Chapter 6

ON STRAWSON ON KANTIAN APPERCEPTION: 'SELF-REFLEXIVENESS' AND THE UNITY OF CONSCIOUSNESS

6.1 *Introduction*

P.F. Strawson's take on a core element in Kant's argument, in the Transcendental Deduction of the *Critique of Pure Reason*, for the unity of consciousness as a necessary condition of the possibility of experience in general is a familiar one. Strawson argues in his classic *The Bounds of Sense* (Strawson 1968) that what can be salvaged from Kant's argument amounts to a certain self-reflexiveness between the unity of consciousness as condition of the possibility of experience and the perception of an objective world that is perceived as being external to, and thus distinct from, oneself as the one perceiving. The self-reflexiveness basically consists in the reciprocal relation between perceiving and perceived. While there must be someone doing the perceiving, the perceiving points to something or some *thing* that there must be to be perceived. More in particular, the unity of consciousness, which constrains one's *subjectively* perceiving something to be the case, must be seen as necessarily dependent upon what is *objectively* perceived, namely a unitary world of particulars in and through which the subject follows a path of experience. The self-reflexiveness at issue is in essence tantamount to a conceptual connection between the unity of the subject and the unity of the object; indeed, as Strawson (1968:96) observes, it provides the 'direct analytical connexion' that one wishes to see in Kant's argument. The argument for this connection is generally presented as a 'transcendental argument'.[1]

These insights, which Strawson believes lie hidden behind Kant's sometimes arcane reasoning and should be able to be extracted from it by virtue of scaling away the speculative excess, are crucial. They cast light on the exigencies of a philosophically

[1] Notice that Kant does not talk about transcendental arguments as such, although like Strawson I believe that Kant does have an argument that proceeds from the premise of self-consciousness to objectivity. I also agree with Strawson that the Transcendental Deduction is not merely an explanation of objective experience, but a proof too (cf. Strawson 1968:88). I scrutinise Kant's own argument in Schulting (2018b). I shall use the designation 'transcendental argument' in a broadly Strawsonian sense, as referring to Kant's own argument.

meaningful construal of the systemic thrust of Kant's argument in the Deduction. However, I am not engaging here with the standard debate on the nature of transcendental arguments and their wider philosophical relevance. Nor do I reflect on the broader philosophical implications of Kant's arguments (or Strawson's take on them). Nor am I interested in Strawson's arguments per se, *outside* the context of Kant interpretation; I do therefore not examine Strawson's reasoning to see how it measures up against current philosophical insights, while there is no doubt that it has been influential beyond Kant interpretation, especially so in the Oxford school of thought. My reason for considering Strawson's construal lies in the fact that it singularly directs our attention to an element of Kant's argument that is universally misunderstood—including, I shall argue, by Strawson himself—as a result of which, I contend, a proper understanding of Kant's main claim in the Deduction has not been possible up until now. This element concerns the precise meaning or sense of the *unity* of consciousness, and hence the meaning of the premise of the so-called transcendental argument. My interest is therefore mainly an interpretative one, which though is crucial to the *philosophical* understanding of the central Kantian argument. So my criticism cannot be dismissed as of *merely* interpretative value. Failure to get the interpretation of Kant's argument for the unity of consciousness right results in the futility of whichever *philosophical* argument that takes specifically and explicitly Kant's argument as its point of departure.

I fully concur with Strawson's line of reasoning that an analytical connection and hence a reflexivity of sorts exists between, on the one hand, the subject of experience (and accordingly the premise of the putative transcendental argument) and, on the other, that of which one has experience, that is, objects in the spatiotemporal world (and accordingly, the conclusion of that argument). (Note that the possibility or factuality of *experience* as such has no central role in the argument; what constitutes the reflexivity is rather the *connection between* the *subject* of experience and *the objective* as that of which she possesses a concept, given the possibility or factuality of experience.)

However, I take issue with Strawson's reconstruction of this reflexivity. My claim is that Strawson's reconstruction of Kant's argument suffers from a modal fallacy concerning the premise of the argument. Further, I believe that the fallacy in Strawson's reasoning jeopardises the viability of constructing the kind of argument that Strawson wishes to extract from Kant's reasoning, and which is supposed to show the analytical connection. I cannot of course, in the space of a chapter, attend to every aspect of the problem. I shall therefore concentrate on Strawson's reconstruction of the *premise* of the argument

concerning the analytical connection between subject and object. Much, if not all, depends on the way the premise is construed.

In Section 6.2, I survey Strawson's reconstruction of the transcendental argument. In Section 6.3, I examine Strawson's reading of the analyticity of the principle of apperception or self-ascription and note some essential differences with what I take to be Kant's own position. I subsequently point out that Strawson crucially fails to distinguish between two distinct ways of conceiving of the unity of consciousness of an experiencing subject (Section 6.4); and that as a result Strawson does not heed the difference between two kinds of claim regarding the *necessary* unity or unifiability of representations (Section 6.5). I thus address three closely related aspects of apperception: analyticity, unity, and necessity. In the course of pointing out the aforementioned failure regarding modality, I indicate that it affects the constructing of a transcendental argument in the way that Strawson does. I shall also suggest that a priori synthesis, which is firmly disallowed by Strawson's argument for analyticity, avoids the problem regarding modality. (Note that I cannot give a full account of a priori synthesis here.)[2]

6.2 *Strawson's Core Argument*

Let me first review Strawson's core argument. In outline, Strawson's reconstruction of the 'direct analytical connexion', which is laid bare by the so-called transcendental argument (T), goes thus:

> T1. Any experience is necessarily unifiable in a unity of consciousness, by virtue of the possibility of self-ascription, the condition of which is transcendental apperception, so as to show self-sameness of representation or a belonging to a single consciousness, viz. an analytic unity of consciousness. (cf. Strawson 1968:98)

> T2. Unity of consciousness requires another unity, viz., an *objective* unity, the connectedness of which presupposes the employment of concepts of the objective, these concepts being the rules that govern the connection of experiences so as to allow a differentiation of the objective from the subjective and hence to enable re-identification of one's variant representations. (cf. Strawson 1968:87)

[2] For this, see Schulting (2018b).

Note that T is Strawson's *reconstruction* of Kant's argument; it is not necessarily Kant's argument. In fact, vital steps in Kant's argument are lacking in T. T is expounded in Section 7 of Part 2.II of Strawson's *The Bounds of Sense* (Strawson 1968:97ff.). Secondly, that the argument concerns a 'direct analytical connexion' means that the relation between the premises of T is one of deductive inference. Unlike a standard logical inference or syllogism, however, T's premises are not independent or independently had but are, in some way, necessarily related. T is a prototypical case of *a priori* conceptual analysis. Essentially, what T says is that it is actual knowledge of the objective world, or in effect the concepts employed for such knowledge, that furnish the material condition of the possibility of consciousness of 'my diverse experiences as one and all my own' (Strawson 1968:107); that is to say, the unity of my consciousness derives ultimately from the unity of the world. It is in this way that my experience must be seen to depend on that which I experience, or put differently, what I experience enables my very experiencing. This concerns what Strawson calls the conceptualisability of experience, in that any experience gives rise to the thought of experience (see Strawson 1968:107). The *thinkability* of experience consists in the ability to differentiate the experience, as an objective event, from the experiencing. The relation of dependency between the experiencing and its object thus manifests that there obtains a self-reflexiveness of experience. This, then, yields the following conclusions:

T3. The self-reflexiveness of experience, viz. the differentiability between experience and that of which the experience is, shows that there exists an 'analytical connexion' between experience and objectivity. (from T1 and T2)

T4. Objectivity and the knowledge thereof are shown to be the necessary condition or the enabling condition of experience in general. (from T3)

If Kant's deductive argument is reconstructed in this way and T is a fair take on Strawson's transcendental argument, it seems clear that there is no need for an *a priori* synthesis of one's identical representations so as to first ground rationally the concept of objective experience, or indeed the concept of unitary consciousness. What suffices is to demonstrate by means of inferential reasoning the incontrovertible truth of a unified objective world as the basic enabling condition of experience in general; and to show that we thus have knowledge, in virtue of the ability to form epistemically warranted judgements, of what Strawson calls 'objects in [the] weighty sense' (1968:73). At any rate, the concepts of the

objective are not to be associated with the categories, for according to Strawson (1968:87, 88) there is no intrinsic link between the epistemology of experience and the forms of logic. Furthermore, Strawson's reading, which rejects *a priori* synthesis, is fully coherent with naturalism about the objective nexus of appearances, the existence of which must be presupposed so as to allow the a posteriori synthesis of representations in conscious experience (cf. Guyer 1980:205, 208–9, 1987:142; Hossenfelder 1978:100–2). Transcendental apperception, which is the condition of the self-ascription of representations and grounds unitary consciousness, is thus seen merely as a necessary conceptual tool in the explanation of natural phenomena and their connections and a subject's experience thereof.

There are all sorts of highly interesting elements implicit in the argument's chain from T1 to T4 that on the face of it are problematic and therefore require clarification. I want to concentrate on a problem concerning premise T1. This regards a modal fallacy. The fallacy issues from a mix-up between two kinds of conceiving of the unity of consciousness, as well as from a certain (mistaken) understanding of analyticity that Strawson reads into Kant's own position on the analyticity of the principle of apperception (B135 [AA 3:110]). The mix-up I am referring to is evidently not a conflation of the two unities that Strawson clearly distinguishes, namely the *analytic* unity of consciousness and an *objective* unity, although in some way the latter distinction is related to the mix-up that worries me. Let me first turn to the aspect of analyticity. In Section 6.4, I shall then come to speak of unity, after which, in Section 6.5, the modality of the apperception principle will be addressed. But first analyticity.

6.3 *Analyticity*

I want to argue that Strawson's unity argument (T1) suffers from a modal fallacy. I believe this is partly due to a particular conception of analyticity, which wants to dispense with *a priori* synthesis. The premise of the transcendental argument that I outlined above, and which concerns the unity argument (T1), is presented as amounting to a strictly analytic principle. This is the principle of the self-ascription or apperception of representations. I shall argue that the way Strawson construes this principle is problematic. The intimate link presumed to exist between self-consciousness and objectivity (T2) is, as a result, vulnerable to the shortcomings of this particular construal of the analyticity of the principle of self-ascription or apperception. However, whatever the case may be regarding the inferences drawn from T1, let us consider the principle of self-ascription in more detail, as a run-up to

an assessment of the modal fallacy in Strawson's unity argument further below (in Section 6.4).

Strawson apparently takes Kant's premise, to wit the familiar proposition 'The: I think *must be able* to accompany all my representations' (B132 [AA 3:108.19]), with which Kant launches his operative argument in the B-Deduction, to express the analytic principle PS:

> **PS=Principle of Self-ascription**: $(\exists z)(z$ is a thinker at time $t) \longrightarrow \{(\forall x)(x$ is a representation $\wedge$ x is being represented at $t) \longrightarrow \Box\ [(\exists y)(y$ is a thinker 'I' $\wedge$ x is self-ascribed or at least self-ascribable under certain conceptual constraints by $y)]\}$

PS defines self-ascription or apperception, which explains the possibility of subjective experience or self-consciousness. It says that, for all representations, if there is a thinker 'I' at t and there is a representation that is represented at t, then that representation must either be self-ascribed or at least be self-ascribable by 'I' (setting aside issues of self-sameness or identity). There are grounds for thus formulating apperception in the text of Kant's Deduction.[3] For Kant himself states regarding the possibility of self-ascription of representations:

> This last proposition [viz. that unity of consciousness is only possible through synthesis, D.S.][4] is […] itself analytic […] for it says nothing more than that all my representations in any given intuition must stand under the condition, under which alone I can ascribe [*rechnen zu*] them as my representations to the identical self and thus grasp them together [*zusammenfassen*] as synthetically combined in one apperception through the general expression I think. (B138 [AA 3:112.13–19]; trans. modified)

---

[3] I disregard the putative differences between the account of apperception in the A- and B-Deductions. For an extensive account of apperception, see Schulting (2018b); see also Chapters 4, 5, and 7, this volume.

[4] Not surprisingly, this reference to an earlier section in the text is ignored by commentators who stress the analyticity of the self-ascription principle. Strawsonians, who insist on the analytic nature of the principle, namely fail to assess Kant's assertion, earlier at B135, that 'this principle of the necessary unity of apperception is, to be sure, itself identical, thus an analytical proposition, yet it *explains* [*erklärt*] as necessary a synthesis of the manifold given in an intuition, without which that thoroughgoing identity of self-consciousness could not be thought' (B135 [AA 3:109.19–23]; emphasis added, translation emended. A good example of the difficulties that the Strawsonian runs up against in regard to Kant's analyticity claim as a direct result of this failure of assessment, is provided by Cassam's reflections on this score (Cassam 1987:375ff.).

PS is routinely taken to explain a de dicto necessity: I *must* be able to ascribe to myself any representation that I have, I being the subject of any of a series of representations that I ascribe to myself, for which certain conceptual conditions for unification should be met (cf. A122 [AA 4:90.18–20]). By implication, *de facto* self-ascription establishes, a posteriori, a synthetic existential unity of all representations so ascribed as belonging to the unity of consciousness of the self who so self-ascribes; however, nothing in the way of a priori synthesis seems thereby required.[5]

Presumably, the analyticity of PS has to do with self-ascription being criterionless, that is, immune to error through misidentification: one knows and cannot fail to know the conditions under which one ascribes one's representations to oneself (cf. Strawson 1968:92, 93, 98, 165). Contrary to the application of concepts to objects, with respect to one's own representations ostensibly no identificatory criteria are required that first enable their self-ascription and no possibility of error exists: the concept of self applies to one entity and one entity only, namely the self that I am when ascribing my representations to myself. More precisely, the extension of the concept of self consists of just one possible particular instance at any one time at which the concept is instantiated by some self who is self-consciously aware of her own representations. The concept of self or the 'I', if it can be called a concept,[6] is as such a universal representation but at the same time applicable *only* to the one particular individual which it is instantiated by. Further, in any case of representing, I a fortiori know, by way of self-ascribing any representation that I have, that I am the one representing. As it appears, the analyticity of PS would thus concern the logically trivial truth that the ability to conceive of one's representations as one's own is reciprocal with the capacity to employ the indexical 'I' in all cases of such conceiving. That is to say, there is an analytic, conceptual relation between a representation and the agent of representation, which is the self-ascribing representer, or the thinking 'I'. Paul Guyer (1980:209) puts it

[5] According to Guyer (1980:212), the notion of apperception, as a consciousness of one's self-consciousness, can only be retained once freed from the 'encumbrance of a priori synthesis'. In other words, on the analytic reading of the principle expounded by Kant at B138, any synthetic unity that would be involved could only be a posteriori, viz. a unity of all actually ascribed representations, not of all possible (i.e. past, present and future) representations (cf. Strawson 1968:96). At any rate, this is gainsaid by Kant's view, expressed in the omitted passage in the above quotation from B138, that the synthetic unity at issue is an *a priori* unity that is the very ground of any analytic unity of consciousness (see B135; cf. B134 [AA 3:110.9–11], where Kant speaks of 'synthetic unity of the manifold of intuitions, *as given a priori*' being 'the ground of the identity of apperception itself').

[6] Kant calls the 'I think' a concept at A341/B399 and elsewhere. But at A345–6/B404 Kant says that of the 'wholly empty representation *I* [...] one cannot even say that it is a concept'.

quite explicitly by contending that Kant holds that '[w]hatever is to count as a representation at all must be fit for self-ascription'.[7] Guyer continues: '[The 'I think'-proposition] asserts […] that I cannot have a representation which is not subject to these conditions [i.e. the conditions for self-ascription, D.S.]. To put it bluntly, Kant asserts that I cannot have a representation which I cannot recognize as my own.'[8]

On this account, it seems that not only the conditions for representing (that is, for having representations) and the logical conditions for self-ascription of representations are conflated, but also the conditions for representing and those for self-consciousness given that, as Strawson (1968:108) asserts, transcendental self-consciousness is the a priori form or condition of self-ascription. A representer could thus not be otherwise than an at least potentially self-conscious representer. I believe that this view of the analyticity of the apperception principle is flawed, for it provides no ground for assuming that any agent of representation is *eo ipso*, even if only potentially, a *self-conscious* subject or that the subject who envisages her own future states of affairs has complete knowledge of future states of affairs as involving herself.[9]

Notice that, as regards T1, Strawson (1968:92) claims that experiences are necessarily unifiable in that they must satisfy the conditions of belonging to a single consciousness. This is a rather different claim from the one regarding the criterionless nature of self-ascription. It seems, then, that Strawson confuses two different arguments: one concerning the logical conditions governing the self-ascribability of one's *own* representations and another for the necessary unifiability of representations *simpliciter*, only the former of which would prima facie amount to the tautology, or the analytical connection, that Strawson considers to be the nub of Kant's argument. This confusion relates to a modal confusion of which Strawson is guilty with respect to premise T1, namely with respect to the sense in which one should understand the unity of consciousness to which one ascribes one's representations. This is what I shall argue in Sections 6.4 and 6.5. But let me first return to Strawson's understanding of the analyticity of self-ascription.

To illustrate the austere conception of apperception as a condition for representation in terms of PS, which dispenses with a priori synthesis, consider Malte Hossenfelder's

---

[7] Strawson speaks similarly of any representation's potentiality for self-ascription (Strawson 1968:101, 114, 117).

[8] Notice that Guyer is critical of this view that he attributes to Kant.

[9] Not all future representative states need be ones that I self-consciously represent, lest the conceptual condition for self-ascription be seen as concerning an ontological necessity, implying a necessary coexistence of representer and self-consciousness (cf. Ameriks 2000b:249).

interpretation of the 'I think'-proposition (Hossenfelder 1978). Having noted that there are ostensible intrinsic problems with Kant's appeal to a priori synthesis and assuming that the principle of self-consciousness is a tautological principle as previously defined (PS), Hossenfelder (1978:100–1) attempts to cast light on the analyticity of the principle by suggesting that we substitute 'to represent' for the verb 'to think' in Kant's proposition 'The: I think must be able to etc.'. The proposition would then read:

> The: I represent must be able to accompany all my representations.

Only in this way, Hossenfelder argues, can the analytic character of PS become explicit, for quite clearly its denial logically entails a contradiction. A representation is always represented by a representer, who, at least according to Kant's principle of apperception, *ex hypothesi* ascribes the represented representation to herself (presumably the 'I') as the representer. It is a trivial conceptual truth that any representation requires a representer. The premise of Kant's argument is then tantamount to nothing more than the unpacking of what is already contained in the concept of 'representation'. Hossenfelder thus reduces apperception to a conceptual principle of representation *tout court*. We can translate Hossenfelder's substitution reading of PS as:

> **PS′**: $(\exists z)(z$ is a representer at time $t) \longrightarrow \{(\forall x)(x$ is a representation $\wedge$ x is being represented at $t) \longrightarrow \Box\ [(\exists y)(y$ is a representer 'I' $\wedge$ x is self-ascribed or at least self-ascribable by $y)]\}$

However, Kant himself does not regard the principle of self-consciousness as simply a principle of *representation*, so that the analytic (conceptual) relation obtains between represented and representer.[10] But even if disregarding this historical point, it is not true to say that said conceptual connection is *eo ipso* substitutable, in all possible cases, for the relation between a representation and a self who self-ascribes her representations to her identical self. First, a representer could just be representing without self-ascribing

[10] Notice that Kant's phrase at B131 (AA 3:108.20–1) continues: '…for otherwise something would be represented in me that could not be thought at all', which would make no sense on Hossenfelder's substitution proposal. Kant's suggestion is that in case the 'I think' would not accompany my representations something would still be represented (in me) but I would not think it, which is trivially true, but not in Hossenfelder's sense.

representations at all—this would amount to first-order representing without a second-order representing of one's representing by virtue of the self-ascription of representations to one's identical self (or to a perceiving without apperceiving).[11] More intriguingly, a representer could be representing representations, or indeed ascribing representations to herself (through a self-reference of sorts), without however thereby self-ascribing them to a self in the strict sense, by which I mean the *same* self (*de re*) to which she also ascribes other representations (over time, involving diachronic identity). It is possible even that a representer could effectively (*de re*) ascribe representations to 'others' when in fact she *believes* that she is ascribing them to herself (*de dicto*) (cf. e.g. A363 [AA 4:228.32–229.04] and A363–4n. [AA 4:229]). (This involves problems concerning the metaphysical status of the identical self to which one ascribes representations, which I must set aside for present purposes.)[12]

If we look at the cases of satisfaction of Kant's apperception principle, then we learn by analysis that apperception cannot be a condition of representing *tout court* (as on PS′). It is not at all the case (i) that all possible representations are necessarily accompanied by an 'I think', nor (ii) that all representations necessarily entail the (transcendental) unity of apperception; and nor (iii) do all of them effectively belong, necessarily, to the thoroughgoing identity of my self-consciousness (in the possessive[13] sense). This can be demonstrated in a breakdown of the 'I think'-proposition into its logical modalities.

Assume necessary possibility P1: *de facto*,[14] the 'I think' accompanies all my representations. If P1, then, *ex hypothesi*, it must also be possible that:

P2: the 'I think' does not accompany all my representations

and/or:

---

[11] Cf. Anth §5, AA 7:135.

[12] I believe that the substitution by some of *de se* modality for the distinction of *de dicto*/*de re* in the case of self-consciousness glosses over the problems involved in attempts to determine the ontological status of the self underlying apperceptive self-consciousness and is therefore wholly stipulative. See in general the authoritative account of Kant's metaphysics of the self in Ameriks (2000a).

[13] I borrow this way of putting it from Ameriks (2000b:281). On the aspect of possession of one's representations, see Chapter 7.

[14] Notice that this adverbial phrase indicates that here an analysis, *ad oculos reflexionis*, of the possible cases of satisfaction of the 'I think' proposition in terms of its logical purport is concerned (cf. Deppermann 2001:130); it is not suggested that an actual occurrence of empirical consciousness is at issue at this point in the argument (cf. Reich 1992:27), even though Kant says elsewhere that the proposition itself is empirical (B420 [AA 3:274:15–20]). See further the Appendix to Chapter 5.

> P3: the 'I think' does not accompany any other representations that happen to occur and are so occurrent in the mind at any time *t* at which the 'I think' is not instantiated

and/or:

> P4: the 'I think' does not accompany any other representations that happen to occur and are so occurrent in the mind at any time *t* at which the 'I think' is not instantiated, and which are also interminably barred from being able to be so accompanied, i.e. such representations that evanesce immediately after having been prompted and leave no significant traces for possible retention and 'taking up' by an act of apperception (some representations may simply not be able to be retained or retrieved).

P2 is obviously spurious, for it is logically inconsistent for me, as the subject of thought, to assert that 'I' am thinking (*de facto*)—or to assent, whilst thinking, to the proposition 'I am thinking'—and yet *not* to accompany my representations that I am thereby thinking. The possessive pronoun 'my' in the predicate 'all my representations' refers rigidly. Those representations are my representations that I accompany as such by effectively thinking them.

By contrast, P1 is analytically true; it expresses quintessentially the principle of identity, which is the first principle of discursive reason.[15] The totality of my representations that are occurrent share the same common mark 'I think' just in case I am accompanying them (as my representations 'all together' [*insgesamt*], as Kant puts it [cf. B132]), by means of the act of thinking, precisely when I am in the business of thinking (representing in a particular way).[16]

---

[15] Cf. UD, AA 2:294. Cf. also B408 (AA 3:268.7–9).

[16] Whilst it would seem that I can think only one thought at a time, the nature of discursive thought, according to Kant, is such that every singular thought, which is accompanied by an 'I think', consists of several representations taken together and thus thought simultaneously, under one common denominator (the 'I think'), *as* same, viz. as 'all my representations' in terms of a compound thought; unity always implies multiplicity, which in turn entails synthesis to the extent that one's various representations are identical or equal, namely related to the identical 'I think'. Kant makes this clear in the course of §§ 15 and 16 of the B-Deduction (for more discussion, see Schulting 2018b).

P3 reflects the case of a representer R representing any arbitrary occurrent representation *x*, *y*, or *z*. Whilst in this case P1 is not satisfied, R would nonetheless be the representer of *x*, *y*, *z*, even if not aware of herself as in the business of representing and a fortiori being self-aware of doing so. R does not accompany her representations in the transcendental way, but merely in the empirical way by just having them in any arbitrary array peculiar to her actual physio-psychological stance at a particular time. Strictly speaking, R does not *think*. Although Kant does not explicitly, at least not in the *Critique*,[17] venture an opinion on the possibilities P3 and P4, of which it is further open to question if they are anything more than merely formally distinguishable, these are surely logically inferable from the 'I think' proposition. This is confirmed by some of Kant's assertions in the text of the Deduction. P3/4-representations are representations, which, as Kant puts it, are 'nothing for me' (B132), which is consistent with the rigid reference of the indexical 'my' of P1-representations.

The determiner 'all' in the predicate 'all my representations' creates an ambiguity, for Kant's proposition could, superficially, be construed such that it posits that the 'I think' does not effectively accompany *all*, but only *some* of my representations, which could lead one to presume that P2 is not strictly speaking false. This is indeed the route that most interpreters take. Elsewhere (Schulting 2018b, ch. 9), I have argued that this view is mistaken and runs into exegetical difficulties. At the systemic level, in any case, (i) it is logically nonsensical to assert, from a first-person perspective, that whilst I am thinking, I am thinking only *some* of *my* representations that are occurrently represented; (ii) of representations that are not occurrently represented I cannot tell whether they could be mine unless they are effectively represented by me, that is, accompanied by the 'I think', and so, by implication, unaccompanied representations are not strictly speaking *my* representations. This excludes readings of the apperception principle which hold that representations are at any rate potentially subject to transcendental apperception, as a great many commentators believe. In Schulting (2018b), I also explain that the predicate 'all my representations' is a single complex representation, which as such, and only as such, is accompanied (*de facto*) by the 'I think'. I call such a representation r_{all} as opposed to an r_{each} (2018b:242).

There are many more intriguing sides to Kant's principle of transcendental apperception that call for further analysis and exegetical backup. I provide these elsewhere

[17] In his *Anthropology*, Kant provides ample concrete examples of P3/P4-representations. See Chapter 4, this volume.

(Schulting 2018b; see also Chapters 4, 5 and 7 in this volume). Here I want to return to the particular problem that I set out to address, namely the modal fallacy that I claim issues from taking apperception as an analytic proposition in the terms proposed by Hossenfelder (PS′), which are implicitly endorsed by Strawson. This fallacy can be brought to light by further focusing on two interconnected features of apperception: first, the kind of *unity* of consciousness that is established by the self-ascription of representations and, secondly, the *modality* involved in making a claim regarding this unity. Only a particular modal claim is compatible with a more strictly defined principle of apperception (which I introduce below), which is in accordance with the breakdown of Kant's 'I think' proposition provided above.

6.4 *Unity*

As regards *unity*, Strawson fails to notice that one can take Kant's argument for the unity of consciousness, which is established by the act of apperception, in two ways, only one of which is correct. One can take it either (i) as an argument for the psychological or existential unity of singular representative states $r_{each}1...r_{each}n$ in terms of mental states had by a representer and which as such are aggregated in any arbitrary sequence in conformity with the way they are prompted to occur, through psycho-physiological patterns or brain states as their proximate causes, by external objects (viz. a unity of representations in a possessive or *de re* sense); or one can take it (ii) as an argument for the unity of representative states in terms of certain states *recognised* and *identified by the representer herself as* together constituting a unitary compound of representations r_{all} that belongs to the representer as her own (a unity in the *de dicto* sense), whereby it should be noted that the representer here is an epistemic agent (a thinker) and not just a representer. It is the latter kind of unity (ii) that, I contend, Kant is in fact arguing for. The difference between these two kinds of construing the unity of consciousness amounts to the difference between arguing for (whereby R stands for representation and UC for unity of consciousness)

> **UC1**: For all R, R is united in UC, given certain conceptual constraints that have to do with the capacity for self-ascription and material constraints connected with the way the world is

and arguing for

UC2: For all R, *if* R is self-ascribed and recognised by a self-ascribing representer (i.e. a thinker 'I') as belonging to her, *then* R is united in UC (regardless of material or psychological[18] constraints)

Not heeding the distinction between UC1 and UC2 is tantamount to committing a modal fallacy with respect to the (unitary) relation between representations and the principle of self-ascription (PS). This needs to be made explicit (see Section 6.5 below).

Let me first take a closer look at how distinguishing between UC1 and UC2 affects the Strawsonian construal of apperception in terms of PS (or PS′). The recognition alluded to in UC2 is of course not a case of actively reflecting on the part of a psychological subject on her mental states (by way of muttering to herself, as it were). Instead, it points to a function[19] performed by the occurrently representing self in that by being self-consciously aware of her identity as the performer of this function, that is, as an epistemic agent rather than in the modality of being primitively aware of one's conceptually indistinct environment, she knows the conditions under which, rather than having merely subjective validity, her representations acquire objective reality and thus become cognitively or epistemically relevant. This cognitive or epistemic relevance is in the first instance just the objective purport that representations that are self-ascribed by the 'I' have *for* that 'I' (see Chapter 7). Identity of self-consciousness is a rule for or function of recognising that one's representations belong together in a particular unitary form, namely in what Kant calls the transcendental unity of consciousness or the original-synthetic unity of apperception.[20] Kant calls this function a priori synthesis. But this conception of self-consciousness, as including *a*

[18] Strawson of course would equally deny that the argument for the unity of consciousness has anything to do with *psychological* constraints per se; the constraints of self-consciousness at issue are rather conceptual. However, I believe that, given his reading of the premise in terms of PS, the conceptual constraints that Strawson wants to argue for in effect are the necessary, if not yet sufficient, conditions of *empirical* consciousness *simpliciter*, and therefore psychological.

[19] Cf. A108 (AA 4:82.12), where Kant speaks of the 'identity of [a] function'; a few lines further down (AA 4:82.21) Kant, similarly, speaks of the 'identity of [the mind's] action'. Notice that by 'function' Kant understands 'the unity of the action of ordering different representations under a common one' (B93 [AA 3:85.18–19]), which is precisely what is meant by 'the synthesis of recognition in the concept', as the heading of the section, in which the phrase 'the identity of the function' occurs, reads. For more discussion on the role of recognition, see Schulting (2017), ch. 6.

[20] Kant identifies the transcendental unity of consciousness as an objective unity in contrast to a subjective unity at B139 (AA 3:113). Strawson's argument for the objective unity as that on which the unity of consciousness is transcendentally dependent, comes close to Kant's talk of transcendental unity in intent, but not in execution, for contrary to Strawson, who confuses transcendental apperception with the capacity for subjective consciousness *simpliciter* and differentiates it from an objective connectedness, Kant's objective unity is none other than the principle of transcendental apperception itself.

priori synthesis, does not comport well with the principle of self-ascription as defined above, namely PS (let alone PS′), which stipulated that any representation whatsoever is subject to self-ascription. We must now redefine PS as:

> **PS′′**: $(\forall x)(\forall e)(x$ is a representation $\wedge$ x is being represented at time $t) \wedge (e$ is an epistemic agent$) \longrightarrow \Box \{[(\exists y)(y$ is an identical thinker 'I'$) \wedge (y{=}e) \wedge (x$ is self-ascribed by $y)] \longleftrightarrow [(\exists z)(z$ is an analytic unity of all representations recognised and retained by e after $t) \wedge (e$ recognises x to belong to $z)]\}$

PS′′ is a better translation of the earlier quoted passage at B138 than PS, for it takes into consideration Kant's explicit stipulation (in particular in the lead-up to B138) that certain a priori conditions, namely, a priori rules for recognition that together amount to a priori synthesis (i.e. the categories),[21] must be satisfied in order for self-ascription to an identical self first to be possible. This explains why, contrary to what Strawson believes, a priori synthesis must be seen as closely linked up with there being an analytic unity of representations at all.[22] PS′′ is effectively tantamount to a biconditional, for not only is self-ascription conditional on the recognition of a unity of representations (z), but z is also only possible under the condition of self-ascription. In fact, self-ascription is nothing but the constitution of z through the act of recognition. There is thus an analytical unity of representations z and an 'I' thinking it *if and only if* 'I' effectively self-ascribe all my representations in accordance with the a priori rules for recognition (i.e. a priori synthesis). It is this biconditional relation between the self and her representations that she self-ascribes (by virtue of recognising their same- or oneness) that determines the analyticity of PS′′, for which the condition of a priori recognition by means of a rule for unification must thus first be met, to wit a priori synthesis. (This latter requirement, which first makes the principle of self-ascription analytic, would appear to indicate that self-ascription is not criterionless, as on PS.) This implies that one is not licensed to argue that for all representations that are had by a representer it necessarily holds that they are ascribable to the *same* representer, nor a fortiori that all (possible) representations *eo ipso* belong to an analytical unity of consciousness of representations that are recognised by an epistemic

[21] For an extensive account of the categories as the combined set of rules of a priori synthesis, see Schulting (2018b).

[22] How this works precisely, in all its Kantian technicalities, is a topic on which I have elaborated elsewhere (see Schulting 2018b).

agent to belong together in accordance with a priori rules.[23] For a representer is not always an epistemic agent or indeed a thinker.

So how, then, can Strawson vouch for the metaphysically intemperate claim that on his reading *all* representations that one has (or potentially has) are self-ascribable, and that they thus make up an analytic unity of consciousness (as on UC1)? Clearly, there is no analytic relation between *all* representations had (or potentially had) and the condition of self-ascribability by a self-same self; so PS is not really analytic, as Strawson and Hossenfelder would have us believe (hence Hossenfelder's substitution proposal PS′ in an attempt to make the principle's analyticity more plainly visible). Strawson disregards the conditional necessity underlying the apperception argument (as observed by PS′′), which first establishes sameness or identity of self-consciousness with respect to one's representations, from which, in a second step, objective connectivity is analytically derived. This fallacy regarding modality can be made more formally explicit in terms of a failure to distinguish between two kinds of modal claim about the unity of representations established by self-ascription. This is the topic of the next section.

6.5 *Necessity*

Let me summarise. I have argued that Kant's premise is such that it cannot be about the trivial truth that every representation requires a representer—a truth that would indeed be conceptual. If the analyticity of the principle of apperception were to concern a merely conceptual truth (as on PS and PS′), the considerable attention given it in the literature over the course of two centuries would be undeserved. Instead, the premise is about what is required for a representation to be part of a unitary representation that has an objective validity, i.e. that is 'something to me' as the identical subject of all my representations—rather than something that just exists, just is a mental occurrence and therefore has merely subjective value to the occurrent representer. Objective validity here just denotes the satisfaction of the conditions for sameness or identity for a set of representations that belong together insofar as I self-ascribe them. This concerns the fact that a representation has the quality of objective validity if and only if it shares the same mark as is shared by other representations that belong to the unity of self-consciousness. And it shares this mark when I actually, self-consciously self-ascribe and unify it with all other representations that I concurrently self-ascribe.

[23] In the A-Deduction, Kant might seem to endorse precisely this reading (see e.g. A113, A116, A117n.). I confront this ostensible discrepancy in Schulting (2018b).

Strawson suggests that the transcendental argument starts off from a trivial truth, which consists in the necessary unifiability of one's representations *simpliciter* (T1), leading to a conception of objectivity as their unity's condition of possibility (T2). This is problematic. If it is true that all representations that one has are necessarily unified, or at least unifiable,[24] in and by a single subject of representation (oneself), then it is not clear why such a claim would logically require—as Strawson thinks in virtue of the inferential force of the putative transcendental argument—a concept of the objective as a means of distinguishing between one's representing and the object of one's representation (the 'thinkability of experience' requirement, adumbrated earlier in Section 6.2). Furthermore, a concept of the objective is a representation no less than any other representation, so what difference does its being invoked as enabling condition for the unity of consciousness make regarding the differentiation requirement, that is, the differentiation of the objective from the subjective?

In other words, why, as per Strawson's reasoning, must objectivity or the concept of the objective figure as the ground of the unity of consciousness if it is the case that, on account of PS, *all* of one's representations are united as a matter of course? It is far from clear on which grounds Strawson believes the unity argument to rest on the conceptualisability criterion, for conformably to construal PS of the premise of the transcendental argument Strawson in fact just posits UC1, suggesting no further condition under which the subclass of *concepts or representations of the objective* is capable of being differentiated from the broad class of *all representations*. Strawson stipulates that the former are necessary to satisfy the requirement of the self-ascription of representations. But stipulating that they are necessary falls short of specifying *the condition under which* concepts of the objective enable a differentiation between what is subjective and what is objective; and to all appearances this is what Strawson fails to do.

There is a modal issue here requiring our attention,[25] for UC1 is tantamount to the following modal claim (where AN stands for Absolute Necessity):

---

[24] It could be countered that a distinction should be observed between the claim (A) *necessarily*, all R are unified in and by S (where R stands for representation and S for subject, viz. the thinking 'I') and the claim (B) *necessarily*, all R are unifiable in and by S. Notice that Strawson also does not appear to respect the difference between A and B. However, for the purposes of indicating the metaphysically intemperate sense in which Strawson construes the premise of self-ascription, this distinction is not relevant, for in both cases A and B a claim is made with regard to the modally absolute sense in which R is related to S: no R is not subject to the condition under which it is either united or unifiable in and by S.

[25] I want to cash out the puzzling dual modality in the verbal phrase 'must be able' in Kant's 'I think' proposition. I analyse this feature, which concerns the deduction of the categories of modality, in detail elsewhere (Schulting 2018b), ch. 6.

UC_{AN}: $(\forall x)(x$ is a representation$) \longrightarrow \Box\ [(\exists y)(\exists z)(y$ is an identical thinker 'I'$) \wedge (z$ is an analytic unity of representations$) \wedge (y$ thinks z by way of self-ascribing her representations$) \wedge (x$ is united or at least unifiable in z with all other representations self-ascribed by $y)]$

On account of UC_{AN}, no criterion for identification of singular representations is needed so as to differentiate them from representations that do not share the mark constituting their sameness (z) by being thought by y. Singular representations share *by implication* a unitary mark that identifies them as belonging to a representing self who self-ascribes them (hence the widely held belief that self-ascription is criterionless). All representations a self has are subject to UC_{AN}. UC_{AN} underlies PS; recall that, according to PS, all representations are necessarily ascribed or at least ascribable in and by an identical thinker 'I' and so, by implication, belong *eo ipso* to a single unitary consciousness (UC1). What is unclear, however, is in what way, assuming PS, Strawson thinks that the objective unity (premise T2 of the transcendental argument), as a means of differentiation of the subjective from the objective, constitutes the necessary ground of the subjective unity of representations (premise T1). That is to say, it is unclear, on account of UC_{AN} and given T's analytic nature, how T2 can be shown logically to be inferable from T1. What is the nature of the grounding relation? More precisely: What is it that makes, logically, T2 is analytically derived from T1 such that T2 is necessarily entailed by T1?[26] Surely, it cannot be T2 itself. It must be some analytically explicable criterion inherent to T1 for the inference to work. I see nothing in T1, if construed as amounting to UC1, that points to such a criterion.

However, in order to prevent metaphysically intemperate claims of the kind that Strawson seems committed to and the resultant argumentative lack of clarity as regards the inferential relation between the premises of T, let us suppose that the claim regarding the premise of the unity of consciousness comes down to a mere conditional (in conformity with PS′′). The conditional would comport with construal UC2 and can be formulated as follows (where $_{CN}$ stands for Conditional Necessity):

[26] Notice that although the conclusion of any arbitrary syllogism is analytically (hence necessarily) entailed, its premises of course need not themselves be necessarily related. This is different with the type of inference that is a transcendental argument (T), whose premises *are* necessarily related because they express what Strawson calls an 'analytical connexion'. My question thus concerns the force of T in terms of how each of its premises are conceptually (analytically) linked.

UC$_{CN}$: $\Box$ $\{(\forall x)(x$ is a representation) $\wedge$ $(\exists y)(y$ is an identical thinker 'I') $\wedge$ (y self-ascribes x) $\longrightarrow$ $[(\exists z)(z$ is an analytic unity of all representations self-ascribed by $y)$ $\wedge$ (y thinks z) $\wedge$ (x is united in z by $y)]\}$

It is evident that UC$_{AN}$ and UC$_{CN}$ spell out two distinct modal claims that should not be confused. In the case of UC$_{AN}$, as we have seen, no representations are excluded from being unified, or unifiable, in and by a single self-ascribing subject of representation (an identical thinker 'I'). In the case of UC$_{CN}$, a condition is specified to the effect that representations are united so as to show an analytic unity of consciousness *if* they are taken together by the subject by way of self-ascribing her own representations. Unlike UC$_{AN}$, with UC$_{CN}$ the assertion regarding unity requires an antecedent condition for its satisfaction, viz. an act or function of identification that the thinking self operates by way of self-ascription of representations; self-ascription is thus a condition for unity that I argued is not fulfilled by mere representing alone. This suggests that representations are not unified (in the strict sense) as a matter of course, nor necessarily subject to a condition of unifiability for that matter. With UC$_{CN}$, no claim, then, is made with regard to the putative existential unity of representations (UC1), whereas with UC$_{AN}$ a claim *is* made to the effect that representations could not *be* otherwise than united or at least potentially united in and by the unity of consciousness, which boils down to an existential claim as to the unifiedness or unifiability of representations (by implication, unruly representations that fail to fit into the unity of consciousness are ruled out on UC$_{AN}$). Also, on account of UC$_{CN}$, the analytic relation between T1 and T2, between the subjective and the objective, can be made clearer, for on this reading both subjective and objective representations, that is, self-ascribed representations as well as representations of the objective, are grounded on the same condition of recognition for unitary representation. (This condition is a priori synthesis, whose discussion I have had to bracket here.)[27]

6.6 *Conclusion*

We are faced with a dilemma. If Strawson wants to argue that there is an analytical connection between the subjective and the objective, then he needs something more than the trivial truth to which the premise of self-ascription, as Strawson construes it (PS),

[27] For an extensive account, see Schulting (2018b).

amounts. For, as I have argued, there is nothing intrinsic to that trivial truth that leads us, inferentially, towards the conclusion of T, that is, that objective connectivity is a necessary condition of the self-ascription of representations. The most Strawson can get out of an argument relying on PS is a short argument from representation (or experience, which for Strawson is equivalent to representation) to that of which the representation or experience is, i.e. to objectivity. This is in effect what Strawson argues in terms of the thinkability or conceptualisability requirement. In this way, however, Strawson fails to explain the self-reflexiveness, or 'analytical connection', between self and objectivity *from* the premise of self-ascription. But we have also seen that Strawson not just advances the conceptualisability argument, but in fact argues for a different claim that reveals a commitment to UC_{AN}. These two claims are clearly in tension.

However, if Strawson were to concede to the conditional construal UC_{CN} of the premise of T, then the anti-sceptical force a transcendental argument is presumed to have is significantly diminished.[28] For the argument would then boil down to a hypothetical argument regarding the unity of representations that will not cajole a sceptic into conceding defeat.[29] On that reading, a radical sceptic could still persist in the conviction that only representations exist that are not unified in the strict sense of being self-ascribed to an identical self (over time) and so do not belong to an analytic unity correspondent with the objective ways of the world, but instead are nothing but mere aggregates of representations $r_{each}1...r_{each}n$ with no *intrinsic* common mark between them that would constitute their sameness. Of course, Strawson wants to avoid this result at all costs. Therefore, given how

[28] At times, Strawson does appear to understand the argument such that it concerns a mere *de dicto* claim with regard to the necessary requirements for representations to have objective reference —e.g. Strawson 1968:89; notice the implicit conditional structure of Strawson's claim here that '[w]e could not employ any ordinary empirical concepts of objects unless our manifold perceptual experiences possessed the kind of coherence and interconnexion which is required for the application of such concepts' (emphasis added). In like manner, he argues that '*if* any phase of experience is to count as a phase of experience of the objective, we must be able to integrate it with other phases as part of a single unified experience of a single objective world', thereby ruling out 'unruly perceptions' (Strawson 1968:89, emphasis added). However, as Strawson observes (1968:92), the Deduction argument in terms of a proof is not 'simply a matter of [giving] the definition of 'experience' that experience involves knowledge of objects'. Hence, the premise of the real proof in the Deduction cannot be the actuality of knowledge of objects, and so for Strawson the argument cannot in effect be a conditionally construed inference.

[29] A sceptic could point out that, taken thus, the transcendental argument would appear to rest on a *petitio principii*, for what had to be proved, viz. objective experience as a condition of subjective experience, is already assumed to be a fact in the premise.

Strawson understands the purport of the transcendental argument, a conditional construal of its premise is not what he appears to have in mind.[30]

I cannot of course in the space of a chapter assess all aspects of Strawson's influential reconstructive strategy for reading Kant's argument. What I have tried to show is that Strawson's premise (T1) reveals an ambiguity regarding modality. The fallacy of reasoning resulting from it can, I have suggested, be avoided if one heeds the conditional purport of Kant's argument, which would include a commitment to a priori synthesis. But Strawson rejects outright the latter and appears to ignore the former. Consequently, he fails, to my mind, to provide the genuine analytical connection between the subjective and the objective that he rightly wishes to highlight.

[30] However, even though objective experience, or the concept thereof, is presupposed in the argument's premise as a fact, a fact that the sceptic does not accept, on UC_{CN} it would still be problematic for a sceptic to make a claim as to the denial or impossibility of a necessary unity of representations correspondent with an objective unity, for a sceptic in the business of making such negative claims must nonetheless *eo ipso* be in the business of forming identical thoughts of her own through self-ascription, and hence would be subject to self-refutation.