PhiMiSci
Philosophy and the Mind Sciences

Look who's talking!

Varieties of ego-dissolution without paradox

Sascha Benjamin Fink[a] (sfink@ovgu.de)

Abstract
How to model non-egoic experiences – mental events with phenomenal aspects that lack a felt self – has become an interesting research question. The main source of evidence for the existence of such non-egoic experiences are self-ascriptions of non-egoic experiences. In these, a person says about herself that she underwent an episode where she was conscious but lacked a feeling of self. Some interpret these as accurate reports, but this is questionable. Thomas Metzinger (2004, p. 566, 2018), Rocco Gennaro (2008), and Charles Foster (2016, p. 6) have hinted at the self-defeating nature of such statements if we take them to be genuine reports: Apparently, the reporter (a) explicitly denies her existence during the selfless experience, but (b) implicitly affirms her existence as a witness to that selfless experience in order to give a first-person report about it. So the content of such a report conflicts with the pragmatics of reporting. If all self-ascriptions of non-egoic experiences are self-defeating in this way, then they cannot count as evidence for the existence of non-egoic experiences. Here, I map out why such strong conclusions do not directly follow: What look like self-ascriptions of non-egoic experiences may occur for a number of reasons. Only some explanations for such utterances rely on a change in consciousness. Of those that do rely on a change in consciousness, only one (total ego-dissolution) is incoherent. But its alternatives do not lead to contradictions. I argue that the most likely change in phenomenality that leads to self-ascriptions of non-egoic experiences is not one where a felt self disappears, but where it expands.

Keywords
Consciousness · Ego-dissolution · Ego-expansion · Multiple selves · Non-egoic experiences · No-Self · Self-refutation

This article is part of a special issue on "Radical disruptions of self-consciousness", edited by Thomas Metzinger and Raphaël Millière.

We are led to believe that some experiences are *non-egoic*: They feel like something, but lack a felt self; they are clearly phenomenal, but the ego is not part of this experience; no one is in the experience, but it still has qualitative features. If there was a feeling of self in experience before one went into a non-egoic episode, we call

[a]Otto-von-Guericke-University Magdeburg

Fink, S. B. (2020). Look who's talking! Varieties of ego-dissolution without paradox. *Philosophy and the Mind Sciences*, *1*(I), 3. https://doi.org/10.33735/phimisci.2020.I.40
 https://philosophymindscience.org ISSN: 2699-0369

this process *phenomenal ego-dissolution.*[1] Non-egoic episodes are characterised by (i) the absence of a feeling of self, and (ii) a noticeable contrast to a feeling of self experienced previously. Apparently, people reporting selfless experiences did not always lack a feeling of self; instead, it is unusual, something that contrasts with one's common ways of experiencing, something remarkable.

Why should we think that there are such experiences? Some people claim that they have undergone (or are undergoing) an episode during which they themselves experienced in a non-egoic way. These claims come in the form of a reflexive first-person ascription, where a person ascribes *to herself* some mental event – in this case, an episode of phenomenal ego-dissolution, either in the past or in the present.[2] Such *self-ascriptions of (episodes of) non-egoic experience* are what I focus on here. But because this phrase is cumbersome, we can use the abbreviation SANE for them. Strictly speaking, SANE could be merely mental and need not be verbally expressed. But if they never were expressed, they would not encourage others to think that such non-egoic experiences exist. It is the verbal expression of such self-ascriptions of non-egoic experiences that counts as intersubjective evidence for there being non-egoic experiences. This is what I focus on here: utterances which, in their most likely reading, involve or are reasonably taken to express SANE. Such verbal expressions are what I mean when I use the abbreviation SANE. And they are primarily responsible for the SANE paradox I focus on (section 3):[3] That SANE are self-refuting, that their structure entails their own falsity.

This paper presents examples of SANE and their phenomenological profile in section 1, characterises their importance for a research programme on non-egoic experiences in section 2, and analyses the argument for them being self-refuting in section 3, where the underlying assumptions for this claim are made explicit.

Now, I do not think that SANE need to be self-refuting. In fact, we have diagnoses available on several levels – pragmatic (section 4.1), cognitive (section 4.2), and phenomenological (section 5) – to explain why people utter SANE. The SANE paradox only comes about by embracing (i) a connection between the phenomenal *feeling of self* and the underlying *mechanisms facilitating self-reference* and (ii) that there is only one unique phenomenal feeling of self. Each of these two claims can be rejected separately, leading to competing diagnoses about what happens in those phenomenal episodes some SANE refer to. Should we keep all of these diagnoses in our tool box? In section 5, I argue against such a pluralism: We should expect at most one of these diagnoses to be accurate.

In order to defend the feasibility of researching phenomenal ego-dissolutions

[1] Conceptually, such episodes of phenomenal ego-dissolution can be *transient* (if one regains a feeling of self afterwards) or (otherwise) *permanent.* Some Buddhist monks may claim that they achieved permanent ego-dissolution. But most cases appear to be of the transient kind. Therefore, I will restrict myself to these. All theoretical points I make, however, will apply to the permanent kind as well.

[2] That is, she ascribes to herself the property of *experiencing non-egoically* at a time t and a location l of her physical body.

[3] There are, however, cognitive analogues.

Fink, S. B. (2020). Look who's talking! Varieties of ego-dissolution without paradox. *Philosophy and the Mind Sciences, 1*(I), 3. https://doi.org/10.33735/phimisci.2020.I.40

and non-egoic experiences, we must give an account of how SANE can offer evidence for some form of phenomenal ego-dissolution. In order to do so, we need to understand better the argument why SANE are supposed to be self-defeating. Then we can catalogue the ways in which we can reject this argument. These are the goals of this paper.

1 SANE introduced: Self-ascriptions of episodes of non-egoic experiences and phenomenal ego-dissolution

Some people say that they have had episodes in their mental life where they felt like no one at all, where their selves went missing, where there was no feeling of self. These episodes, however, were not of unconsciousness, akin to anaesthesia or deep sleep. Instead, they had phenomenal experiences with felt qualities, but these felt as if they were happening to no person, no self. I call such experiences *non-egoic experiences*, and a process of leading into them a *phenomenal ego-dissolution*, because, intuitively, there was a prior feeling of self that then vanished.

People express their beliefs about having had such experiences in verbalised *self-ascriptions of (episodes of) non-egoic experience*, which I abbreviate as SANE.

SANE can be found in a wide range of circumstances. In psychiatry, ego-dissolution is seen as a symptom of certain psychopathologies (Sass, Pienkos, Nelson, & Medford, 2013; Simeon & Abugel, 2006). Elyn Saks (2007, pp. 12–13) has provided us with one of the most vivid examples in *The Center Cannot Hold*:

> And then something odd happens. My awareness (of myself, of him, of the room, of the physical reality around and beyond us) instantly grows fuzzy. Or wobbly. I think I am dissolving. I feel – my mind feels – like a sand castle with all the sand sliding away in the receding surf. *What's happening to me? This is scary, please let it be over!* I think maybe if I stand very still and quiet, it will stop.
>
> This experience is much harder, and weirder, to describe than extreme fear or terror. Most people know what it is like to be seriously afraid. If they haven't felt it themselves, they've at least seen a movie, or read a book, or talked to a frightened friend – they can at least imagine it. But explaining what I've come to call "disorganization" is a different challenge altogether. Consciousness gradually loses its coherence. One's center gives way. The center cannot hold. The "me" becomes a haze, and the solid center from which one experiences reality breaks up like a bad radio signal. There is no longer a sturdy vantage point from which to look out, take things in, assess what's happening. No core holds things together, providing the lens through which to see the world, to make judgments and comprehend risk. Ran-

Fink, S. B. (2020). Look who's talking! Varieties of ego-dissolution without paradox. *Philosophy and the Mind Sciences*, *1*(I), 3. https://doi.org/10.33735/phimisci.2020.I.40

> dom moments of time follow one another. Sights, sounds, thoughts, and feelings don't go together. No organizing principle takes successive moments in time and puts them together in a coherent way from which sense can be made. And it's all taking place in slow motion.
>
> Of course, my dad didn't notice what had happened, since it was all happening inside me.

Saks calls this experience disorganised, marked primarily by a loss of mereological, spatial, or temporal unity ("reality breaks up", "no core holds things together", "random moments of time follow another"), a loss of determinacy ("my awareness [...] grows fuzzy [...] wobbly"), or a loss of coherence. This is tied in with a reduced form of specificity of the self-boundary: "The 'me' becomes a haze." She appears to see the dissolution of her feeling of self as the cause of the other forms of disorganisation: The fact that a core, a centre, a "sturdy vantage point" is missing leads to the lack of unity in these experiences. Without a feeling of self, the senses do not cohere among each other or with cognition (e.g. judgments and comprehension of risk). Even time seems to lack a steady, appreciable pace or sequence. The feeling of self, it appears, is associated by Saks with diachronic and synchronic phenomenal unity. And, notably: This radical change in experience is not behaviourally noticeable for the people around her.

This experience appears to be distant from our everyday way of experiencing, hardly describable, alien. The difficulty is apparent in a specific reluctance to use indexicals as ways to orientate and locate oneself in time or space: "I feel" – herself being involved in feeling – is replaced by "my mind feels", where she is merely the owner of the mind, herself apparently dissociated from what she owns, like a traveller who is far away from the home she owns. Then again, "my" becomes replaced by "the 'me' ", a thing one *has*, not what one *is*. Elsewhere, all indexicals are avoided, when "one's center" – still intended as referring to an indeterminate someone who has a centre – is replaced by "the center", which lacks any indication of being the centre of a person. This creative sensitivity to the subtleties of language is what makes the passage so captivating. But we should be wary of the possibility that literary quality may have trumped veridicality here.

What supports Saks's description, however, is that other reports converge on similar themes, for example reports of psychedelic experiences (see especially Letheby & Gerrans, 2017; Millière, 2017; see also Deane, 2020, this issue, as well as Letheby, 2020, this issue):

> I realized then that I wasnt [sic] myself, I wasn't anything anymore. I had been broken down into nothingness, into oblivion. I felt as though the darkness was more of a something than I. I'm fucking dead I thought. There is no other explanation. I felt nothingness, like my brain had just been paused and I was still taking in my surroundings unwillingly. I felt comatose.[4]

[4]Erowid report #87426. https://erowid.org/exp/87426.

Fink, S. B. (2020). Look who's talking! Varieties of ego-dissolution without paradox. *Philosophy and the Mind Sciences, 1*(I), 3. https://doi.org/10.33735/phimisci.2020.I.40

During this 5-MeO-DMT-induced experience, Rew (17 years old at the time), felt like he had dissolved into "nothingness", that even immaterial darkness was more existent than he was. No feeling of self seems to be at the core of this experience. Additionally, even though he was perceptive ("I was still taking in my surroundings"), he apparently did not feel like an agent in this process (he did so "unwillingly").

While such experiences appear to be frightening when they happen spontaneously, some people try to attain them deliberately: non-egoic experiences are the goal of certain spiritual practice, a sign of having mastered certain meditative or contemplative techniques (Berkovich-Ohana, Dor-Ziderman, Glicksohn, & Goldstein, 2013; Dor-Ziderman, Berkovich-Ohana, Glicksohn, & Goldstein, 2013; Josipovic, 2010), indicating a high level of enlightenment, awakening, or connection with the divine. What is a symptom of pathology in one context is a pinnacle of spirituality in another.

Reoccurring features of SANE are:

(CEASE) A feeling of self is missing in experience.

(ACCESS) The absence of a feeling of self is cognitively accessible.[5]

Without CEASE, the experience would not be non-egoic. Without ACCESS, we would not have any reports or memories of these experiences. Both are common themes in SANE. If taken at face value, SANE suggest that we can undergo non-egoic episodes in meditation, schizophrenia, or psychedelic highs.

If non-egoic episodes occur in such a wide range of circumstances, are they anything special at all? Intuitively, we might think so, but some philosophers argue that experiences are *mainly* egoless, at least to some degree. For example, in *La transcendance et l'ego*, Jean-Paul Sartre reflects on his problems with Husserl and public transport:

> No one would deny for a moment that the *I* appears in a reflected consciousness. [...] [This reflective consciousness, however,] *modifies* the spontaneous consciousness. Since, in consequence, all the non-reflective memories of unreflected consciousness show me a consciousness *without a me* and since, on the other hand, theoretical considerations concerning consciousness which are based on intuition of essence have constrained us to recognise that the *I* cannot be a part of the internal structure of *Erlebnisse*, we must therefore conclude: there is no *I* on the unreflected level.[6] When I run after a streetcar, when I

[5] Especially: accessible for conceptualisation and conceptual processing, e.g. belief and memory formation, reasoning, deliberation, evaluation, etc. Only conceptualisation will lead to verbal reports about non-egoic experiences.

[6] Obviously, it does not follow that if a modifies b, then whatever holds for a cannot hold for b, because this would exclude mutual modification. Let us read this passage simply as indicating a distinction between "reflective" and "non-reflective" consciousness.

Fink, S. B. (2020). Look who's talking! Varieties of ego-dissolution without paradox. *Philosophy and the Mind Sciences*, *1*(I), 3. https://doi.org/10.33735/phimisci.2020.I.40

> look at the time, when I am absorbed in contemplating a portrait, there is no *I*. There is consciousness *of the streetcar-having-to-be-overtaken*, etc., and non-positional consciousness of consciousness.
> (Sartre, 1960, pp. 48–49)[7]

Sartre argues that there is a basic, unreflected level of living consciousness, the *Erlebnisse*. These are experiences of simple perceiving without apprehending (like being absorbed in looking at a portrait) and of actions without explicit intentions (like overtaking a streetcar). On this level, there is no self in the experience. A felt self is introduced only by reflection. Here, one has to mark oneself as the thinker of thoughts, the initiator of actions, the person inhabiting a body, and one has to dissociate oneself from other thinkers, agents, and bodies. In a weaker reading, Sartre argues that even if it is undeniable that a self exists on a reflected level (where we can remember and report on *Erlebnisse* as being my experiences, think about our lived experiences, and have cognitive access to these experiences), this does not strictly entail that a self is also part of our *un*reflected experiences. Then, non-egoic experiences are a possible kind of phenomenal consciousness outside of self-involving conscious cognition. Owen Flanagan (1992, p. 178)[8] and David Hume (1739, p. I.4.vi)[9] seem to go even further and claim that experiences *standardly* lack a felt self, that there is nothing it feels like to be someone, no self-quale or feeling of self.

But if experiences are *always* non-egoic, we cannot account for a very prominent feature of Sane: There is a clear change from what is felt *before and after* to what the experience is like *during* the non-egoic episode. This is a third recurring feature of Sane:

(Difference) There is a felt difference in the feeling of self between the time before/after and during non-egoic episodes.

Without Difference, a non-egoic experience would not be remarkable, but they clearly seem to be. The moments before and after contrast sharply with those during the episodes of ego-dissolution. We try to capture this difference by questionnaires like the 5D-ASC (Studerus, Gamma, & Vollenweider, 2010) or the EDI

[7] In the original: "Personne ne songe à nier que le Je apparaisse dans une conscience réfléchie. […] la réflexion modifie la conscience spontanée. Puisque donc tous les souvenirs non-réflexifs de conscience irréfléchie me montrent une conscience sans moi […]: il n'y a pas de *Je* sur le plan irréfléchi. Quand je cours après un tramway, quand je regarde l'heure, quand je m'absorbe dans la contemplation d'un portrait, il n'y a pas de Je. Il y a conscience du *tramway-devant-être-rejoint*, etc., et conscience non-positionnelle de la conscience." (Sartre, 1936/1992)

[8] "The illusion is that there are two things: on one side, a self, an ego, an 'I', that organizes experience, originates action, and accounts for our unchanging identity as persons and, on the other side, the stream of experience. If this view is misleading, what is the better view? The better view is that what there is, and all there is, is the stream of experience."

[9] "[…] identity is nothing really belonging to these different perceptions, and uniting them together; but is merely a quality, which we attribute to them, because of the union of their ideas in the imagination, when we reflect upon them."

Fink, S. B. (2020). Look who's talking! Varieties of ego-dissolution without paradox. *Philosophy and the Mind Sciences, 1*(I), 3. https://doi.org/10.33735/phimisci.2020.I.40

(Nour, Evans, Nutt, & Carhart-Harris, 2016), which specifically probe aspects of ego-distortion. It is this felt difference from everyday experiencing that is unaccounted for by philosophers who deny a feeling of self in general. Sartre seems to be providing an account of something other than radical non-egoic experiences because following a street car is hardly as dramatic as the experiences described by Saks and Rew seem to be. A theory that denies the existence of a feeling of self in general or suggests that non-egoic experiences are extremely widespread does not explain the kind of data gathered in the validated questionnaires. Something special appears to be going on during these episodes.

2 SANE and their role in researching non-egoic experiences

We may be tempted to focus directly on the question: What can we learn from SANE about the specific features of non-egoic experience? – and simply proceed with a research programme centred on non-egoic experiences themselves.

But maybe we should start by asking about the epistemology of ego-dissolutions: How can we *know* what happens in such episodes? How can we know about the specifics, the dynamics, the onset and offset of these episodes? How can we determine which phenomenal features and structures are affected? What evidence do we have that there is a difference in experience, not merely in reporting? What grounds do we have to think that phenomenal ego-dissolution exists and is not some kind of artefact?

What could count as primary evidence for non-egoic experiences? Probably, introspecting non-egoic episodes ourselves. However, few of us seem to have had them – and, contrary to Hume, Sartre, and Flanagan, even fewer of us describe our everyday experiences as non-egoic. Such experiences appear to be rare in the general population but also in the life of any individual. They are also hard to induce reliably in an experimental setting. Therefore, we should not expect a great amount of data from introspection. Additionally, in states where individuals might experience ego-dissolution, a broad range of cognitive capacities also appear to be altered, so we should not expect data from reliable, rigorous, or systematic first-person methods. These worries strengthen general doubts concerning introspection or first-person data (see e.g. Schwitzgebel, 2012, 2008), doubts we should take seriously. Introspection – an individual directly registering her own mental states – may lead the introspecting individual herself to believe in non-egoic experiences; but introspection does not provide sufficient reason *for the scientific community as a whole* to accept the existence of such experiences.

The scientific community as a whole will prefer intersubjectively accessible evidence. But phenomenal experiences themselves cannot be intersubjectively observed. So maybe we must consider *indirect* or *secondary* evidence – not direct evidence for the presence of such experiences, but evidence for *others* having ev-

Fink, S. B. (2020). Look who's talking! Varieties of ego-dissolution without paradox. *Philosophy and the Mind Sciences*, *1*(I), 3. https://doi.org/10.33735/phimisci.2020.I.40

idence of the presence of such experiences. Ideally, such evidence comes in the form of non-verbal behaviour. But behaviour appears to be silent concerning the presence or absence of a self, and we currently lack any decisive neural or physiological markers for such states. Alvin Goldman (1997) argues that in order to establish such markers as reliable indicators of consciousness, we have to calibrate and validate them by using first-person reports. Only then could we build a foundation that allows us to use such non-verbal markers as evidence.

It seems that there is no way around first-person reports, at least in an initial stage, whether free-form or as questionnaire responses. Our core evidence for the existence and nature of phenomenal ego-dissolution is therefore verbalised self-ascriptions of episodes of non-egoic experiences, which I abbreviate as SANE. Usually, these will be presented as recollections of having been in a non-egoic state, but sometimes (e.g. in psychopathological cases) they might be presented as reports of an ongoing non-egoic episode.[10] Some contain possessive phrases (e.g. "*my self* was absent"), while others use the indexical "I" to claim nonexistence (e.g. "*I* did not exist").

If SANE are the foundation for establishing that non-egoic experiences exist and what they are like, then that's bad news: Such reports seem to be self-defeating, some claim. If so, we have no evidence for the existence of phenomenal ego-dissolution and, *a fortiori*, any SANE actually speaks *against* phenomenal ego-dissolution: That they exist indicates that they are false, goes the argument. Hence, no research programme on non-egoic experiences can get off the ground: The very evidence taken to prove their existence and their nature, SANE, indirectly proves their absence.

Why should SANE be self-refuting in this way? Here is a rough sketch.

Grammatically, SANE can come in the past tense ("I did not feel like anybody") or in the present tense ("I do not feel like anybody"). In character and form, past tense SANE resemble reports about remembered feelings, dream reports made upon awakening, or retrospective reports about psychedelic trips; SANE given in the present tense (e.g. during meditation) resemble introspective reports about ongoing experiences (see section 1).

So SANE resemble reports. But are they reports? Our natural stance is to *interpret* SANE as reports. Thus, we treat them as assertions, and therefore: as aiming at truth. We presume that the experiencer witnessed her non-egoic experience and afterwards describes what she witnessed more or less veridically (or, at least, describes it with the full intent to be veridical).[11] Thus, if we treat them as reports, we treat them as giving us insight into non-egoic states of consciousness.

[10] For example: At the workshop on radical disruptions of self-consciousness at Frankfurt's FIAS in October 2018, Aviva Berkovich-Ohana showed a video clip of a meditator talking *during* meditation about "falling into space", which may be understood as a report about entering a non-egoic experience: "a sense that there is no need for center [...] as if am falling out of the center. [...] no need to be located anywhere."

[11] Otherwise, we would interpret them as fabrications, fictions, or falsehoods. If they are non-veridical assertions, little is gained for a research programme on non-egoic experiences. Thus, I focus on whether they can be veridical – or something close enough.

Fink, S. B. (2020). Look who's talking! Varieties of ego-dissolution without paradox. *Philosophy and the Mind Sciences*, *1*(I), 3. https://doi.org/10.33735/phimisci.2020.I.40

This interpretation is problematic: How can someone witness something that, essentially, does not allow for the presence of anybody to witness it? If the reporter was present, then it wasn't a non-egoic episode; if it was a non-egoic episode, there cannot be anyone to report on it. Thus, if we interpret SANE as reports, they seem self-defeating, goes the claim.

Either we reject this argument or we reinterpret SANE as something other than reports. I do not buy into this criticism in general: Some SANE some may be meaningful, honest, maybe even truthful. But there is something right: The most prominent understanding of ego-dissolution simply cannot be squared with SANE being reports, I argue. But this simply tells us what ego-dissolutions cannot be, not that they cannot exist in general.

So, what are the hidden assumptions that lead to SANE being self-defeating?

3 SANE rejected: The self-defeating nature of first-person reports of non-egoic episodes

SANE face a very reasonable doubt: They appear to be paradoxical. Why? Because these reports express a claim that the speaker herself has had a non-egoic episode, but come with a corresponding *de se*-belief of the speaker that *she herself* is or has been in such a state. So by reporting, one apparently contradicts the report: If you don't exist, then who's talking about being no one? Who's remembering this? Who's reporting this? Who was the witness?

The notion that SANE are self-undermining has been expressed at least three times. First, by Thomas Metzinger (2004, p. 566) in his *Being No One* (see also his 2018):

> Autophenomenological reports given by human beings about selfless states [...] will usually not impress philosophers much, because they contain an inherent logical fallacy: How can you coherently report about a selfless state of consciousness from your own, autobiographical memory? [...] Such reports generate a performative self-contradiction, because you deny something that is presupposed by what you are currently doing.

If Metzinger is right and such claims or self-attributive beliefs that one has had a non-egoic experience are self-contradictory, then SANE could never be reports and such self-attributive beliefs could never be knowledge, *even if there were non-egoic states*. Non-egoic states, if they exist, are (so he claims) inaccessible for autobiographical memory and self-attributive *de se* beliefs. If we have them in autobiographical memory, they are skewed by the processes of memory formation and retrieval (see also Metzinger, 2018, p. 13).[12] Therefore, we must distrust such re-

[12]Note that on page 13 he writes about "full-absorption episodes", which may differ from non-egoic experiences. What they share is their principled ineffability: "A full-absorption episode cannot

Fink, S. B. (2020). Look who's talking! Varieties of ego-dissolution without paradox. *Philosophy and the Mind Sciences, 1*(I), 3. https://doi.org/10.33735/phimisci.2020.I.40

ports and self-attributive beliefs. However, if we distrust SANE, then we have no intersubjective reason to believe that such non-egoic experiences exist. At the very least, we must be agnostic vis-à-vis phenomenal ego-dissolution, despite the existence of SANE.[13]

Rocco Gennaro (2008) makes a similar point:

> So then a real problematic case would be one where there is a claim to have a [pure conscious event], and thus a truly *introvertive* [i.e. looking inside the mind, SBF] mystical experience, but where there is no conscious I or self present. But it is very unclear that there are such cases. [...] [W]e do indeed find reference to an "awareness of self" and the conscious employment of the concept "I". [...] More theoretically, it seems to me that anyone having a truly introvertive experience must be consciously employing the I-concept. For one thing, the practitioner is clearly taking the mental state to be *her own* as opposed to someone else's. For another thing, it is difficult to understand how practitioners can later remember and describe these events without having employed conscious I-thoughts during the alleged [pure conscious event], that is, without having experienced the event as one's own.

Gennaro stresses the necessity of a *de se* element in such beliefs: Subjects must attribute these experiences to themselves. They did not happen to nobody or to someone else. Thus, mechanisms for cognitive self-reference must have been active in such episodes, making them egoic. There must be some "awareness of self". Therefore, these claims do not prove the existence of non-egoic episodes.

The problem has been raised, more casually, by Charles Foster (2016, p. 6) in his *Being a Beast*:

> J.A. Baker [author of the book *The Peregrine*] pursued his peregrines to the point of assimilation with them. His express purpose was to annihilate himself [and become a peregrine]. [...] As a method, dissolution creates great literary difficulties. If J. A. Baker really disappears, who is left to tell the story? And if he doesn't, why should we take the story seriously?

be reported, because the self-referential mechanisms of forming an autobiographical memory are suspended. Therefore, only the process of entering into it or of emerging out of it can be faithfully represented in the autobiographical self-model; the episode itself is not a part of the subject's inner life narrative."

[13] That SANE are self-defeating is good news for those who hold that consciousness is essentially self-involving, like e.g. Zahavi & Kriegel (2015). Such a position appears to be a pretty safe bet if there can never be a datum speaking against it thanks to the self-undermining nature of SANE. But it can hardly be so easy. First, it seems to make the self a necessary feature of phenomenality for *conceptual* reasons. But usually, the idea that consciousness is always self-involving is posited as a *phenomenological* claim, something we have to discover by first-person methods rather than linguistic analysis. Second, SANE pop up widely: They are symptoms of certain psychotic episodes, appear to capture aspects of psychedelic trips, and are used to describe the essence of deep meditative states. It seems odd to explain all of this away by attesting to widespread conceptual confusion.

Fink, S. B. (2020). Look who's talking! Varieties of ego-dissolution without paradox. *Philosophy and the Mind Sciences, 1*(I), 3. https://doi.org/10.33735/phimisci.2020.I.40

Indeed, why should we take such reports seriously? An answer ought to be given before we invest time and effort into offering an explanation for phenomenal ego-dissolution itself. I defend the idea that such reports, while they appear self-contradictory, are not necessarily problematic at their core.

So what, exactly, is the problem? Let me start with what is *not* the problem before presenting the problem's different flavours.

Contrary to Gennaro, the problem is not raised by the first-person indexical "I" or the first-person possessive "my". Consider Sartre's phrase:

(Sartre) When I run after a streetcar, when I look at the time, when I am absorbed in contemplating a portrait, there is no *I*.

Who was running? Sartre was! Who wasn't there? Sartre. This aftertaste of paradox remains if we replace "I" with its referent:

(Sartre*) When Sartre runs after a streetcar, when Sartre looks at the time, when Sartre is absorbed in contemplating a portrait, there is no *Sartre*.

What raises the problem is, apparently, that asserting that one does not exist, if true, precludes the possibility that one can assert it. By saying that one does not exist, one marks oneself out as a liar. Or a lunatic, because denying one's own existence precludes one's ability to be rational. Or a driveler, because by saying that one does not exist, one talks nonsense. Apparently, so the argument goes, one cannot truthfully assert a Sane because its truth undermines any possibility of asserting it. At the heart of the Sane paradox is therefore an infelicity of performance, as diagnosed by Karl-Otto Apel (1976, p. 73).[14] You say that you don't exist, but: Look who's talking! The problem is not a logical one because what is expressed in the utterance can be perfectly true: It is possible that a certain person a does not exist. But while everybody else can express this, a can not. So the paradox is not in what is said, not in the syntax or even the semantics of what is expressed; it arises because of an infelicity in performing the speech act of assertion. We may follow Apel and call it a *performative fallacy*. The contradiction does not arise simply because of the proposition being expressed, but by expressing it.

Why is it a *fallacy* at all, not merely an infelicity? Because from what is said ("I do not exist.") and the fact that this is said, we can derive a contradiction: What is said entails that the referent of the name or first-person pronoun does not exist; but the fact that it is said requires the existence of a speaker; and the speaker is the referent of the first-person pronoun or name. It is perfectly possible that Apel does not exist; it is just impossible that Apel can assert his own nonexistence truthfully. The contradiction arises not as a result of the proposition p itself, but only if the person this proposition p is about expresses, assents to, asserts, believes

[14] Apel bases his analysis on Wittgenstein's *On Certainty*, Hintikka's *Cogito, Ergo Sum*, and Stegmüller's *Metaphysik, Skepsis, Wissenschaft*.

Fink, S. B. (2020). Look who's talking! Varieties of ego-dissolution without paradox. *Philosophy and the Mind Sciences, 1*(I), 3. https://doi.org/10.33735/phimisci.2020.I.40

p. Any stance of the speaker towards p that requires honesty or veridicality is itself precluded by p. Call this the Sane paradox.

The Sane paradox involves some important presuppositions:

(Unique) For each individual,[15] there is one and only one self that is the referent of self-reference and fulfils the functions of self-reference, or only one mechanism facilitating all kinds of self-reference.

(PhenSelf) A self (or any mechanism facilitating self-reference) necessarily comes with a phenomenal feeling of self.

Unique is an ontological commitment to selves as concrete entities and their distribution. PhenSelf is a thesis that connects a *feeling of* self to something that is felt, namely an underlying self. Not all feelings-of require that there be something real they are feelings-of: I can have a feeling of being followed without being followed. But according to PhenSelf, the feeling of self is a necessary correlate of a self, so if the feeling of self is missing, the self is gone with it.

Together with Cease, we can see how the problem arises: If there is one and only one self (Unique) and if this self necessarily shows itself in experience (PhenSelf), then – if a feeling of self is missing (Cease) – any self-reference is impossible. But self-reference seems to be necessary in order for this state to be cognitively accessible *as one's own*.

This immediately suggests several ways of dealing with Sane: Either we can accept these assumptions or we can reject them. If we accept them, we give a cognitive or pragmatic diagnosis. A pragmatic diagnosis works on the level of communicative practices: What is said in a Sane is not what is meant (but what is meant is meant truthfully). A cognitive diagnosis rejects implicatures as tools for explanation and instead targets the mental state of the speaker: what is said in a Sane is what is meant, but what is meant cannot be rationally believed. So it is either not believed at all or it is irrationally believed. I will address these diagnoses in a moment.

But we may also reject any or all of the assumptions of the Sane paradox. This opens the door to phenomenological diagnoses, where the focus is on phenomenal features being present or absent and why they are present or absent. Such phenomenological diagnoses therefore often cross the border between mere phenomenological description and theory about the mechanisms underlying phenomenal consciousness. If we reject Unique, we suggest that many selves can be present in one body, and even if one of these selves does not show itself in phenomenality, other selves may still be present and do their work. Call this the strategy of *multiplicity*. If we reject PhenSelf, we suggest that a feeling of self can be lost without the mechanism of cognitive self-reference being impeded. Call this the strategy of a *merely cognitive self*. If we reject Cease, then people retain their feeling of self, but apparently fail to notice that it is still there. I will introduce a

[15] If we want to qualify this: For each individual at a moment in time t.

Fink, S. B. (2020). Look who's talking! Varieties of ego-dissolution without paradox. *Philosophy and the Mind Sciences, 1*(I), 3. https://doi.org/10.33735/phimisci.2020.I.40

version of this as the strategy of *ego-expansion*. We may also reject DIFFERENCE and say that there is no felt difference between non-egoic experiences and our normal way of experiencing. I will consider this strategy as well as the others in 5. There, I will also discuss problems with forming autobiographical memories of non-egoic experiences. But before we go so far into diagnosing radical disruptions of self-consciousness, let me start with the pragmatic and cognitive diagnoses.

4 SANE diagnosed: What is said vs. what is meant, believed, rational, true

4.1 A pragmatic diagnosis

One conclusion one might draw from the performative fallacy is that no SANE has any meaning because SANE cannot have truth conditions. So either they are nonsense or they share their meaning with all other contradictions.

This is the wrong conclusion to draw. Yes, if we accept all five assumptions, what is said in a SANE leads to a contradiction, but this does not entail that they are meaningless. Remember that the contradiction follows not from what is said, but from saying it. Therefore, the proposition expressed by what is said can be the case. The utterance is perfectly intelligible – it simply is necessarily false when I say it.

The question is then: Why should someone say something that is immediately falsified by saying it? Maybe what is said is not what is meant. If people assert that they don't exist, they violate conversational principles (Grice, 1991), namely the maxim of quality: "Try to make your contribution one that is true." If a SANE cannot be true, it triggers conversational implicatures: the SANE must mean something other than what the speaker says. Just as obviously false statements like "I wasn't myself that night", "There is no time like the present!", or "We're busy doing nothing" can be uttered in a meaningful way, so can SANE.

What might a speaker mean with a SANE? They may intend to convey things like "I don't feel well" or "Nobody is noticing me", or express desires like "I wish I weren't here", or boast that "I got totally lost in what I was doing, paying no attention to myself", or "Look, I have acquired enlightenment, admire me!", and so on. None of this requires any form of actual ego-dissolution.

Instead of being a truthful report, what appears to be a report about a non-egoic state performs a different conversational function. It therefore does not require anybody to make the ontological commitment that non-egoic experiences exist. In fact, the hearer can decode the message *only* because such reports are openly self-contradictory, similar to a case where a speaker excuses her previous erratic behaviour by saying "I wasn't myself". Some variants of SANE might then be a complex way to communicate one's unease with oneself. This is a good fit for cases of apparent ego-dissolution in depression.

Fink, S. B. (2020). Look who's talking! Varieties of ego-dissolution without paradox. *Philosophy and the Mind Sciences, 1*(I), 3. https://doi.org/10.33735/phimisci.2020.I.40

However, this pragmatic diagnosis has two drawbacks. First, it does not explain the details with which such reports are usually embellished. Reconsider Saks talking about herself dissolving like "a sand castle with all the sand sliding away in the receding surf". These metaphors fulfil no further pragmatic function. Rather, they fit our understanding of SANE as detailed reports. Second, this diagnosis hardly explains cases of *written and kept-hidden* reports of non-egoic states, e.g. in diaries. Given that these writings exist and may be revealed *after* an author's death, we should be open to the possibility that some of them are intended as earnest reports, not as pragmatic play.

If we can reasonably rule out the intention of triggering drastic pragmatic implicatures and deviations of what is meant from what is said, we need another diagnosis for what underlies the expression of a SANE for cases that do not lend themselves to pragmatic explanations. This leads us to cognitive diagnoses.

4.2 Cognitive diagnoses

If all the signs suggest that a speaker actually means what she says in a SANE, despite the apparently self-defeating nature of SANE, then we have to ask what brought the speaker to make such an assertion. The possibilities include: (a) the speaker is confused and does not actually believe what she says; (b) if the speaker shows signs of believing what she says, she must be irrational; or (c) if the speaker shows signs of believing what she says and signs of being rational, she intends to deceive us.

4.2.1 Unbelievable?

Usually, we might think that a SANE cannot be believed because believing means taking something to be at least possibly true. But it cannot be true that one believes any p when one does not exist. Therefore, SANE cannot be believed.

But, as Roy Sorensen (2004, pp. 70–71) argues, we are sometimes led to believe necessarily false statements like contradictions, just as badly designed calculators used to. The old calculators simply rounded off after the last number they were able to display: If you divide 1 by 3 and multiply by 3, you get 0.99999, but not 1. One might defend this by ascribing to the calculator the clever "belief" that 1=0.99999… But this ascription fails if we magnify the miscalculation. A calculator should believe that

$$1 \div 3 \times 9,999,999 = 1 \times 9,999,999 \div 3$$

thanks simply to the commutative properties of multiplication.[16] But if we do the calculation on these old machines, we get 3,333,333 for the right-hand side of the equation, and 3,333,332.97 for the left. Obviously, 3,333,332.97 $\neq$ 3,333,333. We

[16] Most old calculators do not have brackets, but perform calculations in the sequence we type. The reading with brackets is $(1 \div 3) \times 9,999,999 = 1 \times (9,999,999 \div 3)$.

Fink, S. B. (2020). Look who's talking! Varieties of ego-dissolution without paradox. *Philosophy and the Mind Sciences*, *1*(I), 3. https://doi.org/10.33735/phimisci.2020.I.40

also know that we may have contradictory preferences. Obviously, preferences should be transitive: If we like a more than b, and b more than c, we should like a more than c. But, as Amos Tversky (1969) showed, we apparently do not work like this. Instead, we sometimes prefer c to a. Under some circumstances, our inbuilt tendencies to accept some contradictions become apparent. But unless this is made obvious to us, we may still assent to self-contradictory statements with conviction. Believing contradictions is not unusual, either for us or for machines.

So just because SANE are contradictory, a person need not be confused about what she said in a SANE. She can believe what is said. We simply have to ascribe to her a *false* belief. Falsity due to self-contradiction does not add a lot to this because we already know that delusions can be self-contradictory (Bortolotti, 2009). So nothing speaks against explaining away non-egoic states by declaring "reporters" of SANE to be delusional. And because the falsity of these beliefs is apparently so obvious, we might as well question their rationality. But do we have to?

4.2.2 Irrational?

If a person offers a SANE, she apparently disqualifies herself from being rational – only the insane come up with SANE. Why? In order to *rationally* believe a proposition, one needs to have adequate reasons to believe it. While one can have a *practical* reason to believe something self-contradictory (e.g. someone may pay me a large sum if I do), one cannot have a *theoretical* reason to believe a contradiction, because logical consistency trumps other reasons. Therefore one cannot rationally believe something self-contradictory. This suggests that our reporters of SANE must be irrational or delusional. They have gained false beliefs about their own past or present and do not let go of them despite strong and obvious inconsistencies.

But someone need not be irrational just because they believe in contradictions. Roy Sorensen (2004, pp. 146–147) even argues that reason *demands* belief in infinitely many contradictions. Consider the very reasonable meta-belief that one of my beliefs is *false*. This precludes my belief system from being coherent and, taken as a whole, fully true. It is therefore rational to believe that one believes a contradiction, if one believes that one believes a contradiction or something false. This merely shows that one can rationally believe in contradictions, but it does not identify *which* of one's beliefs is inconsistent. Yet "*detected* contradictions are instantly abandoned" (Sorensen, 2004, p. 155).

However, this is not what happens in the case of a SANE: Meditators likely remain steadfast that they have had non-egoic experiences even if they are aware of the SANE paradox; Saks still reports on her episodes of ego-dissolution, and tries different literary devices to sidestep the paradox. Neither she nor the meditators abandon their beliefs in their past non-egoic experiences.

We may take this as strong evidence of irrationality. But it raises more issues to explain. We need to give an explanation for the *stability* of such beliefs despite their obvious self-contradictory nature. What is also needed is an explanation

Fink, S. B. (2020). Look who's talking! Varieties of ego-dissolution without paradox. *Philosophy and the Mind Sciences, 1*(I), 3. https://doi.org/10.33735/phimisci.2020.I.40

for their *ubiquity* in a broad range of different circumstances, from meditation to psychedelics to psychopathologies. Irrationality will not likely be a one-size-fits-all solution here. Should we really think that meditators try to achieve irrationality as a goal? This is easily achieved without countless hours of sitting still.

I am suspicious of the idea that all SANE are best interpreted as irrational delusion. Delusions are often highly idiosyncratic: I might believe that I am Napoleon, while you may believe that you are Jesus. Even when there is some similarity (you and me both believe we are famous dead people), our delusions differ massively in their details. In contrast, SANE appear to be very homogenous across individuals. What is therefore also needed is an explanation of the *specificity* of this form of irrationality. We need some explanation: Why do specifically *de se* beliefs about non-egoic episodes arise, and not other self-contradictory beliefs like "I am saying nothing"? Why are they so symptomatic of specific pathologies, but not of all pathologies?

While some SANE might be due to delusion or irrationality, it is unlikely that all are. But if SANE cannot be true and the speaker is neither delusional nor irrational, then are they merely acting *as if* they believed in non-egoic episodes?

4.2.3 Untrue

One diagnosis of SANE is that they are not made in honesty. If we believe this, we may feel pressured to call a SANE "reporter" a liar or bullshitter: What look like SANE are false assertions. Moreover, they are not contingently false, but *necessarily* so, whenever they are stated. Why should people assert something necessarily false if they do not want to trigger the corresponding implicature?

One explanation is that such speakers are trying to deceive. Given that having non-egoic experiences is sometimes seen as a mark of high levels of spiritual enlightenment, which come with high social status in some religions, there are high incentives to lie about having non-egoic experiences in order to gain such status.

However, the lie is extremely blatant, for it is logically impossible to make such an assertion in earnest. This makes it a bad candidate for a lie. A good lie expresses something that can at least possibly be the case. Otherwise, nobody buys into the deception. It simply is too obvious a lie if one claims that $2 + 2 = 5$. Yet some impossibilities are harder to track. Depending on context, even an impossibility might be bought as being the truth by a specific audience. For example, one might try to impress a stranger by claiming to be that famous mathematician who discovered the largest prime number; or one might claim that there is an all-powerful, all-knowing, and all-loving being that allows for suffering in the world but cannot create a stone too heavy for it to lift. Some SANE might be similar deceptions.

However, seeing such "reports" as deceptions for the reporter's own benefit fails to explain why such reports regularly appear in certain psychopathologies. Social status is largely reduced here. Thus, there is some incentive to *hide* one's mental illness, not advertise it by spewing SANE.

Before I get into this, let me emphasise: It would be foolish to think that all

SANE are to be judged equally. Some may be lies to achieve higher social status, but not all; some may be delusions, but not all; some may stem from irrationality, but not all; some may be nonsense, but not all; some may be made in order to express by implicature that one does not feel at ease with oneself, but not all. We have to look at the specific context and speaker to decide.

If we accept the SANE paradox raised by Metzinger, Gennaro, and Foster, our research programme on non-egoic experiences reduces to a study of pragmatics, delusions, lies, and doxastic failures. But it is not a study of phenomenal experiences. In order to give such a phenomenological diagnosis, we will need to reject some of the presuppositions underlying the paradox. If we can, then these utterances may express something that can be the case. So, can SANE actually be accurate descriptions of certain phenomenal experiences?

5 SANE revived: Four ways to lose one's self

5.1 Rejecting the SANE paradox

The pragmatic and the cognitive diagnoses are all available to explain why a SANE may be given by a person. Each of these diagnoses allows us to accept the presuppositions of the SANE paradox and remain agnostic as to whether phenomenal ego-dissolutions exist at all.

Even though we should accept these diagnoses as parts of a toolbox, they do not lend themselves to building a research programme focused on the *phenomenality* of ego-dissolution. For this, we have to reject the paradox.

Let me re-address the presuppositions behind the SANE paradox. Non-egoic experiences are supposedly marked by the following features:

(CEASE) A feeling of self is missing in some experiences.

(ACCESS) The absence of a feeling of self is cognitively accessible.

(DIFFERENCE) There is a felt difference in the feeling of self between the time before/after and during non-egoic episodes.

Two additional assumptions give rise to the SANE paradox.

(UNIQUE) For each individual, there is one and only one self that is the referent of self-reference and fulfils the functions of self-reference, or only one mechanism facilitating all kinds of self-reference.

(PHENSELF) A self (or any mechanism facilitating self-reference) necessarily comes with a phenomenal feeling of self.

Fink, S. B. (2020). Look who's talking! Varieties of ego-dissolution without paradox. *Philosophy and the Mind Sciences, 1*(I), 3. https://doi.org/10.33735/phimisci.2020.I.40

Only together do they preclude the possibility of reporting on non-egoic experiences.

But we might reject Unique, Cease, or PhenSelf and thereby preempt the Sane paradox. If not *the* feeling of self, but merely *a* feeling of self dissolves, then another mechanism facilitating self-reference might bring about relevant *de se* beliefs. If only one feeling of self exists but it doesn't dissolve, then *de se* beliefs can occur. If self-referencing mechanisms need not show themselves in experience, then the relevant *de se* beliefs can occur.

But before I go into these non-paradoxical accounts, is there maybe a way to account for all the features of Sane while accepting the assumptions leading to the paradox? Someone may defend the idea that one can provide Sane from episodes of full ego-dissolution by relying on specific forms of memory. I do not think that this works well. Let me explain why.

5.2 Can we defend total ego-dissolution against the Sane paradox?

Total ego-dissolution would be the case if Unique, Cease, and PhenSelf were all accepted. So there is only one singular self or mechanism facilitating self-reference and self-consciousness which necessarily manifests itself in a specific feeling of self and which is inactive or nonexistent during such episodes. If this were the case, then we should not get any Sane when a feeling of self ceases. In this framework, if we ascribe a non-egoic experience to *ourselves*, then there must have been a self or a mechanism facilitating self-reference in order to bring about this *de se* ascription.

Some may accept these theses, but still find the Sane paradox unconvincing, specifically when it comes to memory. It might be easy to see why one cannot report on oneself as currently undergoing an episode of total ego-dissolution. One would need to speak of oneself in the third-person singular or use a definite description that happened to refer to oneself (e.g. "the speaker of this sentence"). For the type of *de se* reference necessary for a *self*-ascription, however, we would need some way of modelling ourselves in order to refer to ourselves *as ourselves* – and not merely to ourselves *by accident*.

But one might think this: Even though one could not report on such experiences while one is undergoing them, one might remember them and report on them afterwards as having happened to oneself. Metzinger (2004) apparently rejects this possibility, because "How can you coherently report about a selfless state of consciousness from your own, autobiographical memory?" Certainly, *auto*biographical memory relies on self-reference across time, which relies on a model of the progression of a self. But why shouldn't some other form of memory suffice? Who else could this experience have happened to, when *all* the experiences that I can remember are necessarily my own? Could I not validly *infer* later on that this experience was my own? If so, we need not reject any of the premises. Instead the Sane paradox is restricted: Only present-tense Sane

Fink, S. B. (2020). Look who's talking! Varieties of ego-dissolution without paradox. *Philosophy and the Mind Sciences, 1*(I), 3. https://doi.org/10.33735/phimisci.2020.I.40

are impossible, but we may still report veridically about non-egoic experiences, if only *retrospectively*.

This might seem like a reasonable reply, but only because it relies on a misunderstanding: Semantic memory is not the same as episodic and autobiographic memory. That is, some memories of facts will not be *de se* memories (Conway, 2005; Conway & Pleydell-Pearce, 2000; Klein & Loftus, 2014; Klein & Nichols, 2012). I remember a lot of things that are facts, but not about me, e.g. when the Second World War ended, that Paris is the capital of France, that Joaquin Phoenix had a brother called River. Even if a fact concerns ourselves, a memory of that fact is not automatically a *de se* memory (Conway, 2005; Klein & Loftus, 2014). For example, you might remember that someone knocked over the lamp during the party last night, but you might fail to remember that it was *you* who knocked it over.[17] As David Lewis (1979) argues: No amount of knowledge of facts, even up to divine omniscience, can replace knowledge *de se*. And it is only knowledge *de se* that allows us to locate ourselves in the world. God would fail to know where and who she was if she had only factual omniscience. If one wants to defend the possibility of veridical reports from total ego-dissolution, one has to show how one can form *de se* memories of one's own episode of total ego-dissolution. How could that be?

Two stages are the most promising candidates during which memories may be imbued with a *de se* aspect: the *formation* and the *retrieval* (or recall) of the memory.[18] In cases of total ego-dissolution, the *de se* aspect of memory cannot arise during formation, because at that point a self (or a mechanism facilitating self-reference) and therefore the basis for self-reference is missing. This is Metzinger's, Gennaro's, and Foster's point. Therefore, the *de se* aspect must be inserted during reconstruction, injected during retrieval, added on during recall.

Unfortunately, there is no specific noticeable trace of what exactly is "added on": A false memory of me knocking over a lamp feels just as real as a veridical memory of me knocking over a lamp. Sometimes, it does happen that we "add ourselves" to memories of affairs that did not involve us (Hyman Jr, Husband, & Billings, 1995; Loftus & Pickrell, 1995). Someone might reply that these are therefore not memories at all, but pseudo-memories that are simply phenomenologically indistinguishable from real ones; what makes something a memory is not how something feels, but whether what is remembered happened as it was remembered. Yes, there can be false memories with a *de se* component added later; but there can also be *true* memories with a *de se* component added later. This is obvi-

[17] Quite a few thrillers involving amnesia rely on this distinction: One remembers that something happened, but one forgets that it happened to oneself. For example, *Shattered* (1991), directed by Wolfgang Petersen with Tom Berenger, Bob Hoskins, and Greta Scacchi, distributed by Metro-Goldwyn Mayer, uses this plot.

[18] Memory storage does not seem like the right kind of stage for a *de se* aspect to creep in, mainly because adding the *de se* aspect seems to involve specific mechanisms, namely those facilitating self-reference. Storage, usually understood as a passive, even if memory-altering, process, is usually too unspecific to introduce this specific aspect. Still, most of what I say about formation and retrieval would apply to *de se* introduction during storage as well.

Fink, S. B. (2020). Look who's talking! Varieties of ego-dissolution without paradox. *Philosophy and the Mind Sciences*, *1*(I), 3. https://doi.org/10.33735/phimisci.2020.I.40

ously right. As an example: I cannot remember things from my early childhood. Most of what I know from that time, I know from pictures and hearsay. Say I see myself on one of these photos in the arm of my uncle Karl smoking a pipe. I know that it happened to me, but I know this not by remembering the experience itself. However, after a while, I might forget how I formed this knowledge that my uncle Karl held me while smoking a pipe. I might form a *de se* memory about this fact that I take to be one formed in my early childhood. Earlier, it was a fact about myself that I knew, stored in semantic memory; now I remember it as a *de se* fact with a phenomenological *de se* component that was added on later – and thus it has been turned into an episodic memory.

The problem is not that there cannot be true or accurate memories of facts with a *de se* component added on later. The problem arises from the "adding on" itself. For if this "adding ourselves in" is part of the reconstruction, then such memories fail to show that there actually was an episode of ego-dissolution – for it is indistinguishable from the outside or the inside whether the *de se* aspect was added on later or accurately remembered as being present during the experience. If the *de se* aspect was previously present, then the experience was egoic. If one agrees that it was a later mis-reconstruction, addition, or embellishment, then one should be doubtful that the rest of the experience was remembered accurately. We cannot be sure that we are learning something about the deep structure of non-egoic experiences if the reports are edited. But one cannot distinguish between these cases.

But what if the *de se* addition is justified by another memory system? What if one does not remember a proposition about someone and "add on" the *de se* component; what if instead one remembers *imagistically*?[19] That is, one remembers the sensations felt during that episode *as sensations*, as one conjures up the scene of what it was like to sit in one's mother's kitchen as a toddler. In this remembered stream of sensations, a felt self could be missing. But because sensations always have to be mine (I don't feel anybody else's), I can reasonably *infer* that these sensory impressions with a lack of felt self must have been mine. What is added is a possessive *de se* component, not a phenomenal one: The experiences were mine, but I was not in these experiences. But again: What ensures that the *de se* component was not there in the remembered experience?

[19]What I am talking about, in effect, could be interpreted as *iconic* memory. But I am concerned about the distinction between visible and informational persistence in iconic memory (Coltheart, 1980). Visible persistence is supposedly very short-lived (less than 300 ms) and therefore is not suited to playing the role attributed to "imagistic memory" here. Informational persistence is longer-lived, but preserves only the information in the visual stimulus. It is *about* something visual, but is not visual itself. It is therefore not like remembering in the form of a mental image, which is specifically what would be needed for this type of defence; otherwise, the objections against semantic memory of such episodes can be applied. It seems that there is no obvious candidate in psychology that is able to play the role of "imagistic memory". But one would need such a form of non-semantic memory to get this defence off the ground. Therefore, I discuss the issue despite a lack of match to empirical psychology.

Fink, S. B. (2020). Look who's talking! Varieties of ego-dissolution without paradox. *Philosophy and the Mind Sciences*, *1*(I), 3. https://doi.org/10.33735/phimisci.2020.I.40

What if someone holds that a self is a necessary requirement for experiences in general? Any form of consciousness, we are often told, involves a *first-person* perspective. And what is a first person other than a self? Maybe it is the perspectival nature of experience which must point at someone, but not at anyone: at a self as the owner of the experience, as that whose experience it is, as the centre of the perspective. The feeling of self is not something extra, but is somehow ingrained in the structural composition of phenomenal sensations because of their perspectival nature. Maybe this justifies the inference that those were my experiences if I become aware of the perspectival nature of my imagistically remembered sensations.

I agree that our sensations, especially our auditory and visual experiences, are perspectival. But I do not think that this forces us to accept a self outside experience, a mechanism facilitating self-reference, or even a feeling of self as part of our experiences. Our experiences are perspectival and thereby indicate a centre in experience. But the perspectival nature of experiences does not necessarily indicate *oneself* as the centre; it just indicates *someone*. Consider, as an analogy, *Being John Malkovich*.[20] In this movie, people can enter the mind of John Malkovich through a door in the $7\frac{1}{2}$th floor of a Manhattan building. When a person enters through that door, we see the succeeding scene unfold from John Malkovich's first-person perspective. Obviously, the visual imagery suggests that these experiences are someone's, but they need not be *mine*. Neither set-up nor framing specify whether these visual experiences are those of the person entering the mind of John Malkovich or of John Malkovich himself. And when seeing the scene as an observer, I need not feel like I am John Malkovich. The perspectival nature of the visual sensations is preserved, but they do not necessarily therefore feel like my own perspective, when I see the scene. The same goes for the people undergoing the experience in the movie: They do not confuse themselves with John Malkovich, but simply partake in his stream of consciousness – they experience this stream as John Malkovich's. Therefore, conceptually, the perspectival nature of sensory imagery is not sufficient to indicate that these sensations are one's own, that the vanishing point of this perspective coincides with *my* self. The difference between experiencing a perspective and experiencing a perspective *as being my own* is then not one of sensory impressions at all. *Selfhood* is not a sensory quale and not part of the imagistic array I remember or experience. This does not mean that there is no feeling of self at all, but simply that there is no *sensation in any sensory modality that is that feeling of self*.

If the perspectivity of our sensory experiences does not suffice for a feeling of self inside experiences or necessarily indicate a self outside our experiences, then attributing a remembered sensory experience to myself is something problematic for allegedly non-egoic episodes. In order to remember this stream of consciousness *as being my own*, a *de se* component is again "added on" to the perspective in recall. And again, we – whether as memorisers or as external observers – cannot

[20]Universal 1999, directed by Spike Jonze, written by Charlie Kaufman.

Fink, S. B. (2020). Look who's talking! Varieties of ego-dissolution without paradox. *Philosophy and the Mind Sciences, 1*(I), 3. https://doi.org/10.33735/phimisci.2020.I.40

determine whether the *de se* component was present in the first place or added on later.

So if there are non-egoic experiences, then we cannot have first-person reports of them, for this requires reference to oneself *as* oneself. If there are such experiences, then we cannot have autobiographical memories of them for the same reason. If they exist, then no purported autobiographical memory would prove their existence, for these would be indistinguishable from mis-reconstructions or mis-categorisations of egoic episodes as non-egoic ones. While this does not prove the nonexistence of episodes of total ego-dissolution, there is no way to demonstrate their existence by providing "reports". This puts them in the same category as unicorns, whose existence one also cannot prove by pointing to "reports" and drawings of them.

We should therefore consider alternatives to total ego-dissolution where we do not need to posit something being "added on" to a memory. Then, it might be less controversial how we can see Sane as indicators, in some sense, of non-egoic episodes.

5.3 Multiple feelings of self: Plurality or modularity

Let us accept Cease and PhenSelf but reject Unique: A feeling of self can temporarily cease to exist but there could be many differentiable phenomenal "selves" in our experiences. These might include a feeling of ourselves *as a body*, another *as a thinker of thoughts*, another *as a person in social relations*, and so on – each likely associated with a dissociable cognitive mechanism that facilitates this form of self-consciousness. Obviously, these distinct feelings of selves in each of us may influence one another, but they could still be dissociable. If they are dissociable, one feeling of self can cease to exist while others persist. Then, one can reasonably say that one of these – *a*, but not *the*, phenomenal self – dissolves while a system retains other mechanisms sufficient for *de se* beliefs or cognitive self-reference. Even though he does not endorse it, this position may be developed from work by Raphaël Millière (2017), who, in order to give an account of what happens in ego-dissolution, helpfully distinguishes between different forms of self-consciousness, e.g. the disruption of self-referential thoughts, of narratives about oneself, of body ownership, of bodily self-awareness, and of bodily self-location (and extension).[21]

We can construe the thesis of "multiple selves" (i.e. multiple feelings of self at one time in one system) in two ways: as modularity or plurality. In the case of modularity, there is no single homogeneous mechanism bringing about a feeling of self; different modules specialise in different tasks, among them those mentioned by Millière (2017). Each module can bring about a kind of self-reference and self-relatedness independently from others, and each of these forms may, if we accept PhenSelf, manifest itself in phenomenal consciousness in a distinct way. But just as the knife, saw, and corkscrew still contribute to one Swiss Army knife, each

[21] See also Gennaro (2020, this issue), as well as Millière (2020, this issue).

Fink, S. B. (2020). Look who's talking! Varieties of ego-dissolution without paradox. *Philosophy and the Mind Sciences, 1*(I), 3. https://doi.org/10.33735/phimisci.2020.I.40

self-module contributes to a larger, integrated, and unique feeling of self, which allows for mediation between these modules and a coherence between these feelings as indicating one and the same self. A disturbance in the modules affects the character of a feeling of self, but there is still *one* feeling of self. Because each self-module can fail independently *and* because such disturbances are noticeable phenomenologically, there can be reports of such partially non-egoic experiences – but only if some of the mechanisms necessary for self-reference are intact, for they enable the formation of first-person reports (e.g. the modules for cognitive or linguistic self-reference) or autobiographical memories.

In the case of plurality, these modules are not part of a larger, integrated feeling of self. Each contributes to forms of self-reference, but each produces its own distinct feeling of self – without any overarching unique feeling of self that they are part of. Some of these can be disturbed while others remain unaffected and therefore enable first-person reports and autobiographical memories of (partially) non-egoic experiences. Simply, there is no unique feeling of self, but multiple feelings of self, each associated with its own distinct mechanism facilitating some form of self-reference.

In both cases, the expression "I" in first-person reports will be ambiguous: It can refer to the bodily self, the social self, the cognitive self, or the reflexive self. If so, then the performative self-contradiction could be prevented by disambiguating. Consider Sartre's statement again:

(Sartre) When I run after a streetcar, when I look at the time, when I am absorbed in contemplating a portrait, there is no *I*.

In this example, the first two instances of "I" obviously refer to a physical body that runs and looks, the third "I" refers to a cognitive self, while the fourth "I" refers to the reflexive self – some being present in experience, some not. If we disambiguate in this way, the paradox ceases to arise:

(Sartre**) When *Body-I* run after a streetcar, when *Body-I* look at the time, when *Cognitive-I* am absorbed in contemplating a portrait, there is no *Reflexive-I*.

This partial breakdown in the chorus of the selves could be what is expressed in Sane. Non-egoic experiences then need not be *total* ego-dissolutions; they are only *partial* disruptions of feelings of self and their associated self-reference-enabling modules.

But I see some drawbacks. First, this explanation does not seem to fit Saks's report, where she claims that *the* centre (singular) does not hold. The definite article "the" here suggests a uniqueness: one and only one centre. Apparently, there is one coherent feeling of self for her, and it has gotten lost. Nor did only some of the feelings of self (as in the case of plurality) get lost; nor did the feeling of self change its character (as in the case of modularity). If the uniqueness expressed

Fink, S. B. (2020). Look who's talking! Varieties of ego-dissolution without paradox. *Philosophy and the Mind Sciences, 1*(I), 3. https://doi.org/10.33735/phimisci.2020.I.40

by Saks is not just loose talk, we should search for alternative explanations that do not rely on a multiplicity of selves or feelings of self.

The second drawback I see is: The account needs to explain how different feelings of self interact in order to give rise to the appearance of *one integrated* self, a feeling of self as inhabiting this body rather than that, *and* as the thinker of thoughts, *and also as* the experiencer of sensations, *plus* as the person addressed by others, and so on. If such an explanation is not provided, this approach remains unconvincing because this high degree of integration is the most prominent features of selves: their *individuality* – their lack of partition. A feeling of self should share this *undividedness* in order to be a feeling *of one self as an individual.* If such an explanation for overarching integration of feelings of self into one is provided, the multiplicity thesis loses some of its drive. Because now, there is apparently one overarching process that forms a coherent feeling of self out of several. The multiplicity thesis also becomes ambiguous: It is unclear what is disturbed in non-egoic states – is it the integrating process or the parts being integrated?

A third drawback I see is that because this explanation relies on multitudes, either of modules contributing to a larger feeling of self or of unintegrated feelings of self, it is not very parsimonious. With Occam's razor, we should prefer an explanation that covers the same territory with fewer entities postulated.

We can also consider an evolutionary take: The amount of sugar one's body burns is a constraint on one's evolutionary success. Finding the right proportion, the sweet spot, between energy conservation on the one hand and access to resources (like food, status, or mates) on the other hand is an evolutionary imperative. So we have to show what an organism gains something by spending energy. If each of the multiple mechanisms for self-reference bringing about a feeling of self demands energy to be active, then the multiple-selves account looks like a very costly model in comparison to one that claims only one unique mechanism facilitating all feelings of self. An explanation without such pluralism might then be preferable, if not for theoretical reasons then for reasons of energy conservation.[22]

Both of these reasons for rejecting the multiplicity account are comparative: Of two or more accounts, they tell us which to prefer. So we should look at further alternatives. If multiplicity is the only game in town, it wins by default. But wouldn't it be cool if there could be more to ego-dissolution than just the disturbance of some-but-not-all feelings of self?

[22]There is also an evolutionary reason speaking *against* a unique mechanism facilitating self-reference: robustness. One mechanism, if disturbed, would lead to a global break down of self-reference. Given that locating oneself, discriminating one's body from others, and other forms of self-reference have an evolutionary advantage, any disturbance of a unique mechanism facilitating self-reference would come with high costs to the organism. In comparison, if there are many independent mechanisms, a break down in one could leave a sufficient level of functionality for continued survival. The question is how to weigh robustness against energy conservation. For this, we would need to compare theories. What matters here is that there are comparative constraints on theory selection (e.g. Occam's Razor), and that they suggest looking for alternatives to compare to.

Fink, S. B. (2020). Look who's talking! Varieties of ego-dissolution without paradox. *Philosophy and the Mind Sciences, 1*(I), 3. https://doi.org/10.33735/phimisci.2020.I.40

5.4 Ego-expansion

Let us accept UNIQUE and PHENSELF but reject CEASE: There is only one feeling of self but it does not dissolve. It is the dissolving of a unique mechanism for self-reference or a unique self-model, indicated by the loss of a feeling of self, that leads to the SANE paradox, so rejecting CEASE disarms the paradox. So what happens in such states if it is not a feeling of self dissolving?

Here is one option:[23] Rather than contracting into nothingness, the feeling of self expands until everything in consciousness becomes part of it or attached to it. Instead of being *nothing*, in these states the feeling of self is associated and involved with *everything*. In this way, our engine for self-reference is still active – indeed, it is hyperactive. If the feeling of self is still there and an indicator of the activity of a mechanism facilitating self-reference, then we can get first-person reports and self-ascriptions of experiences of ego-expansion.

But why would they be about ego-*dissolution* if the feeling of self is not dissolving? The problem might arise from an ambiguity in how we conceptualise such experiences: In both the case of ego-dissolution and ego-expansion, the self/other distinction loses its unique meaning. Why? An empirical concept is uniquely meaningful only if there are some things that fall under its extension and some that do not. If everything falls under the extension of "feeling of self", it is not the feeling *of self* that dissolves, but the *boundary* between a feeling *of self* and the feeling *of anything else* in consciousness. It is the concept of "self" that becomes meaningless, because it lacks any boundary and allows for no distinctions in one's mind during full ego-expansion. The categories *self* or *belongs to me* become meaningless if everything is experienced as belonging to me. In a common but naïve stance, where one projects what is in consciousness onto the world, one *is* everything and one with the universe. One's concept of "self" would be extensionally indistinguishable from the concept of *things being identical to themselves*. The same holds in the case of ego-dissolution: The category of *other* or *not belonging to oneself* becomes meaningless because it encompasses everything in one's mind. So the self/other distinction loses its meaning under both circumstances. "Self" needs "other" to make sense because the two are mutually exclusive opposites. In cases where everything falls under *self* or everything falls under *other*, it is just as meaningful to claim "The self ceased to exist" as it is to claim "The other ceased to exist". In both total ego-dissolution as well as full ego-expansion, the term *self* becomes empty just as the term *other* becomes empty. Thus, first-person reports of "ego-dissolution" might be adequate descriptions of a conceptual dichotomy dissolving.

So in cases where there is no empirical distinction between the concepts "self" and "other" available to oneself, it might make just as much sense to say that everything is *other* (report of ego-dissolution) or to say that everything is *me* (report of ego-expansion). But which is the experience underlying this report? Ego-dissolution or ego-expansion? Total ego-dissolution is disqualified because it does

[23] I am grateful to Franz X. Vollenweider for suggesting this diagnosis.

Fink, S. B. (2020). Look who's talking! Varieties of ego-dissolution without paradox. *Philosophy and the Mind Sciences, 1*(I), 3. https://doi.org/10.33735/phimisci.2020.I.40

not easily square with SANE. (Or, at least, it makes assumptions one need not share.) Therefore, if SANE are indicators of a feeling of self-alteration, and if there is only one feeling of self, we are left with ego-expansion as a suitable alternative.

In the extreme case of ego-expansion, everything going on in consciousness would become part of the self-model, such that nothing could be meaningfully referred to as "other". Only unconscious processes would be exempt from such a hypertrophying self. We would feel as involved in pushing the clouds across the sky as we would in moving our legs. Reports collected by Berkovich-Ohana et al. (2013, pp. 5–6) reflect this tendency for ego-expansion when meditators speak of the bodily space being larger, bodily sensations being wider, a "sense of expansion", a "center of space [becoming] endless", that there were "little bodily boundaries compared to the usual feeling", and so on. Data by Preller et al. (2018) are suggestive of an explanation: They found that "LSD induces hyper-connectivity predominately in sensory and somatomotor areas" (Preller et al., 2018, p. 3). Together with an increased hyperconnectivity in the sensory thalamus, a classic neural hotspot for consciousness and its unified nature, this can be interpreted as: more and more active neural correlates of sensory conscious states are incorporated into and associated with one's own bodily self.[24]

This oceanic explanation of non-egoic states clearly marks such episodes as *egoic*: A feeling of self would be present! And it explains the misconceptualisation that happens in SANE as benign. It might also explain why certain people who claim to have achieved self-annihilation do not simply stop speaking but instead use a plural form. Consider Foster (2016) talking about Baker, the author who wanted to capture what it is like to dissolve and become one with a flock of peregrines: "If Baker is to be believed, it worked [...] the pronouns changed from 'I' to 'we'."[25] This change to the first-person plural may indicate that a felt self has expanded and now encompasses more than what the speaker usually associates with himself. Ego-expansion might thus be a suitable alternative to the multiplicity account and to total ego-dissolution.

An obvious problem for the ego-expansion account is that it does not straightforwardly explain differences in describing these experiences. If we believe that the feeling of self is unitary, then only two options remain for attaining experiences without a self/other distinction, namely total ego-dissolution and ego-expansion. Total ego-dissolution, if we buy into the SANE paradox, is undetectable or ineffable. But still, we have to explain why subjects differ in their

[24] Note that "ego-inflation" in the EDI is associated with "unusually elevated self-assuredness and confidence" (Nour et al., 2016, p. 3), reflected in claims like "I felt more important or special than others" or "I felt especially sure-of-myself". "Ego-inflation" therefore captures something different than "ego-expansion". "Ego-expansion" is a claim about an increase in experiences associated with oneself, but this need not come with self-assuredness or confidence. Thus, phenomenal ego-expansion (in the sense I in which use the term) would not be captured by "ego-inflation" in the EDI questionnaire.

[25] In *The Peregrine*, Baker does shift from using "I" as a way of referring to the narrator to using "we".

Fink, S. B. (2020). Look who's talking! Varieties of ego-dissolution without paradox. *Philosophy and the Mind Sciences, 1*(I), 3. https://doi.org/10.33735/phimisci.2020.I.40

replies to questionnaires or how they describe their experiences: Some answers suggest an "anxious ego-dissolution", some an "ego-inflation" or an "oceanic boundlessness". In this account, we have to explain all in terms of ego-expansion. How might this work?

I am sympathetic to ego-expansion, so I propose the following empirical hypothesis: Subjects conceptualise their experience of ego-expansion as "no self" or "all self" depending on where they are focusing when the expansion happens. Consider this: Our attention can be on our thoughts, our body, our breathing, our digestion, or other aspects of our body and mind. But it can also attach itself to external events, like the sky, water flowing, sounds of neighbours arguing, and so on. It can be on self-related thoughts (*I still have to feed the cat*) or thoughts where the content is unrelated to us (*There is always a prime number between* n *and* $2n$). We know that under the influence of LSD, our attention is altered. Subjects who have taken LSD often focus with extreme intensity on small details, often at the cost of ignoring everything else. Say that while the feeling of self is slowly expanding, one focuses not on one's own hand or breathing but on a feature of the environment (e.g. the veins of a leaf) or a mathematical truth. When attention is shifted again, as the person tries to get back to herself, she realises that she does not find herself where she expected herself to be; she has lost the familiar feeling of her usual, expected self-boundaries because they slowly expanded while she was distracted. The mismatch of predicted self-boundaries in experience and the lack of felt self-boundaries is then interpreted as ego-dissolution – because if it is not where it always is, it must be gone! Compare this to a case where people who have taken LSD focus on their own body or on their body in relation to the environment. The process of a feeling of self expanding is felt *as a process*: We can attend to and appreciate the dynamics and characteristics of its unfolding. As the endpoint of maximal ego-expansion is reached, *we* (not I) know how we got there. The mismatch of prior felt self-boundaries and the lack of felt self-boundaries is interpreted as an expansion of self because we witnessed our feeling of self expanding. What I suggest is that whether we categorise such an episode as ego-dissolution or oceanic self-boundlessness depends solely on how aware we were of the process of expansion. This suggests that people who are more introspective or sensitive, focusing on their body and mind, or more trained in keeping their attention partially on themselves might experience more episodes of oceanic boundlessness than people who lack such training or are more externally focused or do more rumination (i.e. thinking non-self-related thoughts and getting lost in them).[26]

Coming back to the Sane paradox: We can expect first-person reports of such

[26] I suspect that experienced meditators should feel more ego-expansion rather than self-annihilation due to their mental training in attending to their attentional focus. Unfortunately, Buddhist literature appears to talk more about ego-dissolution. This might be a cultural artefact attributable to the status of specific writings and ways of describing one's experience in this spiritual practice. Or it may also speak against my account. Here, unfortunately, I lack the space to clear up this apparent mismatch.

Fink, S. B. (2020). Look who's talking! Varieties of ego-dissolution without paradox. *Philosophy and the Mind Sciences*, *1*(I), 3. https://doi.org/10.33735/phimisci.2020.I.40

episodes of ego-expansion. And we can understand why some of them might rather be phrased in terms of ego-dissolution. So this interpretation can explain the specific structure and existence of Sane if ego-expansion underlies them. Ego-expansion is comparatively more parsimonious than the multiplicity account. However, there is an even more parsimonious solution available.

5.5 No-ego revelation

For total ego-dissolution, we accepted both that there is only one feeling of self (Unique) and that it dissolves (Cease), which got us the Sane paradox on the condition that mechanisms for self-reference must show themselves in phenomenal consciousness (PhenSelf). For the multiple-selves hypothesis, we rejected Unique and accepted Cease, which led to a non-parsimonious but acceptable diagnosis. For ego-expansion, we accepted Unique and rejected Cease, leading to a more parsimonious explanation in comparison to the multiplicity account. Logically, we could reject both Unique and Cease. However, there are two ways in which we can reject Cease: First, there are many feelings of self and everything stays the same; or, second, there is no feeling of self to begin with, so no feeling of self can dissolve. This second interpretation denies, to some degree, PhenSelf: A mechanism facilitating self-reference does not need to show itself in phenomenal consciousness at all. This, the no-ego account, is the most parsimonious attempt at a diagnosis.

So, what if there never was a phenomenal self? Here, we deny the basic presupposition of phenomenal ego-dissolution, because something has to exist before it can cease to exist. But I argued in the beginning, against some interpretations of Hume and Sartre, that this account fails to capture the *difference* between being inside and outside a non-egoic episode. So, how can we reply to this challenge?

At its core, the no-ego account explains this apparently felt difference as a difference in *beliefs* about one's experience. Certain beliefs, e.g. about the proper grammatical construction of sentences, may *coerce* us into believing that there is a self in consciousness. Because how else could we make sense of statements like this:

(Dream) I dreamt I was Napoleon.

The second instance of "I" in Dream appears to pick something out *in the dream*, i.e. it is about what I dreamt of, not who dreamt it. This suggests that there is a feeling of self in experiences like dreaming. But this leads to paradoxes similar to the Sane paradox for cases like "I dreamt I was dead and gone." or "I sometimes imagine myself as never having been born." How do we explain these away? I suggest: Grammar bewitches us.

As an analogy, think of Mark Twain complaining about "The Awful German Language" (1880). In German, articles indicate a gender: masculine (*der*), feminine (*die*), or neuter (*das*). According to Twain, these grammatical genders do not make

Fink, S. B. (2020). Look who's talking! Varieties of ego-dissolution without paradox. *Philosophy and the Mind Sciences, 1*(I), 3. https://doi.org/10.33735/phimisci.2020.I.40

any sense: "In German, a young lady [*das Mädchen*] has no sex, while a turnip [*die Rübe*] has. Think what overwrought reverence that shows for the turnip, and what callous disrespect for the girl." We can imagine someone so confused by this grammatical artefact that they really think that any young lady is genderless and acquires a gender only after puberty, when she turns into a woman, or that turnips are actually all female. This odd fellow has been misled by grammar.

Just like our odd fellow who superimposed grammatical gender onto external reality, it might be that we naturally force our conceptions about proper grammar onto experiences themselves. It is not that we find a self *in experience*; rather, a self is introduced *in the way our language forces us to report on experiences*.

But, as I said before, such a theory fails to explain a reported change: Things appear to be different before and after in comparison to *during* the non-egoic episode. If there never was a self in experience, what is it that happens in such episodes of falsely labelled "ego-dissolution"?

I suggest an ironic solution: The fundamental irony of spiritual enlightenment is that it is a basic and general fact that phenomenal experiences are always selfless but we hardly ever notice this. It takes special circumstances for one to cognitively register that one never felt like anybody. Nothing changes in phenomenality if we attain these states. We just recognise that something has always been lacking: a feeling of self. This realisation that the idea of being a self is only a side-effect of grammar and not something we actually experience would be deeply ironic: Becoming enlightened by attaining a non-egoic state of experience is nothing special or extraordinary – it simply is our basic mode of being.

But one does change, if only doxastically: Before and after the episodes, one has *false* beliefs about the structure of one's experiences; during these episodes, one gains *true* beliefs about the structure of one's experiences. The difference is a difference in our belief system. And because all we claim is that there is no self-reference *in phenomenal consciousness*, we can still hold that there are *unconscious* processes that facilitate self-reference. We can therefore have purely cognitive *de se* beliefs without any phenomenal component indicating a lack of self-reference in phenomenal consciousness.

Despite claiming that there is no feeling of self, the no-ego account still belongs to the phenomenological diagnoses because it makes a claim about phenomenality: There is no feeling of self and there never was. In these episodes, we realise that we confused our way of speaking with our ways of experiencing. This is the revelation one attains in non-egoic experiences.

Revelations can be short-lived or persistent, frightening or enlightening. But they need not come with a phenomenal difference. People may react differently to them: Some may feel as if they have lost something because a welcomed illusion has been revealed as just that – an illusion. Language still has them under a spell, and they may therefore think that something must have been there before.[27]

[27] If our odd fellow fails to find a turnip's ovaries, he may also think that this turnip has lost its gender – even though it never had one to begin with.

Fink, S. B. (2020). Look who's talking! Varieties of ego-dissolution without paradox. *Philosophy and the Mind Sciences*, *1*(I), 3. https://doi.org/10.33735/phimisci.2020.I.40

Others might truly feel the glow of epistemic progress and accept that language is confusing. They either stop speaking or start using a form of language that allows them to avoid a first-person pronoun, like speaking Japanese, or in English using only plural forms, third-person pronouns, or definite descriptions (e.g. "the speaker") to avoid being bewitched again by language. This doxastic difference between before/after and during the episode can have widespread cognitive effects, like most doxastic differences.

Obviously, we can have reports here because this account of non-egoic experiencing does not exclude the possibility of any form of cognitive self-reference. It just claims that there never was a *phenomenal* self, not that there never was any form of self-reference. All it rejects is that forms of self-reference ever show themselves in experience. Non-egoic states aren't anything special.

The no-ego account is obviously highly counterintuitive. For many, it is an obvious fact not just that they *are* someone, but that it *feels* like something to be someone. That this belief is so widespread may merely underline how deep the confusion runs. We need a convincing error theory for why this delusion that there is a feeling of self is so common. Pointing to grammar may get us only so far. But a proponent of the no-ego account may offer the following explanation, which relies on two theses.

First: In any language, we track cognitive differences in speaker and audience. According to a pragmatics-first approach (see e.g. Bar-on, 2013; Moore, 2016, 2017, 2018; Scott-Phillips, 2015), language was developed for communication and coordination, not necessarily for cognition. In communication, we have an embodied speaker and audience. Both are co-represented during an act of communication. However, communication happens in the real world, not just in our minds – and in the real world, speaker and audience differ. It is good communicative practice to keep track of the intentions and beliefs of the different individuals involved in a communicative act. Only if we are aware that someone differs from us in their beliefs may we be able to deceive or teach. In the framework of Gricean communication, tracking audience and speaker separately is in fact basic to any communicative act. And because language is one of the most important tools we have for categorising and capturing the structure of our experiences, we transpose into descriptions of experience this difference between speaker (i.e., us) and audience (the others).

Second, we are by nature naïve realists: In our natural stance towards our experiences, we take ourselves to be in contact with the real world, not with mental representations or phenomenal experiences *standing in* for the real world, unless we reflect on or are made aware of specific illusions. For this reason, we often pick out experiences by what we take them to stand in for in the world. For example, we may say that we *had an experience as of a lamb* on the horizon – although we only experience a hazy fleck of white. Could be a lamb, could be a Hungarian sheepdog (a Komondor), or even a white car. So we picked out this experience based not on how we experienced it, but on what we think it relates to in the world outside our

Fink, S. B. (2020). Look who's talking! Varieties of ego-dissolution without paradox. *Philosophy and the Mind Sciences, 1*(I), 3. https://doi.org/10.33735/phimisci.2020.I.40

minds. In such cases, there is a "crossing over" from what we believe about the world onto how we categorise our experiences themselves. In this process, we may come to believe that there are certain structures in our experiences even though these structures pertain only to the world outside.

Here is an illustrative example adapted from Palmer (1999, p. 209): We know that each of us has only one left and one right hand. And we think that our visual experiences present us with only one left and one right hand when we look at either one. But because of the set-up of our eyes, anything outside of where we focus produces non-aligned images on our retinas. This non-alignment shows itself as "seeing double" outside of where we fixate (including an area called "Panum's fusional area"), which is sometimes interpreted as *depth*. The effect is most striking if we stretch an arm out and focus on our thumb or on the background. If we fixate on the thumb, we see double in the background; if we fixate on the background, we visually experience two thumbs. But this is true for all visual experiences, thanks to the architecture of our visual system. So we do not actually *experience* one left hand if we fixate on the right; we experience two. Yet in forming our beliefs about experience, we superimpose how we think the world is – where we actually only have one left and one right hand – onto what we think our experiences are like.

This forcing-onto-experience might also happen with "fundamental" structures: structures that stem from a basic feature of our dominant form of communication, namely tracking which intentions and beliefs belong to us as speakers and which ones belong to our audience. It might be that there is no felt self in phenomenal consciousness, but merely cognitive tracking of which mental states are under our control and which ones are outside of it. Obviously, today we exploit language for cognition, not merely for communication. Still, tracking which mental states belong to the speaker and which to others remains important. And this tracking of our own mental states as our own could be superimposed *onto* consciousness even though it is not *in* consciousness.

In this view, there is no feeling of self. There is, however, a self superimposed on an experience, necessary for registering this experience as one being had by the speaker, and not necessarily shared by the audience. In most cases, the speaker will fail to notice this superimposition as a purely cognitive add-on and will instead take it to be an accurate apprehension of the experience itself. I am uncertain whether I buy into this sketch of an error theory, but it certainly is a plausible one which a proponent of a non-egoic account might offer. This error theory explains why the ego-delusion – the delusion that there is a felt self – is so widespread: It exploits a mechanism at the core of every Gricean communicator.

5.6 Summary

How should we proceed if we captured a Sane in the wild? If we have reason to believe that (concerning this specific Sane) what is said reflects what is meant, we reject the pragmatic diagnoses. Then, if we have reason to believe that the

Fink, S. B. (2020). Look who's talking! Varieties of ego-dissolution without paradox. *Philosophy and the Mind Sciences, 1*(I), 3. https://doi.org/10.33735/phimisci.2020.I.40

utterer of this SANE is neither obviously deluded, nor irrational, nor deceiving us, we exclude the cognitive diagnoses. That leads us to the phenomenological diagnoses, where four options are available. In each, the speaker uses this SANE to express something concerning her experiences. Of these, only one (total ego-dissolution) leads to the SANE paradox. This leaves three live options.

Must we favour one of these phenomenological diagnoses over the others for all cases? Or should we expect that each may apply to some but not all cases of "non-egoic" ego-alteration? Should we therefore seek *one general* phenomenological diagnosis, or be pluralists who proceed on a case-by-case basis? This is what I focus on in the last section.

6 Many phenomenological diagnoses or only one?

My goals for this article were twofold. First, I have tried to show that there is not only one diagnosis available to us to explain the occurrence of verbalised self-ascriptions of non-egoic experiences (SANE). Pragmatic and cognitive diagnoses are available to us. In these, we explain why someone says something apparently self-defeating without making any commitment to any unusual experiences. Second, I hope to have demonstrated that the SANE paradox relies on controversial claims: the first, called UNIQUE, that there is one and only one mechanism for self-reference; the second, called PHENSELF, that each mechanism facilitating self-reference necessarily manifests itself in phenomenal consciousness; the third, called CEASE, that a feeling of self dissolves in these episodes. But we can reject either one and give explanations of how this is compatible with the occurrence of honest and somewhat veridical SANE: In phenomenality, a self may be modular, there may be a plurality of selves in each person, the self may expand to the point where the self/world distinction loses all meaning, or there never was a phenomenal self – a fact we come to cognitively realise in such episodes. Total ego-dissolution, however, may be a state achieved only in death, not while we continue to cognise.

Now, are these phenomenal hypotheses in opposition? Do some entail others? Are some incompatible? One might think not, and thereby accept that ego-dissolution is a heterogeneous phenomenon. This would raise many methodological questions about how to differentiate these phenomena empirically. But I believe that such pluralism of diagnoses is misguided: We should expect a homogeneous phenomenological diagnosis for all cases.

Why only one account? If there is a feeling of self, it is a highly abstract feature of consciousness. We should not expect massive variance among individuals, just as we do not expect massive inter-individual differences in the way we experience space, time, or unity (cf. Hohwy, 2011; Fink, 2018). Thus, we should not expect more than one account to be true, but we should expect more than one to be false.

Does one entail another? No. All phenomenological diagnoses differ tremendously. Let me illustrate. Is there any feeling self at all phenomenally present

Fink, S. B. (2020). Look who's talking! Varieties of ego-dissolution without paradox. *Philosophy and the Mind Sciences, 1*(I), 3. https://doi.org/10.33735/phimisci.2020.I.40

or not? Here, all theories agree except for the no-ego account. If there is a feeling of self, is it unique and monolithic? Total ego-dissolution and ego-expansion subscribe to this view, which distinguishes them from modularity and plurality accounts. The no-ego account denies the antecedent here. Lastly, if there is a feeling of self, does it dissolve during these episodes? Here, ego-expansion says no, while total ego-dissolution and the multiplicity account both say yes. So none of these theories can be reduced to any of the others. Therefore, at most one can be true. And if we cannot expect inter-individual variation on such a fundamental level, these accounts are actually in competition.

If the phenomenological diagnoses are mutually irreducible, which one should we prefer? Due to the Sane paradox and the problems with *de se* memory formation, I reject total ego-dissolution as a live option. What of its alternatives? So far, the data do not tell, and each hypothesis fails to account for some aspects of certain reports.[28] Authors like Millière (2017) may favour a multiplicity account. But this account is not very parsimonious because it postulates a plenitude of feelings of self. Additionally, it still needs to explain how they interact in order to give rise to a feeling of one unique and undivided self, a self in the sense of an individual. These are reasons why I prefer its alternatives.

What of the other two? I have a place in my heart for the no-ego account, mainly due to its ironic nature: Thousands of hours of sitting and attending and all you get is just an insight about what was there from the start. But, unfortunately, irony is not evidence. We should, however, worry about attributing to ourselves a massive number of false beliefs. Yet, I find the error theory of why we believe that there is a felt self convincing. It relies, however, on the pragmatics-first account of language and on us being naïve realists. Both are controversial theses. If we reject either, we should favour ego-expansion as the underlying process for "ego-dissolution". Here, episodes of an inflated sense of self are misapprehended, leading to a report either about ego-expansion or about ego-dissolution. This is the account I deem most likely.

However we pick our favourites from amongst the phenomenological diagnoses, we have already taken a step forward. We now have several hypotheses that compete; each makes different predictions and can be supported by different kinds of evidence; and we may have a ranking of which has the highest prior probability. As for which of these will explain the incoming data best – time will tell.

[28] This may be explained to some degree by cultural and social biases in reporting.

Fink, S. B. (2020). Look who's talking! Varieties of ego-dissolution without paradox. *Philosophy and the Mind Sciences, 1*(I), 3. https://doi.org/10.33735/phimisci.2020.I.40

Acknowledgments
This article benefitted tremendously from the input of numerous individuals. I am grateful to all of them. Among these are the participants of the workshop on radical disruptions of self-consciousness at Frankfurt in October 2018. Especially, I want to thank Thomas Metzinger and Raphaël Millière for their great and detailed input and criticism, which improved the paper tremendously, and also for their work in making this special issue happen, and their trust and courage to publish with *Philosophy and the Mind Sciences*, making this our inaugural issue. I also wish to thank two anonymous reviewers and three less-than-anonymous commentators (Chiara Caporuscio, Adrian Kind, Fabian Fuchs) for their supportive and helpful comments. Also, I am grateful to my certainly-not-anonymous co-editors-in-chief, Wanja Wiese and Jennifer Windt, for help, support and guts to simply go and run with the idea of a free open-access-journal. Lastly, I thank Emily Troscianko for her great help and input as lector. All remaining mistakes are mine. Part of this research was funded by the DFG-RTG-2386 *Extrospection*.

References

Apel, K.-O. (1976). Das Problem der philosophischen Letztbegründung im Lichte der transzendentalen Sprachpragmatik: Versuch einer Metakritik des "kritischen Rationalismus". In B. Kanitscheider (Ed.), *Sprache und Erkenntnis (Festschrift für Gerhard Frey zum 60. Geburtstag)* (pp. 55–82). Innsbruck: Inst. f. Sprachwissenschaft d. Univ. Innsbruck.

Bar-on, D. (2013). Origins of meaning: Must we "go Gricean"? *Mind & Language*, *28*(3), 342–375. https://doi.org/10.1111/mila.12021

Berkovich-Ohana, A., Dor-Ziderman, Y., Glicksohn, J., & Goldstein, A. (2013). Alterations in the sense of time, space, and body in the mindfulness-trained brain: A neurophenomenologically-guided meg study. *Frontiers in Psychology*, *4*, 912. https://doi.org/10.3389/fpsyg.2013.00912

Bortolotti, L. (2009). *Delusion and other irrational beliefs*. Oxford: Oxford University Press.

Coltheart, M. (1980). Iconic memory and visible persistence. *Perception & Psychophysics*, *27*(3), 183–228. https://doi.org/10.3758/BF03204258

Conway, M. A. (2005). Memory and the self. *Journal of Memory and Language*, *53*(4), 594–628. https://doi.org/10.1016/j.jml.2005.08.005

Conway, M. A., & Pleydell-Pearce, C. W. (2000). The construction of autobiographical memories in the self-memory system. *Psychological Review*, *107*(2), 261–288. https://doi.org/10.1037//0033-295X.107.2.261

Deane, G. (2020). Dissolving the self: Active inference, psychedelics, and ego-dissolution. *Philosophy and the Mind Sciences*, *1*(I), 2. https://doi.org/10.33735/phimisci.2020.I.39

Dor-Ziderman, Y., Berkovich-Ohana, A., Glicksohn, J., & Goldstein, A. (2013). Mindfulness-induced selflessness: A MEG neurophenomenological study. *Frontiers in Human Neuroscience*, *7*(582), 1–17. https://doi.org/10.3389/fnhum.2013.00582

Fink, S. B. (2018). Introspective disputes deflated: The case for phenomenal variation. *Philosophical Studies*, *175*(12), 3165–3194. https://doi.org/10.1007/s11098-017-1000-8

Flanagan, O. (1992). *Consciousness reconsidered*. Cambridge, MA: MIT Press.

Foster, C. (2016). *Being a beast*. London: Profile Books.

Gennaro, R. J. (2008). Are there pure conscious events? In C. Chakrabarti & G. Haist (Eds.), *Revisiting mysticism* (pp. 100–120). Newcastle: Cambridge Scholars Press.

Gennaro R. J. (2020). Cotard syndrome, self-awareness, and I-concepts. *Philosophy and the Mind Sciences*, *1*(I), 4. https://doi.org/10.33735/phimisci.2020.I.41

Goldman, A. I. (1997). Science, publicity, and consciousness. *Philosophy of Science*, *64*(4), 525–545. https://doi.org/10.1086/392570

Grice, H. P. (1991). Logic and conversation. In M. Ezcurdia & R. J. Stainton (Eds.), *Studies in the way of words* (p. 47). Cambridge: Harvard University Press.

Hohwy, J. (2011). Phenomenal variability and introspective reliability. *Mind & Language*, *26*(3), 261–286. https://doi.org/10.1111/j.1468-0017.2011.01418.x

Hume, D. (1739). *A treatise of human nature*. London: John Noon.

Fink, S. B. (2020). Look who's talking! Varieties of ego-dissolution without paradox. *Philosophy and the Mind Sciences*, *1*(I), 3. https://doi.org/10.33735/phimisci.2020.I.40
 https://philosophymindscience.org ISSN: 2699-0369

Hyman Jr, I. E., Husband, T. H., & Billings, F. J. (1995). False memories of childhood experiences. *Applied Cognitive Psychology*, *9*(3), 181–197. https://doi.org/10.1002/acp.2350090302

Josipovic, Z. (2010). Duality and nonduality in meditation research. *Consciousness and Cognition*, *19*(4), 1119–1121. https://doi.org/https://doi.org/10.1016/j.concog.2010.03.016

Klein, S. B., & Loftus, E. F. (2014). The mental representation of trait and autobiographical knowledge about the self. In T. K. Srull & R. S. Wyer Jr (Eds.), *The mental representation of trait and autobiographical knowledge about the self*: *Vol. V* (2nd ed., pp. 1–50). New York: Psychology Press.

Klein, S., & Nichols, S. (2012). Memory and the sense of personal identity. *Mind*, *121*(483), 677–702. https://doi.org/10.1093/mind/fzs080

Letheby, C. (2020). Being for no-one: Psychedelic experience and minimal subjectivity. *Philosophy and the Mind Sciences*, *1*(I), 5. https://doi.org/10.33735/phimisci.2020.I.47

Letheby, C., & Gerrans, P. (2017). Self unbound: Ego dissolution in psychedelic experience. *Neuroscience of Consciousness*, *2017*, 1–11. https://doi.org/10.1093/nc/nix016

Lewis, D. K. (1979). Attitudes de dicto and de se. *The Philosophical Review*, *88*(4), 513–543. https://doi.org/10.2307/2184843

Loftus, E. F., & Pickrell, J. E. (1995). The formation of false memories. *Psychiatric Annals*, *25*(12), 720–725. https://doi.org/10.3928/0048-5713-19951201-07

Metzinger, T. (2004). *Being no one*. Cambridge, MA: MIT Press.

Metzinger, T. (2018). Minimal phenomenal experience. *MindRxiv*. https://doi.org/10.31231/osf.io/5wyg7

Millière, R. (2017). Looking for the self: Phenomenology, neurophysiology and philosophical significance of drug-induced ego dissolution. *Frontiers in Human Neuroscience*, *11*(245), 1–22. https://doi.org/10.3389/fnhum.2017.00245

Millière, R. (2020). The varieties of selflessness. *Philosophy and the Mind Sciences*, *1*(I), 8. https://doi.org/10.33735/phimisci.2020.I.48

Moore, R. (2016). Gricean communication and cognitive development. *The Philosophical Quarterly*, *67*(267), 303–326. https://doi.org/10.1093/pq/pqw049

Moore, R. (2017). Pragmatics-first approaches to the evolution of language. *Psychological Inquiry*, *28*(2/3), 206–210. https://doi.org/10.1080/1047840X.2017.1338097

Moore, R. (2018). Gricean communication, joint action, and the evolution of cooperation. *Topoi*, *37*(2), 329–341. https://doi.org/10.1007/s11245-016-9372-5

Nour, M. M., Evans, L., Nutt, D., & Carhart-Harris, R. L. (2016). Ego-dissolution and psychedelics: Validation of the ego-dissolution inventory (EDI). *Frontiers in Human Neuroscience*, *10*(269), 1–13. https://doi.org/10.3389/fnhum.2016.00269

Palmer, S. E. (1999). *Vision science: Photons to phenomenology*. Cambridge, MA: MIT Press.

Preller, K. H., Burt, J. B., Ji, J. L., Schleifer, C. H., Adkinson, B. D., Stämpfli, P., et al. (2018). Changes in global and thalamic brain connectivity in LSD-induced altered states of consciousness are attributable to the 5-HT2A receptor. *Elife*, *7*(e35082), 1–31. https://doi.org/10.7554/elife.35082

Saks, E. R. (2007). *The center cannot hold: My journey through madness*. New York: Hyperion.

Sartre, J.-P. (1960). *The transcendence of the ego*. New York: Hill; Wang.

Sartre, J.-P. (1992). *La transcendance et l'ego*. Paris: Vrin. (Original work published 1936)

Sass, L., Pienkos, E., Nelson, B., & Medford, N. (2013). Anomalous self-experience in depersonalization and schizophrenia: A comparative investigation. *Consciousness and Cognition*, *22*(2), 430–441. https://doi.org/10.1016/j.concog.2013.01.009

Schwitzgebel, E. (2008). The unreliability of naive introspection. *Philosophical Review*, *117*(2), 245–273. https://doi.org/10.1215/00318108-2007-037

Schwitzgebel, E. (2012). Introspection, what? In D. Smithies & D. Stoljar (Eds.), *Introspection and consciousness* (pp. 29–47). https://doi.org/10.1093/acprof:oso/9780199744794.003.0001

Scott-Phillips, T. C. (2015). Nonhuman primate communication, pragmatics, and the origins of language. *Current Anthropology*, *56*(1), 56–80. https://doi.org/10.1086/679674

Simeon, D., & Abugel, J. (2006). *Feeling unreal: Depersonalization disorder and the loss of the self*. New York: Oxford University Press.

Sorensen, R. A. (2004). *Vagueness and contradiction*. New York: Oxford University Press.

Studerus, E., Gamma, A., & Vollenweider, F. X. (2010). Psychometric evaluation of the altered states of consciousness rating scale (OAV). *PloS One*, *5*(8), 1–19. https://doi.org/10.1371/journal.pone.0012412

Tversky, A. (1969). Intransitivity of preferences. *Psychological Review*, *76*, 31–48. https://doi.org/10.1037/h0026750

Zahavi, D., & Kriegel, U. (2015). For-me-ness: What it is and what it is not. In D. Dahlstrom, A. Elpidorou, & W. Hopp (Eds.), *Philosophy of mind and phenomenology* (pp. 36–53). New York: Routledge.

Fink, S. B. (2020). Look who's talking! Varieties of ego-dissolution without paradox. *Philosophy and the Mind Sciences*, *1*(I), 3. https://doi.org/10.33735/phimisci.2020.I.40

Fink, S. B. (2020). Look who's talking! Varieties of ego-dissolution without paradox. *Philosophy and the Mind Sciences, 1*(I), 3. https://doi.org/10.33735/phimisci.2020.I.40