one's company. This is something I felt quite acutely when I was younger, especially at those moments when my social worlds collided, and I suddenly didn't know *who* to be. But this needn't be restricted to the social parameters of different contexts. For some, who they are at work and who they are at home is different, and it doesn't matter if their social graphs overlap. The brutal soldier and the loving partner. The office clown and the stern parent. For others, who they are doesn't so much vary with who they're interacting with, but *how* they're interacting with them. The professional (email). The poet (twitter). The troll (youtube). Social media have, if anything, accelerated the proliferation of these digital masks. They even let us *switch* between masks concurrently, casually flitting back and forth between apps and the persons they make us from one moment to the next. Regardless of the when, where, and why, these behavioural fault lines between contexts can grow into the dissociative disconnects that define plurality, be it through psychological trauma, or deliberate cultivation.

Yet these distinct 'personae' tend to persist across time, sometimes going dormant only to be re-awakened later (e.g., who I am among friends who rarely meet). This dormancy isn't an accident, but a ubiquitous feature of the human condition. If nothing else, in sleep we are no one, and in dreams we may be people we've never been before, and never will be again. However you draw the boundaries between 'subjects', they're best seen as *instances* of what I'm calling 'personae'. The same persona can be operative at different times (phases) and in different places (forks). If we're not careful, this threatens to recreate the problem of personal identity at one remove, as the question of how subjects are integrated into unified selves is swapped for how they are bundled into personae.



For now, the key point is this: we needn't see the self as a kernel of personhood that hides *behind* our many masks, as it could as easily be a structure composed by a collection of personae that *yield* control to one another in a way that produces coherent behaviour. To frame this in terms of self-recognition: personae let us recognise ourselves as agents by *making sense* of our actions in some region of our world, while selves stitch these personae together into a single narrative framework that *covers* the world as a whole. Personae are *local*, while selves are *global*.

Let's get back to the analogy between concurrent forks and concurrent persona communicating in the same mind. We don't want to make too many assumptions about the extent to which human cognition is concurrent, but I think we've earned some speculative leeway here. Let's spend it on another variation of the above example: What if there *really is* an imposter, but it's the wayward fork themselves? What if this fork was kidnapped and altered by a nefarious third party, with the aim of infiltrating my mind(s)? What we have here isn't simply a new personae, but a *false* one. If it does get reconnected to the group, then a bad actor will have gained backdoor access to my inner self, and be able to influence my thoughts and actions without my knowledge. This reframes what's going when subjects recognise one another as identical/distinct: it positions self-recognition as *authentication*.

Authentication is ubiquitous in contemporary life. The *prosthetic personae* we adopt online are all password protected, and usually linked together in a complex hierarchy of access: you log in to some apps via Facebook, which itself is linked to your email address, which by now has multi-factor authentication tying it to your phone, which in turn is tied to your physical address, which in-the-last-instance is pinned to you *qua* embodied personage. These systems of authentication are usually seen as tracking something real that underpins them: who you *really* are, your *authentic* self.

But the last scenario suggests this relation can be inverted: that the self can be *constituted* by a system of authentication, rather than merely identified by it. The concurrent communication between forks needed to sustain a coherent pattern of overall behaviour has to be secure. You don't even have to create subtly altered trojan forks to disrupt my (distributed) self, all you need to do is to spoof my (internal) comms. Yet such spoofing isn't necessarily as simple as cracking a password. Though authentication can incorporate tests for unique tokens (e.g., private keys, personal memories, physical bodies), it can equally include checks for behavioural consistency: detecting communications and/or actions that are *out of character*. These could be *static* checks determined by some overarching model of my character, but they could equally involve *dynamic* checks for consistency between forks. These are precisely those checks needed to maintain a coherent identity.

This brings us back to personae. I think most people have experienced seemingly *alien* impulses or *intrusive* thoughts at one time or another. Sometimes we act on or express them without thinking, but part of developing a coherent personality is learning how to filter them, to classify them as 'not me' and voluntarily suppress them. Not all such impulses/thoughts are so radically uncharacteristic. Sometimes they're simply inappropriate to the context we're in. We can feel which facet of ourselves has generated

them and nudge it back into its proverbial box. This is what it means to cultivate the boundaries between different persona. Not simply learning how to compartmentalise those capacities, tendencies, and reflexes that fit us into one environment, but learning how to switch between them smoothly as those environments overlap, keeping our behaviour consistent along the edges of our *patchwork lifeworld*. In the end, it doesn't matter whether we've got some centralised, overarching picture of our character as long as each persona has enough of a picture of its own and those proximal to it that they add up to a *patchwork personality*. The difference is negligible as long as the handoffs are smooth, with local authentication yielding immediate control from context to context, in a way that guarantees global consistency.

The point where these personae become persons in their own right, and this switching becomes a distinct action they perform, rather than the reflex of an overarching self, remains a tricky question. But I think the zones of swarming subjects betwixt zero, one, and many are maybe a little less murky now. However, we've still got a mereological mystery to solve, and that means finally confronting the obvious question directly: **Just what *is* a self, exactly?**

6. Authentic Egos

This is far from the first piece I've written on the question of selfhood. It's a central theme of the talks 'Autonomy and Automation', 'Beyond Survival', and the more recent 'AI and the Artifice of Self', a key problem raised in 'The Reformatting of Homo Sapiens', and even the subject of some neuropunk speculations in the more personal 'Transcendental Blues'. You'll find aspects of the ideas and arguments laid out above anticipated, elaborated, and contextualised in those pieces, but I won't even pretend to synthesise them all here. Instead, I'll look to the very first paper I wrote on the topic, which was inspired by Ray Brassier's attempt to integrate Wilfrid Sellars' Kantian account of rational subjectivity and Thomas Metzinger's account of phenomenal selfhood. Here's the infamous opening paragraph of Metzinger's *Being No One*:

> This is a book about consciousness, the phenomenal self, and the first-person perspective. Its main thesis is that no such things as selves exist in the world: Nobody ever *was* or *had* a self. All that ever existed were conscious self-models that could not be recognised *as* models. The phenomenal self is not a thing, but a process—and the subjective experience of *being someone* emerges if a conscious information processing system operates under a transparent self-model. You are such a system right now, as you read these sentences. Because you cannot recognize your self-model *as* a model, it is transparent: you look right through it. You don't see it. But you see *with* it. In other, more metaphorical, words, the central claim of this book is that as you read these lines you constantly *confuse* yourself with the content of the self-model currently activated in your brain.

I think there's a great deal of merit in Metzinger's functionalist account of transparent self-models—I'll adopt and explain various features of it shortly—but over the years I've come to find fault in the way he frames the subject matter of his theory. It's not just that he presents his replacement of selves-qua-*substances* with selves-qua-*processes* as denying the existence of selves *as such*, but that this is done in a way that exploits the referential ambiguity of the term 'self' in order to then stipulate what it should refer to. Your *self* does not exist, because you *yourself* are really a certain sort of causal system.

Let's try and get a handle on this ambiguity. First, it's important to see that, along with the pronouns 'I' and 'you', the term 'self' can be used in two ways: it can refer *broadly* to every aspect/component of our physical incarnation, and it can refer *narrowly* to that in virtue of which these aspects belong to the same thing, or that through which these components are unified. This is classically understood as an opposition between *appearance* and *essence*, where the former is contingent and the latter is necessary. This is how Descartes was able to leverage it to posit a distinction between extended and thinking substances, because he could hypothetically strip away every aspect/component of the one while leaving the other untouched. However, this isn't the only way to interpret the broad/narrow relation. The various forms of folk-psychology mentioned earlier, followed by the various forms of psychoanalysis and empirical psychology they beget, begin to conceive it in *functional* and then *representational* terms. The self becomes one component amongst others, whose role is to *regulate* the interactions through which the rest compose, which then requires that it *simulate* them. This self-image then gets analysed into its own components, giving us the body schema, the attention schema, and the ego-ideal, amongst other notions.

Metzinger's view is that our tendency to see ourselves as having an essential core that can be distinguished from its complete incarnation is a result of the representational structure of our minds. He takes consciousness to be a simulation of our environment which cannot be experienced *as* a simulation (a *transparent* phenomenal world-model). We can become aware that our experience is misrepresenting locally (e.g., optical illusions), but never globally, because we have no direct access to the underlying machinery generating it. He takes self-consciousness to be a *partition* within this simulation that separates us from everything else in a similar fashion (a transparent phenomenal self-model). This model within a model has at least four distinct functions: i) it must delimit what does and doesn't belong to our body (ownership); ii) it must situate our perspective within the environment (location); (iii) it must distinguish *actions* that we perform from *events* that simply happen (agency); and iv) it must extend the framework of bodily action to encompass mental actions, such as concentration and imagination (attentional agency). These correspond roughly to the body-schema ((i) and (iii)) and attention-schema ((ii) and (iv)) mentioned above. The question is, what does it mean for us to confuse ourselves with the *content* of these representations (an 'ego'), if what they are representing is precisely the causal system that Metzinger says we are?

Metzinger's aim is to explain the source of those intractable introspective intuitions about selfhood that motivate the theories we discussed earlier: continuity of consciousness and personal uniqueness. But it's important to distinguish between intuition and theory here. Though the transparent simulations Metzinger is describing constitute the *immediate* core of our representation of the world, there are more *mediated* forms of representation founded upon and integrated with it. Language is the medium through which such representations get articulated, assessed, and corrected, but this doesn't mean they are intrinsically linguistic. A day trader has to have a fairly complex internal model of the stock market in order to go about their business, and this model may exploit aspects of sensory imagination to function (e.g., *visualising* rising prices graphically, or *feeling* compatible strategies as harmony), but this *mediated immediacy* can be more easily brought into question by heterogeneous forms of information (e.g., *testimony* from colleagues or

mathematical analysis from experts). Theories of personal identity, Metzinger's included, aim to integrate our immediate (first person) intuitions about ourselves into a mediated (third person) model of the world that extends beyond our phenomenal horizon. Metzinger's theory invites us to see some aspects of our sense of self as a low-dimensional simulation of the complex physical system that is the human brain-body system, which on that basis can easily misrepresent it (e.g., the felt presence of a phantom limb), and other aspects as artefacts of this simulation that *cannot even misrepresent* (e.g., the felt certainty that I'm the same person as the teenager who sat his exams, but not the drunk that did things I can't remember and would never do).

For Metzinger, the only facts about identity that could be misrepresented here are whether we're the same functional system, but these facts are essentially indifferent to the content of that self-model. I could be kidnapped and brainwashed in such a way that this content was radically altered, changing both my behaviour and the way I interpret this behaviour, and there simply would be no fact of the matter as to whether I'm still the same person beyond being the same organism. The brainwashing technicians could make the resulting creature interpret itself as continuous or discontinuous with its previous incarnation, without significant consequence. This would seem to suggest that Metzinger is on the side of those bioconservatives who would insist, if they even countenanced their possibility, that my hypothetical concurrent forks are *really* distinct persons, insofar as they are distinct organisms, but that's not quite true. Metzinger sees his work as an extension of the Enlightenment project, both in the sense of disenchanting nature of any such normative valence, and in the sense of taking responsibility for (re)making ourselves as autonomous individuals. He is happy, and even eager, to talk about the plethora of *unnatural* minds that cognitive engineering might allow us to create and/or become.

Here's how these issues are framed in the closing paragraph of *Being No One*, directly in response to the opening one:

> Do you recall how, in the first paragraph of the first chapter, I claimed that as you read these lines you constantly *confuse* yourself with the content of the self-model currently activated in your brain? We now know that this was only an introductory metaphor, because we can now see that this metaphor, if taken too literally, contains a logical mistake: There is no one *whose* illusion the conscious self could be, no one *who* is confusing herself with anything. As soon as the basic point has been grasped—the point that the phenomenal self as such is not an epistemically justified form of mental content and that the phenomenal characteristic of self*hood* involved results from the transparency of the system model—a new dimension opens. At least in principle, one can wake up from one's biological history. One can grow up, define one's goals, and become autonomous.

This acknowledges the problematic ambiguity in the opening, then simply displaces it. There is *no one* to whom the illusion is appearing, but understanding this allows *someone* to wake up and shrug off the constraints that our evolutionary history has imposed upon us. To be fair to Metzinger, he does suggest that this awakening consists in re-engineering ourselves to overcome these constraints, such that the individuals (and maybe distributed systems) that awaken are not quite those that were asleep; but the *normative* questions he raises about how we should go about this, and what it means to do so in a way that

is an expression of *autonomy*, are fundamentally disconnected from his theory of selfhood as such. He ends up closer to our postmodern radicals: **One cannot further the cause of self-determination by dissolving the notion of self.** He's managed to prioritise metaphysics of personal identity over the relevant normative issues even as he eliminates it.

Despite this, Metzinger's account of the self-model is extremely powerful. What's required is a way to treat it, or a more mediated self-understanding that extends it, as representing *something*, even if this is neither the body qua physical system nor the soul qua metaphysical substance. There are two basic insights required to make this work. On the one hand, the self-model must simulate not simply the way the body and mind *really are*, but the way they *ideally should be*. It should incorporate not just the body schema and attention schema, but also something like the ego-ideal. If nothing else, who you try to be is as important to understanding who you are an agent as what you try to achieve is important to understanding your actions. On the other, there needs to be some sense in which the self-model can *misrepresent* not just its physical parameters but which person it incarnates. Our ability to define who we are as autonomous agents must operate under constraints. If nothing else, you can't simply define yourself as having an immortal soul, no matter how deeply invested you are in the idea that you're the reincarnation of some famous historical figure. These insights are tied together by the open-ended oscillation between choosing who we want to be, and learning that this is impossible on the terms we've set ourselves. Being who we are is a *commitment* whose content is progressively elaborated and revised, as we try and fail to write our life stories on the turbulent surface of the world.

Let me sketch a final variant of the fork example to illustrate my point, and reconcile it with Metzinger's: If we picked a random person, split them into five qualitatively identical forks, and put them in a room together, would we have one person or five? Maybe it could go either way. What if we deliberately picked a person who is personally invested in their introspective intuitions about uniqueness? Each copy of this person would believe themselves distinct from the others, and it's reasonable to think that this is *sufficient* to make them so distinct. Their self-image does not permit simultaneous multiple instantiation, even if this creates tricky questions about how the rights and responsibilities of their original get inherited and/or divided between them. Something similar applies to the examples of cybernetic enhancement and mind uploading considered earlier. These might retain uniqueness, but they can undermine the conception of continuity built into a person's vision of themselves, forcing them to redefine who they are in a way that we should respect.

Conversely then, is me believing that, if I were forked in this way that I'd still be one person, sufficient to make it so? This is where Metzinger's approach challenges the permissive stance of postmodern radicalism. The immediate intuitions generated by the transparent self-model are *very strong indeed*. It's one thing to have a theoretically mediated belief that you can be split and recombined into concurrent forks on the fly, and still be the same person, and another to actually act in a way consistent with this belief when you're put in that situation. It might turn out to be impossible to achieve the sorts of consensus between semi-autonomous forks needed to make the arrangement work without either significant training or direct

modification of one's self-model. The requisite forms of recognition, communication, and authentication might need to be hard wired in order to guarantee unity. If nothing else, if my latent fear that reintegration will kill me prevents my forks from going through with it, then there's a nascent crack in my ego-ideal that prevents its realisation. After all, ought-implies-can: if you can't (reliably) hold yourself to certain standards then you shouldn't. No commitment without (some) capacity.

These hypothetical problems are less far away from real ones than you might think. The consistency required between forks to execute a plan of action concurrently is very similar to that required between phases to execute a plan sequentially. Both cases require *self-discipline*: delayed gratification in the latter case (e.g., working hard to pay for a future vacation) might be deferred gratification in the former (e.g., one fork working hard to pay for another to travel the world). I'm sure I'm not alone in occasionally resorting to adversarial relationships in order to tie myself together over time: laying traps for my future self that force me to complete tasks I would otherwise avoid, then cursing my past self before condemning my future self to a similar fate. This is but one strategy for coping with internal constraints on our ability to fulfil our commitments. Another is cultivating and configuring the relationships between distinct personae.

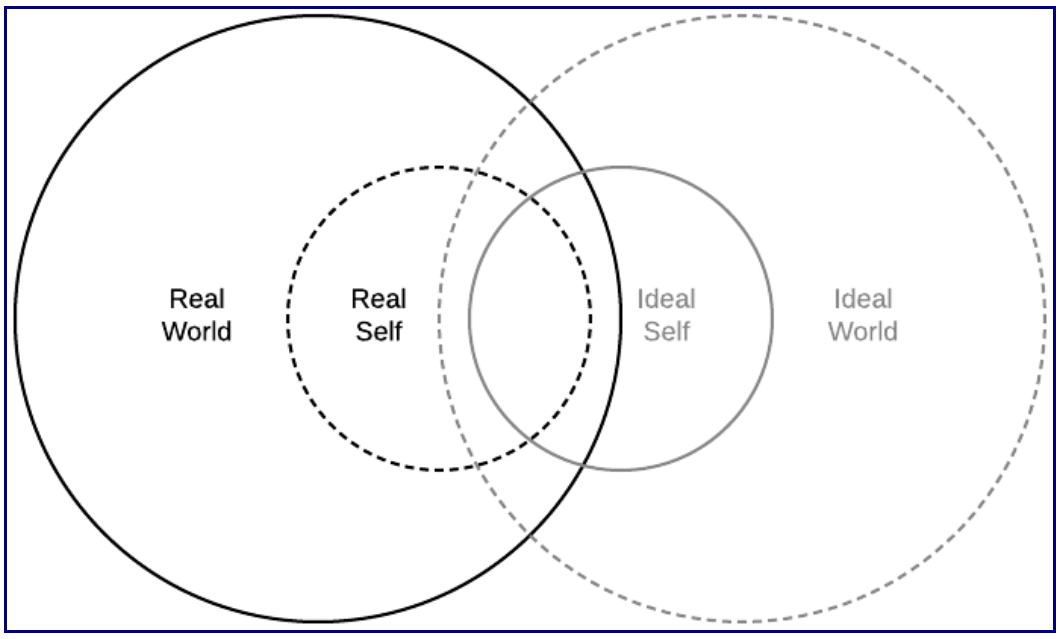
Consider the situation that occasioned this post: I suffer from significant variation in cognitive capacity and executive function. Sometimes I simply cannot be the person I want to be, that I otherwise could be, in a given context. One persona after another decoheres and crashes, as they fail to approximate their ideals, and other, less ambitious ones must take over. At my worst I'm left in the equivalent of safe mode, periodically attempting to restart any version of me that can be something more. There's a non-trivial danger of dissociation here, as these junctures between personae grow into fractures in my ego-ideal. To be more cliché, bipolarity threatens to split me down the middle, leaving two selves that cannot agree on how best to live the life they share. But this division of psychic labour can be codified, refined, and *willed* as such. A plan can be a strategy can be a narrative. There's nothing to say that self-*legislation* can't also be self-*construction*. That's when we call it a constitution.

7. Conclusion: Artificial Souls

So, just how much of one's body can one rationally hate without the belief that one does not hate oneself becoming delusional? This question has taken us on a long and meandering path, but I think we've finally collected all the resources needed to answer it. All that's left to do is assemble them. The core idea is this: **Every rationally autonomous agent has some authority over the identity conditions governing their own persistence**. These conditions are not created from whole cloth, without constraint, and not every decision regarding how to identify is *consistent* let alone *good*. The process of self-determination is neither without friction nor failure, but it permits us to evolve in strange and mutually incompatible ways. However, in order to make sense of this core idea, I have to articulate the key distinctions involved in describing any causal system that might incarnate a person: *self/world*, *real/ideal, implicit/explicit*, and *means/ends*.

Let's begin by combining Metzinger's conception of the relation between the phenomenal self-model and the phenomenal world-model with the account of rational agency as constituted by a representation of the

world as it really is and as it should ideally be sketched earlier. This gives us a fourfold distinction between real world, real self, ideal world, and ideal self. There are three things to bear in mind here. First, that our self-model is a partition within the world-model, insofar as the self is part of the world. This goes for the ideal self/world as much as the real one. Second, that these models can extend beyond the *immediate* world of animal awareness and into the *mediated* world of rational cognition. Reason bootstraps the scientific image of man in the world from out of the manifest one. And third, that these representations/simulations/models need be neither *centralised* in structure or *homogeneous* in format, as long as they can be dynamically integrated in practice. Our cognitive subsystems needn't already be *actually* consistent, as long as we're always striving to make them *rationally* cohere. To diagram this picture of the mind:



The next thing to do is to explain the overlap between these representational structures. There are really two issues here. On the one hand, there's always some *explicit* overlap between the real and the ideal. We can't wish to change everything about the world, or even everything about ourselves, without a complete and total disconnect akin to death. To be the same person living in the same world, only better, requires only that we simulate the *divergence* between them (prediction), and that this divergence *drives* action (prescription). On the other, there's always a greater *implicit* overlap between the systems generating these simulations than we are capable of making explicit. Even our mediated representations remain transparent in the limit: we're only capable of discovering and acting on misrepresentations locally, never globally, for this always requires using some representational capacities to correct others. Whether errors in these representations are mere misunderstandings or wishful thinking is thus somewhat blurry, but the world-model and self-model blur in opposing ways.

The abstraction implicit in the sorts of low-dimensional representations that compose them is typically seen as *instrumental* from the perspective of our world models, but as *teleological* from the perspective of our self-models. Our picture of the world is implicitly *simplified*, but our picture of ourselves is implicitly

idealised. This is what it means to say that who we *are* is an essentially normative matter, even if we can discover truths about our real capacities that force us to revise the commitments implicit in the way we see ourselves. To put this in more classical terms, the distinction between the *body* and the *soul* is really a distinction between *means* and *end*. Kant's idea that rational beings are ends-in-themselves implies Hegel's idea that the body is the original site of property (or sovereignty), qua means to this end. If we desire to remain as we are, this needn't imply we don't wish to undergo transformations of a kind that our self-conception doesn't and perhaps can't comprehend: How would a medieval peasant feel about cellular replacement? How would they feel about blood transfusions? No matter how horrified they'd be about the former, they wouldn't be able to tell a coherent story about themselves if they didn't countenance it. The latter could go either way. Learning more about our body is not necessarily learning more about *who* we are (essence/end), even if it means learning about the constraints governing *how* we are (appearance/means). One means is as good as another, unless it's something we're attached to, at which point we can incorporate it into the end if we wish.

These ideas could fleshed out more. There are things to be said about mutual recognition and the social constitution of selfhood, the ludic structure of narrative identity, the significance of death/survival, and a variety of other things. But I have to stop elaborating somewhere. How should we answer the question then?

The easy version of the question is whether it would possible to replace every component part of my body and still satisfy my conception of who I am, or at least, fall short of it no more than I usually do. *Prima facie*, I've established the feasibility of the examples of forking/merging, mind-uploading, and cybernetic transplantation discussed above, insofar as I'm personally quite happy to treat the matter out of which I'm made as an (in)convenient means to an end. But the harder version concerns the feasibility of substantive changes in the behaviour of my body (and mind). I might not *know* who I am well enough to license such potentially irreversible changes. There might be some bodily inconveniences whose essential character I don't fully appreciate before the transformation, and cannot appreciate afterwards. Shutting off my pain receptors, flensing my amygdala, or modifying the parameters of the way my body fits into its environment (and my mind embeds/extends into it) might break my patterns of behaviour in catastrophic ways while leaving me unable to comprehend what I've broken, especially if I do them all at once. These are experiments in self-realisation that can fail.

My pithy response to these worries is that anything which can't fail isn't an *experiment* worth the name, but that there's no reason we have to perform them all at once. I have personally taken psychoactive medication that modifies the way my neurology works in ways I and even those who prescribed it to me don't fully understand, and given time to test it I've been able to determine that I don't like how it changes me, how it unbalances the compact between my personae. It's likely I'll do this again in the future, perhaps with different results. Who knows? The difference between this and more drastic changes to the neurocomputational underpinnings of my mind are more matters of degree than kind. This isn't to say there isn't some complex boundary the crossing of which would leave something other than Pete in my wake,

only that this most deeply personal of sorites problems can't be solved without a non-trivial combination of experimental self-discovery and soul-searching decision. But this in turn is not different in kind from the ordinary logic of self-construction.

The most peculiar thing about self-knowledge is that, despite our authority to *decide* who we *are*, it's still easier to say for sure who we *aren't*. To be clear, I'm not just talking about identity as a *relation* here (i.e., which other subjects I'm identical to), but about identity as *property* or collection of properties that between them somehow determine this relation (e.g., whether I'm a man, a philosopher, a socialist, etc.). If existence precedes essence, it's *brute*, rather than *bare*. Each one of us is a bundle of drives telling itself a story, and we might not know which narrative path will make them cohere, but we can be certain that whatever else *this* one doesn't work. If we're lucky, we eventually find one that does. The maxim of such personal fallibilism remains Pindar's injunction: "Become who you are."

Crucially, I don't mean to deny that we can find facts about our biology, neurology, or psychology that place fundamental constraints on who we are and can be. To have written such a long piece spurred by my own experience of mental and physical illness only to turn round and reject this would be not a little hypocritical. As a good Spinozist, I encourage everyone to study their body as a causal system to the fullest extent of their ability. It's important to know what a body can't do. But it's important to resist the idea that this tells us anything positive about who we *must* be, to convert our natural history into an authentic origin. Such authenticity is nothing but the anticipation of future nostalgia: *I will have always already been what I am now.* We'd do better to examine the ways in which the systems of authentication out of which we spin our selves enable an openness to becoming otherwise. But that's a project for another time.

Allow me to close by saying something more about the philosophies of embodiment I've been criticising. I don't want to claim that those who strongly identify with their bodies are delusional, because their choices in this matter are entirely their own. However, I will claim that those who think that there are no other options available to them are labouring under a set of constraints that can only be described as *ideological*. I've discussed the various strands of the embodiment paradigm elsewhere, but I think what I've done here is to sketch a peculiar convergence that's the thin end of an ideological wedge. I don't mean those forms of bioconservatism I've mentioned at various points, but the odd combination of postmodern radicalism and new materialism that's surprisingly common in certain quarters in the humanities. My critique of these ideas is hardly new, but I think the position I've sketched here provides a precise way of describing a tendency I've elsewhere called the *sacralisation of the body*.

When the explanatory, normative, and metaphysical critiques of Cartesianism, Platonism, and Abrahamic conceptions of mind and world converge, in a way that simultaneously rejects any substantial self as locus/origin of value and the vector of disenchantment that would strip nature of such value, the distinction between means and end here articulated collapses. The precise form this collapse takes varies, from autopoetic vitalism to ecological animism, but the common thread is that **the body becomes the soul**. This completes the normative circuit from radical anti-naturalism to reactionary neo-naturalism I've been tracing

here: the problematic features of supernatural mental substance they began by critiquing have been transposed onto the natural body, producing new and reproducing old theological impulses.

I'm increasingly of the opinion that the proper response to this is retake the language of souls: not as *supernatural sparks* that determine our fates, but as *artificial designs* that elaborate non-deterministic destinies. The same thing is at stake in self-care, childrearing, and the engineering of autonomous artificial general intelligences: the crafting of souls capable of driving themselves into ever more interesting regions of the space of possible minds. Not so much programs as programmes, to exploit a quirk of British spelling.

Here ends my own little excursion in the theory and practice of soulcraft, stitching my ideas, inclinations, and this old prosthetic persona back together in a new and more durable form. It is my great pleasure to announce that normal service will be resuming shortly.

BEGINNING OF PHASE 3.