This is only a penultimate draft. The final version of this chapter will appear in:
M. Frauchiger [ed.], @@@ - *Themes from Dretske*. (Lauener Library of Analytical Philosophy, ed. by W. K. Essler, D. Føllesdal, and M. Frauchiger, vol. 6) Berlin/Boston: De Gruyter, forthcoming 2019.

Thomas Metzinger

Dretske on transparency

1. Introduction

The epistemic goal of this chapter is to evaluate if Fred Dretske's conception of phenomenal transparency is still valid, and whether it can be refined and extended into the present debate. My main claim is that his model of introspective self-knowledge is untenable as it stands, but that the conceptual foundations provided by his information-theoretic and representationalist approach to the conscious human mind still possess considerable relevance and fecundity. The structure of this contribution is as follows: In section 2 I will briefly look at Dretske's concept of "transparency". Section 3 will offer a representationalist definition of phenomenal transparency, flag three possible fallacies of equivocation, and briefly contextualize "phenomenal transparency" within the history of classical analytic philosophy of mind. In section 4, I will enrich the phenomenological account of transparency by adding two further constraints every future theory should satisfy. Section 5 first looks at the concept of "introspection", separating epistemological and phenomenological readings, and then draws attention to the fallacy of the representational divide and the impossibility of naturalist self-indicators. The final section offers a general model of phenomenal transparency. It also offers two specific examples of how the problem of transparency can be treated using more recent conceptual tools in the specific domains of perceptual presence and action generation. These two examples can also be read as tentative proposals for how Dretske's original points can be connected to the current debate in interdisciplinary philosophy of mind and cognitive science.

2. Dretske on transparency

Fred Dretske was a representationalist. In 20th century analytic philosophy of mind he was perhaps the leading pioneer in fruitfully connecting classical philosophical terminology to the novel conceptual instruments offered by information-processing theories of mind and brain. Let us begin by looking at a series excerpts from *Naturalizing the Mind*:

> **The Representational Thesis**
>
> (1) *All mental facts are representational facts*, and (2) *All representational facts are facts about informational functions*.

> Representational Naturalism (as I shall call the view defined by the Representational Thesis) helps one understand, for example, why conscious experiences have that peculiar diaphanous quality – the quality of always being present *when,* but never *where* one looks to find them. It provides a satisfying account of the qualitative, the first-person, aspect of our sensory and affective life – distinguishing, in naturalistic terms, between *what* we experience (reality) and *how* we experience it (appearance). (Dretske, 1995, p. xiii)

Here, "looking to find" a conscious experience quite obviously refers to a special kind of epistemic agency, which is a *mental* form of agency: it is the volitional, goal-directed control of introspective attention. We can therefore read Dretske as saying that if introspective attention is directed at conscious experiences instantiating a specific sensory or affective character, then it always finds them as part of a phenomenal *Now*, but not necessarily as elements of an *inner* space, a space that can be described as containing directly mind-dependent entities. In other words, the first phenomenological aspect of diaphanousness refers to *temporal* internality ("Nowness", Dretske's "quality of always being present"), while the second phenomenological aspect refers to *psychological* internality ("being in the self-conscious mind").

Dretske also was a naturalist. From this arises the first, possibly problematic, ambiguity: If we want to follow Dretske in being naturalists, then all the carriers and informational processes realizing first-order conscious experiences will be *physical* processes. They will therefore be time-consuming. The same applies to any form of introspective access to those processes. They will be extended not only in space, but also in time. There will be signal transduction times, variable processing speeds in different parts of the brain, nested temporal scales, and ultimately there will be a neurally realized process of arriving at an introspectively available model of the relevant conscious experience. For a naturalist, there is no such thing as true "instantaneousness", because "the quality of always being present" mentioned by Dretske can itself only be a phenomenal quality, a property of how the target of introspection *appears* to us as subjects of experience. Therefore, the *de nunc*-character of introspective access is an illusion. The "quality of always being present" is something that can only appear in a smeared phenomenal moment, for example a Jamesian specious present – it is not instantaneousness in any more robust epistemological or metaphysical sense. What is needed is a representationalist analysis of the phenomenal quality of "instantaneousness", which ultimately leads to the conceptual ambiguity pointed to above.

A second goal for any systematic extension of Dretske's thought is a richer description of the phenomenological *domain* for which diaphanousness is claimed, including its boundaries. A naturalist finds nothing like a "self" anywhere in the brain or the natural world. So, what is needed seems to be an information-theoretic and representationalist model of the second phenomenological aspect: psychological internality or "being in the self-conscious mind." Fred Dretske speaks of our "sensory and affective life", but we have cognitive and intersubjective lives too. We are embodied agents, beings possessing a highly differentiated and fluid motor phenomenology; we are empathic and sometimes compassionate, enculturated and are situated in a specific cognitive niche. In addition, our conscious self-model is grounded in a rich melange of interoceptive qualities: we have vestibular sensations, an invisible peripersonal space around us, and we transparently locate ourselves in multiple spatial and temporal frames of reference. Are all these phenomenological subdomains transparent? Does the original diaphanousness-claim hold for all of them? Are some of them opaque? Is there perhaps an underlying temporal dynamic, i.e., phenomenal trajectories connecting transparent and opaque states? Are there phenomenological *degrees* of diaphanousness?

One advantage of the "Representational Thesis", according to Dretske, is that it helps us to understand the phenomenology of mind-independence while at the same avoiding a homunculus fallacy in developing a theory of inner knowledge:

> It demystifies introspection; the mind's knowledge of itself no longer requires an internal "eye" observing the clockwork of the mind. And it provides an answer – a biologically plausible answer – to questions about the function or purpose of consciousness.
>
> The thesis is less plausible – some would say completely implausible – for sensory affairs, for the phenomenal or qualitative aspect of our mental life. [...] Even if thought, belief, and judgment can be understood as internal representations of external affairs, sensations, experiences, and feelings cannot. (Dretske, 1995, p. xiv)

Limiting intentional content to propositionally structured mental representations is one example of a strategy that was likely already a minority position in 1995 (Churchland, 1989; Clark, 1989) and which has largely been given up in more recent, empirically informed philosophy of mind (Clark, 2015; Hohwy, 2013; Metzinger & Wiese, 2017c). Vague umbrella terms like "experience" and folk-psychological distinctions like "cognition", "perception", "action", and "feelings" are basically a thing of the past, having been replaced by much more fine-grained conceptual instruments. For example, many authors today would see Dretske's "affective states" as predictive models at some level of a neurally realized hierarchical Bayesian model, involving homeostatic cost functions and the uncertainty involved in variable prediction error reduction rates over time (for an example, see van de Cruys, 2017). Given this context, one further desideratum for a systematic extension would be to integrate the phenomenology of cognitive agency (i.e., the specific phenomenal character instantiated by active, controlled, and high-level symbolic thought; Metzinger, 2013a, 2017) with the best current theories of affect, sensory perception, and attention, possibly uniting them in one single formal framework (Friston, 2010).

But let us stay with one of Dretske's central insights for now, the idea of "displaced perception" and its application to the problem of introspective self-knowledge (see also Dretske, 1999):

> Perceptual displacement – seeing that *k* is F by seeing, not *k*, but some other object, *h*, occurs when there is conceptual, but no corresponding sensory, representation of *k*. [...] Perceptual displacement enlarges the number of facts one perceives without a corresponding enlargement in the number of objects one perceives.
>
> If, then, introspective knowledge is a species of displaced perception, it is an instance in which an experience (of blue, say) is conceptually represented as an experience of blue

> via a sensory representation not of the experience, but of some other object. [...] Introspective knowledge of E requires no other sensory representation of objects than those already being represented by E – the experience one comes to know about. (Dretske, 1995, pp. 42–43)

In 1995, I edited an anthology of philosophical texts titled *Conscious Experience* (Metzinger, 1995). In that volume, Fred Dretske's research assistant Güven Güzeldere introduced the "fallacy of the representational divide" as the problem of all metarepresentational theories of consciousness relying on the notion of "inner perception" (Güzeldere, 1995). In a nutshell, any form of perception-like internal meta-representation directed at first-order sensory states could only grasp their non-intentional or "carrier" properties. These "carrier" properties would represent concrete properties of the internal vehicles, but not their content, which, arguably, is an abstract property (see Dretske, 1988 for more discussion on the conceptual distinction between "vehicle" and "content" for representations). In criticizing David M. Armstrong's model of higher-order perception (Armstrong, 1968), Güzeldere writes:

> I think that the reasoning which tries to account for the changes in the attentiveness, vividness, etc. of the externally directed ordinary perception (from the truck driver's auto-pilot mode to his attentive condition) in terms of an internally directed, higher-order perception of lower order perceptual states, is based on a fundamental mistake worthy of a name – I will call it the 'fallacy of the representational divide'.
>
> Mental states *qua* brain states (in a materialist ontology) are the *vehicles* that represent for us the world around, as well as in, us. What we thus become aware of is what those states represent as being a certain way, i.e. their *content,* in terms of the properties of what is represented. The 'fallacy of the representational divide' … [is] a tacit attempt to replace what is being represent-*ed* with that which is the represent-*er.* […] In other words, no matter how much we find out about the intrinsic properties of representational states, we may simply not be able to reach the other side of the 'representational divide', in virtue of this alone, and get to the extrinsic, relational properties of those states. (Güzeldere, 1995, p. 350)

Can the model of displaced perception be applied to the problem of introspective consciousness without creating the fallacy of the representational divide? Andreas Kemmerling (1999) demonstrates that nothing can be a natural meta-representation of some fact x *and* a natural x-representation at the same time, as demanded by Dretske's naturalist model of conscious representation. On closer observation, there are serious conceptual difficulties with Dretske's theory of introspective self-knowledge, as well as with his use of the folk-psychological term "perception" and the application of his theory of displaced perception to the special case of introspective self-knowledge. I will return to these problems in sections 5.2 and 5.3, but for now, let us remain with his highly original and innovative idea of a transparent form of introspection, a form of introspection that does not itself generate any additional self-related phenomenology:

> In the case of introspection, the perception is displaced, yes, but the necessary information is there *whether or not the intermediate representations* are veridical. [...]

> This, I submit, is the source of the “directness” and “immediacy” of introspective knowledge.
>
> If there is an inner sense, some quasi-perceptual faculty that enables one to know what experiences are like by “scanning” them, this internal scanner, unlike the other senses, has a completely transparent phenomenology. [...] Introspection has no phenomenology because the knowledge one gets by it is (itself) experience-less. One can, by introspection, come to know about experience, but the knowledge is obtained without any experiences beyond the ones one comes to know about [...] There is no need for an experience of the experience. All one needs is a belief about it. (Dretske, 1995, p. 61)

According to this model, there are two meta-representational mechanisms by which the human mind can meta-represent “experiences” or first-order perceptual states. The first is transparent inner perception. This generates the phenomenology of immediacy and direct realism because the *mechanism* itself (the “scanner”, to use Dretske’s metaphorical terminology) can never itself become the object of introspective attention. The second is belief formation. This is the process of generating a propositional attitude – a high-level symbolic representation of the fact *that* a certain, perceptually given, state of affairs holds in the world.

This raises the question of whether the process of belief formation itself can be characterized by a distinct phenomenological profile, and if this profile can be distinguished from the stable “holding” of a perceptually grounded belief once it is firmly established. Are the relevant connecting beliefs dynamic entities? Given recent advances in neuroscience, cognitive science, and computational modelling, the question also arises if Dretske’s original idea could perhaps be couched in more precise terms today. I will return to this below.

Let us summarize this overview by noting a number of routes by which Dretske’s original work could be refined and extended:

1. The phenomenology can be made much more precise and comprehensive:
    a. Does the conscious experience of belief formation have its own distinct phenomenological profile?
    b. If Dretske is right and there are examples of a “completely transparent phenomenology”, do we also know of any examples exhibiting a “completely opaque phenomenology”?
    c. Can one and the same mental process *vary* in terms of transparency and opacity?
    d. What other phenomenological constraints should be satisfied by a more comprehensive theory of introspection and phenomenally transparent self-representation?
2. The epistemology of introspective self-knowledge can be made much more precise and comprehensive:
    a. Dretske’s own theory of introspection has non-trivial difficulties – can the underlying conceptual issues be resolved?

    b. A more fine-grained theory of introspection must be developed that clearly distinguishes between phenomenological and epistemological aspects of introspective experience.
    c. Are all kinds of introspective self-knowledge forms of (inner) epistemic agency, or can introspection also be a fully automatic, subpersonal process?
3. The empirical grounding can be improved given knew knowledge:
    a. As an analytic philosopher, Fred Dretske was remarkably open to interdisciplinary cooperation and eager to build bridges between philosophy and the cognitive sciences (Dretske, 1981, p. viii) – are there new insights in the domain of empirical consciousness research that possess direct relevance to his original project?
    b. Do current formal models of mental representation offer new conceptual instruments that could be fruitfully applied on a philosophical level?

3. A representationalist concept of phenomenal transparency

Transparency is a property of phenomenal representations; unconscious representations in the brain are neither transparent nor opaque. *Phenomenal transparency* consists in the fact of only the *content properties* of a conscious mental representation being available for introspection, but not its non-intentional or "vehicle-properties". Introspectively, we can access a representation's content, but not the *carrier* of this content. This is what creates the phenomenal character of subjective immediacy, direct introspective access, and naïve realism in perception.

> **Phenomenal transparency**: Some phenomenal states are transparent in that only their content-properties are introspectively accessible to the subject of experience.

When Dretske speaks of transparent phenomenology, he clearly means it in this sense. The introspective scanner is transparent in that the content it generates, but not the underlying information-processing mechanism itself, is available for attention.

In the philosophical literature, there are at least 3 other uses of the term "transparency" which must not be confused with the notion of "phenomenal transparency" in order to avoid fallacies of equivocation.

3.1 Epistemic transparency

Epistemic transparency is an internal property of the mind as a whole, as in introspective acts. Consider, for example, the certainty implied by the alleged self-evidence and self-verification of the Cartesian *cogito*. "Epistemic transparency" here refers to a modal intuition about internal states experienced as "mental": that systematic misrepresentation or an unnoticeable lack of introspective self-knowledge are impossible (Metzinger & Windt, 2015, section 3.1). As an epistemological thesis, this intuition needs independent argument. Given new empirical results about the evolution of self-deception (Hippel & Trivers, 2011) as well as more than a century of clinical neuropsychology, it has become highly implausible. For example, in Anton-Babinski syndrome – a phenomenal configuration following cortical blindness and also known as "visual anosognosia" – patients will demonstrably suffer from a lack of visual awareness while at the same time consistently denying their own blindness (Anton, 1898; Benson & Greenberg, 1969; Metzinger, 2003a, 234p). They will show all the signs of functional blindness while confabulating about the experiential contents of their own minds. However, while the epistemic-transparency thesis for self-consciousness and introspection has long been empirically falsified, it correctly and interestingly describes important aspects the *phenomenology* of human self-awareness, the phenomenally experienced self-certainty of the subject (or "*Selbstgewissheit des Subjekts*"). This phenomenology needs a conceptual analysis and an empirical explanation of its own (Metzinger, 2003a, 2008).

We can now see how the classical modal intuition is *causally rooted* in the transparent phenomenology of introspection: From a first-person perspective, the classic modal intuition is highly plausible because the specific, abovementioned phenomenal character of subjective immediacy and direct introspective access very often makes systematic misrepresentation *phenomenologically* impossible. This explains why many Cartesian arguments implicitly conflate phenomenological impossibility and epistemological impossibility.

3.2 Referential transparency

"Transparency" can also be read as a property of contexts, where extensional (or 'referentially transparent') contexts are constituted by sentences characterized by intersubstitutivity of coreferential terms *salva veritate*; and an implication towards the existence of the entities mentioned by them. However, referentially opaque contexts exist as well, for example those that are constituted by propositional attitudes, temporal and modal terms, or indirect speech. Non-linguistic creatures could certainly have phenomenally transparent states in the first sense (i.e. enjoy subjective qualities of immediacy, direct perception, and naïve realism), but it is hard to imagine a creature without language creating a Cartesian *cogito* or a self-conception of being an ideal observer with an epistemically transparent mind.

Referential transparency is very different to phenomenal transparency but there may be a deep causal connection between semantics and the speaker's transparent phenomenology of direct perception: transparently perceived sensory objects are perceived as *existing* and as directly *given*. The naturally evolved biological function of phenomenally transparent representation could therefore lie in the representation of facticity and epistemic reliability, and in supporting existence assumptions on different levels of world-modelling (for example via high model evidence and counterfactual depth; see section 5.2). Call this the "reference thesis": The capacity for transparent cognitive and linguistic reference presupposes a phenomenally transparent representation of the referent. I will not develop this point here, but there clearly is a variable phenomenology of existence *as such* (Ratcliffe, 2008) which probably grounds semantic properties like referential transparency by causally enabling the relevant linguistic operations. For example, modern computational models of phenomenal transparency exactly help us understand the functional conditions under which the human brain will represent an object as *present* (Metzinger, 2014; Seth, 2014).

3.3 Transparency in communication theory

Transparency can also be a functional property of information channels (Dretske, 1981, p. 38). The three defining characteristics of this notion of transparency are, first, that it accepts unmodified user information as input; second, that it delivers user information that is unmodified with respect to form and informational content on the output side; and third, that user information may well be internally changed and reprocessed in many different ways, but is always retransformed, without causal interaction with the user, into the original format before reaching the output stage. E-mail is an obvious example: A single message is re-coded into binary and may be broken into many chunks with each taking different paths and "hops" through the internet before being reassembled on the other side. However, the user has no access to the subpersonal mechanisms underlying successful personal-level communication.

Obviously, phenomenal transparency in the sense intended by Fred Dretske is not a property of technical systems. However, it is interesting to once again note the parallel that emerges if we view the neural correlate of consciousness or the conscious model of reality as a *medium:* This medium is transparent insofar as the subpersonal processing mechanisms contributing to its currently active content are attentionally unavailable to high-level introspective processing. Consciousness is an invisible interface. It helps to connect biological information-processing systems. In particular, this is interestingly true for some forms of consciously experienced social cognition *not* mediated by artificial information channels like the internet, like "mind reading" (Carruthers, 2017; Vogeley et al., 2001). Phenomenologically, we sometimes seem to "directly" know what another person thinks or feels, because the corresponding phenomenal person-model has become transparent. Communicating with *ourselves* via conscious, self-directed thought is an interesting special case: Phenomenologically, the degree to which we "directly" know what we *really* think or feel ourselves is characterized by considerable variance (see sections 4.1 and 4.2).

3.4 The wider context: Phenomenal transparency in Analytic Philosophy of Mind

Below are two passages from a now-classical paper by G.E. Moore, titled *The Refutation of Idealism*:

> [T]he fact that when we refer to introspection and try to discover what the sensation of blue is, it is very easy to suppose that we have before us only a single term. The term 'blue' is easy enough to distinguish, but the other element which I have called 'consciousness' – that which a sensation of blue has in common with a sensation of green – is extremely difficult to fix. [...] And in general, that which makes the sensation of blue a mental fact seems to escape us; it seems, if I may use a metaphor, to be transparent – we look through it and see nothing but the blue; we may be convinced that there *is something*, but *what* it is no philosopher, I think, has yet clearly recognised. (Moore, 1903, p. 446)

> [T]he moment we try to fix our attention upon consciousness and to see what, distinctly, it is, it seems to vanish: it seems as if we had before us a mere emptiness. When we try to introspect the sensation of blue, all we can see is the blue: the other element is as if

> it were diaphanous. Yet it can be distinguished if we look attentively enough, and if we know that there is something to look for. (Moore, 1903, p. 450)

Moore's paper is the *locus classicus* for the concept of "phenomenal transparency", a property of some conscious states which he called "transparency" or "diaphanousness". Today, a standard definition of "phenomenal transparency" is that it essentially consists in only the *content properties* of a conscious mental representation being available for introspection, but not the fact that it also has non-intentional or "vehicle-properties" (Metzinger, 2003a, 2003b). In other words, it is not experienced *as a representation*; introspectively, we can access its content, but not the content-formation process itself. The physical carrier is invisible. Often, it is assumed that transparency in this sense is a property of *all* phenomenal states. However, this assumption is incomplete – *opaque* phenomenal representations also exist (see section 5).

The most notable phenomenological examples of opaque state-classes are consciously experienced thoughts: we experience them as mind-dependent and internally constructed, as mental representations that could be true or false. Further, some emotions, pseudo-hallucinations, and lucid dreams are also subjectively experienced *as representational processes*. They also make the possibility of *mis*representation introspectively available. This is an important causal function as it decisively contributes to the general intelligence of an information-processing system. Phenomenal opacity frees a system from naïve realism. There is therefore a spectrum between phenomenal opacity and phenomenal transparency, and any given content can vary along this spectrum. Phenomenally opaque processes sometimes appear to us as deliberately initiated cognitive or representational processes. However, they can also appear as automatic or spontaneously occurring, as limited or even global phenomenal simulations, and they frequently seem not to be under the experiential subject's control (Fox & Christoff, 2018; for more, seeMetzinger, 2003a, 2015a; Metzinger, 2003b; Metzinger, 2013a, 2003b, 2014, 2018).

Moore presents a phenomenological argument against the transparency of consciousness *as such*. Many analytic authors have read Moore as if he were talking about the "transparency of qualia"[1]. However, the relevant point is not about the transparency of awareness or qualia, but rather the fact that consciousness as such can be *made* phenomenally opaque (for a well-researched and substantial discussion, see Hellie, 2007). Moore's much deeper point is that conscious experience *as such* has phenomenal character *sui generis* (namely, a second-order "awareness-of", a relational "signature of knowing"), and that this character can sometimes be detected by introspective attention. In fact, Moore's self-stated goal in introducing "transparency" is actually "to try to make the reader *see* it" (Moore, 1903, p. 450). His relevant phenomenological claims are these:

- **(M1)** There is one most general phenomenal property, which is shared by all sensory qualities:
  → "Consciousness" *as such*, pure awareness, phenomenality *per se.*

[1] Cf. Block (1996, p. 26-27), Kind (2003, p. 229), Tye (1992, p. 160, 2002, p. 139), Kennedy (2009, p. 574-577), Speaks (2009, p. 539), Stoljar (2004, p. 341). For an excellent and lucid l discussion containing these and further references, see Hellie (2007). See Metzinger 2003a, 2003b, 2014 for a more detailed treatment of phenomenal transparency.

- **(M2)** Under standard conditions, this global property is "transparent" (Moore, 1903, p. 446), a "mere emptiness" or "diaphanous":
  - → It is not explicitly experienced, but it *can* be.
- **(M3)** This property is *evasive*:
  - → It "seems to vanish" under attentional agency; i.e., if we actively try to "fix our attention" on it (Moore, 1903, p. 450), thereby creating an Epistemic Agent Model(EAM).[2]
- **(M4)** This property can *become phenomenally opaque*, under two conditions:
  - → we look "attentively enough",
  - → and we "know" that there is a possible object for introspective attention (Moore, 1903, p. 450).
- **(M5)** It is difficult to simultaneously direct and sustain introspective attention to the global property in question and to concrete perceptual qualities:
  - → Moore found no philosopher in the literature who was "able to hold *it* and *blue* before their minds and to compare them, in the same way in which they can compare *blue* and *green"* (Moore, 1903, p. 450).
- **(M6)** Awareness is relational and consciousness is a second-order epistemic process:
  - → Having a sensation is an awareness *of* something, and consciousness is the knowledge that this awareness currently exists (Moore, 1903, p. 449).
- **(M7)** Consciousness is meta-awareness:
  - → The *type* of epistemic relation is identical – "awareness of" is the same relation in sensation as it is in becoming aware of this awareness (Moore, 1903, p. 449).

One could interpret Moore as saying that an aperspectival and non-agentive form of meta-awareness is co-instantiated with all forms of conscious perceptual knowledge; the second-order relation of an awareness-of the current existence of an awareness-of some specific perceptual phenomenal character is mostly unnoticed (i.e., transparent), but that we can make it phenomenally opaque by attending to it.

There are historical precursors to Moore's classical treatment, which in turn influenced Fred Dretske. For example, Thomas Reid used the interesting concept of "unattended instantaneousness":

> We are so accustomed to use the sensation [of hardness] as a sign, and to pass immediately to the hardness signified, that, as far as appears, it was never made an object of thought, either by the vulgar or by philosophers [...] There is no sensation more distinct, or more frequent; yet it is never attended to, but passes through the mind instantaneously, and serves only to introduce that quality in bodies, which, by a law of our constitution, it suggests. (Reid & Brookes, 1858, p. 120)

This is an important observation, because it for the first time directs our attention to the possibility that different *processing speeds* could play a role in understanding the relation between an ongoing sensation "passing through the mind" and the mechanism of introspective

[2] An EAM is an internal representation of the system as capable of epistemic self-control, of actively influencing what it will know and what it will not know. According to subjective experience, we are entities that actively construct and search for new epistemic relations to the world and ourselves. See for example Metzinger (2017); § 2.5.

attention directed at it, a causal interaction often leading to the reification of the sensation’s content and phenomenal transparency.

Ludwig Wittgenstein drew attention not to temporal immediacy and the *de nunc*-character of transparent phenomenal experience, but rather to the absence of a distinct phenomenal character signifying psychological internality terming it *Die Abwesenheit des “Gefühls des In-dich-selber-Zeigens”*:

> Schau auf das Blau des Himmels, und sag zu Dir selbst „Wie blau der Himmel ist!“ – Wenn Du es spontan tust – nicht mit philosophischen Absichten – so kommt es Dir nicht in den Sinn, dieser Farbeneindruck gehöre nur *dir*. Und du hast kein Bedenken, diesen Ausruf an einen Andern zu richten. Und wenn Du bei den Worten auf etwas zeigst, so ist es der Himmel. Ich meine: Du hast nicht das Gefühl des In-dich-selber-Zeigens, das oft das ‚Benennen der Empfindung‘ begleitet, wenn man über die ‚private Sprache‘ nachdenkt. Du denkst auch nicht, du solltest eigentlich nicht mit der Hand, sondern nur mit der Aufmerksamkeit auf die Farbe zeigen. (Überlege, was es heißt, „Mit der Aufmerksamkeit auf etwas zeigen“.) (Wittgenstein, 1971, § 275)

Introspective attention to interoceptive sensations might be an interesting counterexample here. One relevant question is if self-directed guided attention could count as a subsymbolic form of mental demonstration (for a recent discussion, see Hofmann, 2018). Mental demonstration could certainly be limited, or it could systematically fail and leave us unaware of certain critical aspects of our own minds. More recently, Gilbert Harman spoke of the “unawareness of intrinsic non-intentional features”:

> [I]n the case of her visual experience of a tree, I want to say she is not aware of, as it were, the mental paint by virtue of which her experience is an experience of seeing a tree. She is aware only of the intentional or relational features of her experience, not of its intrinsic non-intentional features. [...]
>
> When you see a tree, you do not experience any features as intrinsic features of your experience. Look at a tree and try to turn your attention to intrinsic features of your visual experience. I predict you will find that the only features there to turn your attention to will be features of the presented tree, including relational features of the tree “from here.” (Harman, 1990)

Finally, Sydney Shoemaker is very close to the representationalist definition proposed above. He says that transparency results from a lack of introspective access to “nonintentional features of one’s experience that encode this content”:

> The only thing that seems to answer the description “attending introspectively to one’s visual experience” is attending to how things appear to one visually; and offhand this seem to tell one what the representational content of one’s experience is without telling one anything about what the nonintentional features of one’s experience are that encode this content. One may be inclined to say that one is revelling in the qualitative or phenomenal character of one’s experience when one “drinks in” the blue of a summer sky or the red of a ripe tomato. But neither the blue nor the red is an object of

> *introspective* awareness; these are experienced, perceptually rather than introspectively, as located outside one, in the sky or in the tomato, not as features of one's experience. G.E. Moore once complained that the sensation of blue is "as if it were diaphanous"; if one tries to introspect it one sees right through it, and sees only the blue. In a similar vein one might say that qualia, if there are such, are diaphanous; if one tries to attend to them, all one finds is the representative content of the experience. (Shoemaker, 1990).

These are just examples. If we accept the history of ideas within analytic philosophy itself as a theoretical background context, then we can extract a semantic core for the notion of "phenomenal transparency". It has three major components: Temporal immediacy (Reid), a failure of attention fixation (Moore), and an absence of introspective access to non-intentional features (Harman, Shoemaker, and others). Given this analytic background, let us now extend Fred Dretske's ideas, first by attempting to take the actual phenomenology more seriously than most authors in the analytic tradition have done.

4. Extension 1: Towards a richer phenomenological account

In the following sections, I will briefly enrich our understanding of conscious experience by adding two further constraints: the fact that *phenomenally opaque* representations do exist, and the fact that there is a *variable gradient* of phenomenal opacity versus transparency.

4.1 Examples of phenomenally opaque processes

The most obvious example of a conscious process that is phenomenally experienced *as* representational and *as* mind-dependent is agentive high-level cognition. Consciously experienced cognitive reference is opaque (Metzinger, 2003b): We operate on mental representations while actually being able to attend to the *process* by which they are constructed, rearranged, disambiguated, and so on. However, as recent empirical research in the burgeoning field of task-unrelated and spontaneous thought clearly shows, what we have called "conscious thinking" in the past is an automatic, subpersonal process for at least two thirds of our conscious life-time (for an overview and introduction, cf. Fox & Christoff, 2018; Metzinger, 2015a). Phenomenological examples of conscious thought as a form of unintentional mental behaviour are:

- Daydreaming (the spontaneous flow of imagery and fantasy; perceptual decoupling),
- unbidden memories (unintentional retrieval of episodic memory without meta-awareness),
- automatic planning (mental simulation of possible future actions and goal states),
- autobiographical rumination ("mental time travel", self-directed, task-unrelated cognition), and
- stimulus-unrelated thought (attentional lapses leading to self-generated thought chains).

Nevertheless, genuine mental action on the level of consciously experienced cognition also exists (Metzinger, 2017), and it possesses a distinct phenomenological profile. Hence, any convincing theory of conscious thought will have to explain the distribution of opacity/transparency for classical cognitive agency as well as for spontaneous, task-unrelated thought.

Examples of mental action in this sense are:

- Trying to concentrate (endogenous, volitional top-down control of attentional focus, e.g. in deliberately focusing one's attention on a perceptual object or attaching it to an abstract goal-representation),
- trying to call up a series of images from memory (active control of episodic memory retrieval),
- high-level symbolic thought (active control of semantic memory for type-token binding, categorization, or the construction of part-whole relations),
- mental calculation (mentally simulating arithmetic operations over symbols),
- reasoning, as, for example, in trying to construct a valid argument (mentally simulating logical relations holding between propositions (e.g., implication); generating premises, "forming" conclusions).

All processes in the second category are conscious, and we clearly experience them as representational and mind-dependent. They are also phenomenally opaque because the possibility of *misrepresentation* (Dretske, 1986) is an element of their phenomenological profile. There is epistemic uncertainty – during the relevant inner episode we experience certain aspects of the construction process itself, and it remains unclear if the epistemic goal will actually be reached at the end of the process.

But let us focus on the first category Fred Dretske saw as presenting a particular philosophical difficulty, namely, sensory or perceptual experiences (Dretske, 1995, p. xiii). As it turns out, we also know of many perception-like states in which phenomenal transparency and the corresponding phenomenology of direct realism are suspended, for example:

- Pseudo-hallucinations,
- hypnagogic and hypnopompic imagery,
- synesthetic concurrents,
- lucid dreams,
- derealization syndrome,
- body perception in depersonalization disorder.

Even if we further limit the phenomenological domain to our dominant sensory modality, we find many examples of phenomenal opacity in the visual domain alone:

- Benign visual illusions experienced *as* misrepresentational while they occur,
- low-level visual hallucinations and context-free, geometrical patterns (falling into four different categories: lattices, spider webs, tunnels, spirals),
- intramodal synesthetic concurrents,
- retinal afterimages,

- Charles Bonnet syndrome,
- hypnagogic / hypnopompic imagery,
- the visual phenomenology of Lewy Body dementia or Parkinson's disease dementia.

Transparency is not a necessary condition for phenomenality. An opaque phenomenal representation is one that is experienced *as* a representation, for example in pseudo-hallucinations or during a lucid dream. Typically, such states are subjectively experienced not as representational, but as obviously *mis*representational, leading to thoughts along the lines of "There must be something wrong with my eyes", "I must inadvertently have ingested a potent psychoactive substance", "Maybe I suffer from a neurological disorder like Charles Bonnet syndrome", and in a lucid dream, "I *know* that the dream world does not accurately represent my current physical environment". (Metzinger, 2013b; Windt & Metzinger, 2007). In such cases, there is a subjective sense of epistemic uncertainty which perhaps correlates with the level of prediction error on an unconscious, subpersonal level of processing, very often accompanied by a phenomenology of "unrealness" (which can be local or global), plus a phenomenology of losing control. Any convincing philosophical theory of transparency must satisfy the opacity-constraint: It should make these facts intelligible on a conceptual level. Therefore, in continuing Fred Dretske's line of thought, what we really want is an evidence-based, representationalist and information-theoretic analysis of phenomenally opaque experience (cf. sections 5.2 and 5.3).

4.2 Variance

Let us now turn to Dretske's second example, "feelings" or affective states. If we take our own phenomenology seriously, we find that transparency really is a *gradual* property of phenomenal states. Consequently, we have to do justice to the fact that the distribution of transparency/opacity in phenomenal space possesses a *variance*. Call this the "variability constraint". Examples of variably opaque state-classes on the global level of the phenomenal world-model are derealization in severe stress situations or transition phenomena in psychiatric disorders. On the level of our phenomenal self-model, depersonalization disorder and emotional ambiguity are well-documented cases in which the "realness" or mind-independence of one and the same conscious content – my own body, my current emotional state – can change over time. Let us remain with one single example, the specific phenomenal character of subjective feelings resulting from emotional ambiguity.

Imagine that you have a strong intuition that your partner is cheating on you. Guided by a direct emotional perception as it were, you *just know* (Metzinger & Windt, 2015) that something is wrong, you have an immediate and clear perception of another human being. As with all intuitive knowledge, you do not know *why* or *how* you know that your partner is betraying you, but the accompanying phenomenal experience comes with a high degree of certainty; there is a distinct phenomenal "signature of knowing" (Metzinger & Windt, 2014). It clearly seems there is an obvious fact out there, and there is an accompanying emotional state in your inner life – for example a deep sense of disappointment.

Then the emotional state becomes ambiguous. You begin to realise that you might be suffering from a pathological form of jealousy because you have been psychologically traumatized by painful discoveries of this type in the past. What was subjectively a "direct emotional perception" of another human being, and a transparent social cognition resulting in an intuition characterized by a high degree of certainty, now gradually reveals itself as possibly mind-dependent. It might be a misrepresentation! What was transparently experienced as an element of objective social reality out there is now experienced as an event in your inner life – something that is not directly given, but rather something constructed by internal mechanisms beyond your conscious control. Everyone who has lived through emotionally ambiguous states like jealousy or paranoia knows that the spectrum between full transparency and opacity is wide and that there are processes where we continuously move back and forth along this spectrum, never reaching a stable state. This is the variability constraint, and any good philosophical theory of Moorean/Dretskean "diaphanousness" should make it intelligible.

5. Extension 2: Towards a more fine-grained representationalist theory of introspection

All phenomenal experience is introspection. But not all of it is experienced *as* introspection. Before taking a second look at potential difficulties with Dretske's own theory of introspection, let us separate epistemological and phenomenological readings of the term "introspection" while at the same time taking a closer look at its two major semantic components. These two components are mental self-directedness of an epistemic agent either, first, via conceptual thought and cognitive self-reference or, second, by guided attention.

5.1. What is introspection?

Let us begin by asking, what exactly does the "intro"-component refer to? There are many different notions of internality relevant to the investigation of phenomenal and cognitive mental content: *temporal* internality (constituting the experiential "Now" and the phenomenal *de nunc*-character discussed above), *functional* internality (Clark, 1998; as in the closing of sensorimotor loops, possibly constituting complex phenomenal contents like experiential embodiment or situatedness, see Shapiro, 2014), or *representational* internality (as is relevant for self-consciousness). In our present context, it is of particular relevance to distinguish between introspection as operating on first-order states where the phenomenal content only supervenes on *physically* internal system states, and introspective operations on states which additionally *represent* their intentional content as actually being an internal aspect of the system itself. A weak version of local supervenience would state that both types of relevant first-order phenomenal properties nomologically supervene on contemporaneous and spatially internal properties realized by the central nervous system of a human organism. Here, representational internalism would then be the phenomenological thesis that introspection in some cases creates a representation of first-order phenomenal states *as* internal states of the epistemic agent itself. In other words, it creates a self-model.

Simultaneously, we must apply the distinction between attention and cognition. There are two major forms of mental epistemic agency, namely cognitive agency and attentional agency (Metzinger, 2013a, 2013b, 2017, 6p). Attentional agency is the relevant, and more fundamental variant. Attention is a selection process that episodically increases the capacity for information-processing in a certain partition of representational space. More recent philosophical theories describe it as second-order statistics or an optimization of precision expectations (Clark, 2015; Hohwy, 2013; Metzinger, 2017, p. 9). Sometimes this process can be volitionally controlled and instantiates the phenomenal property of agency (Wiese, 2018, 2018, 170-171, 234-243). Functionally speaking, attention is also internal resource allocation. Attention, as it were, is a representational type of *zooming in*, serving a local elevation of resolution and richness of detail within an overall representation. It generates non-conceptual content for which frequently we possess no transtemporal identity criteria and which is therefore often ineffable (Metzinger & Walde, 2000; Raffman, 1995). Cognition, on the other hand, enables *re*cognition, supports linguistic concept formation and generates a form of content which at least emulates compositionality, syntacticity, systematicity, etc.

Given these elementary distinctions, it will be helpful to distinguish four different notions of introspection, as there are two types of internal meta-representation, a subsymbolic, attentional kind (which only ”highlights” its object by increasing precision estimates, but does not form a mental concept), and a cognitive type (which forms or applies an enduring mental ”category” or prototype-centred region of state-space for its object).

Introspection$_1$ (“external attention”)

Epistemologically, introspection$_1$ is subsymbolic meta-representation operating on a pre-existing, coherent world-model. This type of introspection is a phenomenal process of attentionally representing certain aspects of an internal system state (hence it can be conceptually described as a form of *intro*spection), the intentional content of which is constituted by a part of the world depicted *as external*. In other words, a self/world-prior is already active on the level of the representational hierarchy the agent actively attends to, and attention exclusively targets the world-model, but not the self-model. The accompanying phenomenology is what we ordinarily describe as outward-directed attention or the subjective experience of attending to some object in our environment. Introspection$_1$ corresponds to the broad *folk-psychological* notion of attention.

Introspection$_2$ (“consciously experienced cognitive reference”)

This second concept refers to a conceptual (or quasi-conceptual) form of meta-representation, again operating on a pre-existing, coherent model of the world. This kind of introspection is brought about by a process of phenomenally representing the act of agentive cognitive reference to certain aspects of an internal system state, the intentional content of which is constituted by a part of the world depicted *as external*. Once again, a self/world-prior is already active on the level of the representational hierarchy the agent actively thinks about, and the process of concept and/or belief formation exclusively targets the world-model, but not the self-model. Phenomenologically, this kind of introspection is constituted by all experiences of attending to an object in our environment, while simultaneously *recognizing* it or forming a new mental concept of it: it is the conscious experience of ongoing cognitive reference. A good example is what Fred Dretske (1969) called "epistemic seeing".

Introspection$_3$ ("inward attention", sometimes interpreted as "inner perception")

This is a process of subsymbolic meta-representation operating on a pre-existing, coherent *self*-model, i.e., a dynamic, multimodal representation of the system as a whole (for the notion of a "self-model" see Metzinger, 2003a, 2008). This type of introspective experience is generated by processes of phenomenal representation, which directs attention towards certain aspects of an internal system state, the intentional content of which is being constituted by a part of the world depicted *as internal*. Here, attention is exclusively targeting different layers of the self-model. Good examples are interoceptive experiences like "gut feelings", as in the conscious experience of hunger, thirst, respiration, detectable variations in blood pressure, movement and vestibular sensations caused by bodily movement, emotional or affective experiences, suddenly becoming aware of a mind wandering episode, or choicelessly and non-judgmentally attending to the emergence and disappearance of thoughts during a period of classical mindfulness meditation (Hasenkamp, Wilson-Mendenhall, Duncan, & Barsalou, 2012; Hölzel et al., 2011).

The phenomenology of this class of states corresponds to what is, in everyday life, called "inward-directed attention". As we now see, the phenomenology of Introspection$_3$ is highly differentiated, depending on whether attention is directed to interoceptive, motor, affective, or cognitive content layers in the human self-model. On the level of philosophical theory, it is this kind of phenomenally experienced introspection that underlies classical theories of *inner perception*, like that presented by John Locke or Franz Brentano. And, to come back to Fred Dretske's original point, Introspection$_3$ could of course be a process that sometimes is transparent. On the other hand, it could also be opaque or unconscious.

Introspection$_4$ ("consciously experienced cognitive self-reference")

This type of introspection is a conceptual (or quasi-conceptual) kind of meta-representation, again operating on a pre-existing, coherent self-model (Metzinger, 2003b). Phenomenal representational processes of this type generate conceptual forms of self-knowledge by directing cognitive processes towards certain aspects of internal system states, the intentional content of which is being constituted by a part of the world depicted *as internal*.

The general phenomenology associated with this type of representational activity includes all situations in which we consciously think about ourselves *as* ourselves (i.e., when we think what some philosophers call I*-thoughts; Baker, 1998). On a theoretical level, this last type of introspective experience clearly constitutes the case in which philosophers of mind have traditionally been most interested: the phenomenon of *cognitive self-reference* as exhibited in reflexive self-consciousness. In alluding to Dretske (1969) we might perhaps speak of "epistemic introspection".

Obviously the first two notions of introspection, namely introspective availability, are rather trivial; they define the internality of potential objects of introspection entirely by means of a simple physical concept of internality. In the present context, internality *as phenomenally experienced* is of a higher relevance. However, it may be useful to remember the principle of local supervenience for phenomenal content, which is highly plausible on empirical grounds: For *all* forms of mental representation referred to above it is true that their phenomenal content is fixed as soon as all internal and contemporaneous physical properties of the respective system are fixed. In this weaker sense, all phenomenal experience is a form of introspection.

Now we can say that, if our hypothetical Dretskean "introspective scanner" (Dretske, 1995, p. 61) were a transparent mechanism, then we would make the prediction that for the case of Introspection$_1$ and Introspection$_3$ the first-order target states would appear as mind-independent and directly given. We also immediately see that this point only partly holds for Introspection$_3$: Some target states (e.g., thought chains) are slow enough that we become aware of the fact that this is an introspective form of self-representation, which might in principle involve misrepresentation.

5.2 Metarepresentational theories of consciousness, displaced perception, and the fallacy of the representational divide

In the philosophical tradition, consciousness has often been understood as a type of meta-level, higher-order knowledge: a form of inner knowledge which can accompany mental processes. The English word "conscience" is derived from the Latin *conscientia*, which originally meant "jointly knowing", "knowing together with" or co-awareness, but also consciousness and conscience (Metzinger, 2010). Interestingly, throughout most of the history of philosophy, consciousness had a lot to do with conscience (i.e., a higher-order *moral* judgement of one's own actions and inner states), with Descartes then separating "conscience" and "consciousness", thereby constituting the modern concept of consciousness in the 17th century. What metarepresentational theories of consciousness have in common is that they attempt to construe phenomenal content as second-order intentional content. This line of thought is also used by some contemporary approaches which analyse phenomenal consciousness as a specific type of meta-representation. The basic thought is often that consciousness arises when the inner states of a system which already carry representational content *in turn* become the content of higher-order meta-representational states. However, there remains the basic problem of linking this terminology to fine-grained phenomenological descriptions of introspective experience and our everyday folk-psychological vocabulary with its strong Cartesian bias. If they are formulated in the latter form at all, such theories of consciousness typically take two forms: theories of higher-order *perception,* and theories of higher-order *thought*. In Western philosophy of consciousness, a line of inner-perception models runs from Aristotle to John Locke, Immanuel Kant, William James and Franz Brentano, and resurfaces in the works of authors such as David Armstrong, Peter Carruthers, Paul

Churchland and William Lycan. What these thinkers have in common is the suggestion that consciousness may be a special kind of inner perception: the perception of one's own mental processes.

Fred Dretske's theory of phenomenal transparency is a highly original variation of this strategy that applies his theory of displaced perception to the problem of introspective knowledge.

As such, it is a theory of *object representation*:

> Perceptual displacement – seeing that *k* is F by seeing, not *k*, but some other object, *h*, occurs when there is conceptual, but no corresponding sensory, representation of *k.* (Dretske, 1995, p. 42)

This theory of object representation is then applied to the special case of introspective knowledge. The object represented is an "experience":

> If, then, introspective knowledge is a species of displaced perception, it is an instance in which an experience (of blue, say) is conceptually represented as an experience of blue via a sensory representation not of the experience, but of some other object. [...] Introspective knowledge of E requires no other sensory representation of objects than those already being represented by E – the experience one comes to know about. (Dretske, 1995, pp. 42–43)

According to Dretske, the first object E is "conceptually" represented, with the help of a *sensory representation* of a second object. This creates an obvious tension as it leaves the true nature of Dretske's "transparent scanner" open. It is also unclear both whether the first-order "experience" is phenomenal, and what makes it conscious in the first place. While the general idea seems very clear – introspective knowledge is an inference from a first-order perceptual belief *de re* (a combination of introspection$_3$ and introspection$_4$) – it now becomes difficult to specify the precise content of the relevant connecting belief in a non-circular way and to later justify it as a kind of introspective *knowledge* (for a very lucid analysis see Aydede, 2003, § 2).

The problem can be made very clear from a related perspective, by taking a second look at what Güven Güzeldere says about the ambiguity of all metarepresentational theories of consciousness based on the classical vehicle/content distinction and employing folk-psychological terms like "perception" or "belief":

> What does `being conscious of one's belief' mean? Being conscious of the *content* of that belief, being conscious of t*he fact that one has a belief* with the content that it has, or being conscious of *the belief state qua the vehicle* that it is (i.e. the mental state that does the representing, not what it represents)? (Güzeldere, 1995, p. 340)

Even if experiences could be "objects" of a transparent inner form of perception, by "conceptually" representing them *as* objects, the general problem would remain. Recall Güzeldere's notion of the "fallacy of the representational divide":

> The ‘fallacy of the representational divide’ [… is] a tacit attempt to replace what is being represent-*ed* with that which is the represent-*er.* […] In other words, no matter how much we find out about the intrinsic properties of representational states, we may simply not be able to reach the other side of the ‘representational divide’, in virtue of this alone, and get to the extrinsic, relational properties of those states. (Güzeldere, 1995, p. 350)

Theories of inner perception, “displaced” or not, have undisputable advantages: They can explain the existence of unconscious first-order perceptual states, and they provide us with an intuitive model of introspection which possess great *prima facie* phenomenological plausibility, a model which therefore strongly resonates with our folk-psychological idiom, our accustomed way of speaking about our own phenomenal experience. But why should the causal properties of a first-order intentional state change just because it becomes the object of a second-order intentional state via displaced perception? Conscious percepts, whether mediated by a Dretskeian “transparent scanner” or not, clearly have a very different functional profile from unconscious perception. In addition, this approach generates the “inner sense”-problem: What exactly is the sensory *modality* which, due to its transparency, makes no contribution to the phenomenal character of the first-order state except making it “phenomenal”? If it episodically became opaque, what exactly would the contribution be? If the content of my first-order state were “blueness”, but a *misrepresenting* (Dretske, 1986) transparent scanner would erroneously represent it as red in a displaced manner, what exactly would be the content of experience? We can all experience hallucinations, but – according to Dretske’s model – could we also hallucinate experiences?

5.3 The impossibility of natural self-indicators

Andreas Kemmerling, in a substantial and careful criticism of Fred Dretske’s theory of introspective self-knowledge, shows how the analogy with displaced perception does not really work, and demonstrates that a naturalist theory of meta-representation presupposing objective dependency relations, indicator functions, etc. faces fundamental difficulties (Kemmerling, 1999). For example, if the experience a conscious subject has in perceiving a shirt as blue at the same time indicates to the subject *that* it is representing it as blue, then this creates another type of tension. As Kemmerling points out:

> [T]he indicator and the indicatum, in this case, would have to be identical. (But this is an undesired result. For we should either accept that, for analytical reasons, nothing is a natural sign of itself [...], or concede that something may indeed indicate itself, yet only in a special and peripheral sense. (It is hard to see how one natural sign could be an indicator of itself without every natural sign being a self-indicator; self-indication if we allowed for it would be inevitably universal in the realm of the natural, and in this sense would be trivial.) (Kemmerling, 1999, p. 316)

If we proceed to the idea of a “connecting belief”, we now need the assumption that something could at the same time be a meta-representation of the content of some first-order perceptual state *and* an indicator of whatever this state refers to – and that the two functional roles could never be dissociated. But, in a naturalistic world containing natural brains, this may of course happen. If we call the example of the shirt's being blue "x", Kemmerling says the following about the connecting belief:

> As an x-meta-representation, its job is not to track x at all. As a meta-representation, it doesn't focus on x; it (virtually) focusses on something else, namely r_0's representing x, – something which may well not be in phase with x. [...]
>
> Nothing can be a natural x-meta-representation and a natural x-representation at the same time. For the functional roles definitive of an x-meta-representation and an x-representation diverge. There are possible cases in which the alleged meta-representation is bound to say "Yes", but the relevant object-state does not represent x. (Kemmerling, 1999, p. 323)

Recently, Frank Hofmann proposed a new model based on transparent *mental demonstration* which avoids rationality constraints, the mysteries of direct "acquaintance" with "experiences" as such (whatever that may be or presuppose), or positing an inner sense (Hofmann, 2018). The core idea is that "mental pointing" is not a relation between a subject and a mental state or "experience", but rather a relation between the subject and a type-level content element of this experience. Of course, it is highly implausible that one should not be able to directly demonstrate one's experience *as such*, to mentally point to an experience *as an experience* (and not only to an appearing property, by forming a mental "property demonstrative") – the practical phenomenology developed in non-Western philosophical traditions has shown this possibility many centuries ago (Hasenkamp et al., 2012; Hölzel et al., 2011; Lutz, Slagter, Dunne, & Davidson, 2008). Nevertheless, Hofmann's new model of not pointing to the mental state itself but only to an aspect, to a specific content *property*, is a new idea that relates directly to the critical discussion above, because it could in principle yield phenomenal transparency:

> Mental demonstration is 'transparent', if you like. It does not get a grip on the experience itself, but 'goes through' to its content. (Hofmann, 2018)

In principle, Hofman's model could solve the problem with Dretske's "connecting belief" (as criticized by Aydede, 2003; Kemmerling, 1999), namely, by introducing a demonstrative judgement into a more elaborate hypothetical cognitive architecture that then adds as further building blocks both ordinary conceptual classification and relevant background-knowledge (instead of the classical "connecting belief"). Here is how Hofmann describes the relation to Dretske's original idea:

> How about Fred Dretske's model of *displaced perception*? – A certain description of how we can acquire self-knowledge seems to be true of both the model of displaced perception and the demonstrative transparency model: 'We come to know about one thing by knowing some other thing in the first place.' Just as we come to know that the gas tank is empty by knowing that the gas gauge has a certain reading, we can come to know what we are experiencing by knowing (recognizing) of a content element of our experience that it is of such and such a kind. This is of course true of both models, since both models assume that there is a certain kind of transition or inference. But, of course,

> the demonstrative model has a quite distinctive component which is lacking in Dretske's model of displaced perception: mental pointing and demonstrative reference. We start from a demonstrative judgment based on mental pointing, and this is a crucial ingredient in gaining introspective self-knowledge. (Hofmann, 2018)

This is one example of a recent, and more specific proposal, but it also lacks any connection to empirical research about attentional processing – and therefore we do not learn anything about the inner *nature* of mental demonstration. Given the additional fact that other traditional philosophical theories of mind directly contradict this hypothetical model from a *phenomenological* perspective, along with the obvious problems of Dretske's original conception I sketched above, it now makes sense to ask if there are any other conceptual tools that are in line with Fred Dretske's general strategy, but which might help us in gaining a fresh perspective on the problem of phenomenal transparency.

6. New conceptual instruments: Transparency under predictive processing

Currently, the information-theoretical approach most intensively discussed by philosophers of mind is the so-called "predictive processing" (PP) approach.[3] The statistical/information-theoretic approach to mental representation pioneered by Fred Dretske is currently reaching a new level of maturity. Many new fine-grained and extendable conceptual tools have become available. The new, slowly emerging version of the original project of an "information-theoretic cognitive semantics" is now firmly grounded in a wealth of neuroscientific data, and it simultaneously opens a new perspective on the phenomenology *and* epistemology of perception.

Under the PP approach – which many philosophical authors interpret as representationalist (see Gładziejewski, 2016; Kiefer & Hohwy, 2018; Wiese, 2017b, 2018; Williams, 2018) – there is one information-theoretic notion, which is of maximal importance: "variational free energy" (Friston, 2010). In generating mental properties and intelligent behaviour, all the human brain ever does is to minimize variational free energy, or prediction error (Clark, 2015; Hohwy, 2013). PP can therefore count as the latest continuation of Fred Dretske's pioneering work in integrating statistical information theory and philosophy of mind (Dretske, 1981).

Central philosophical intuitions like "displaced perception" and higher-order "belief states" reappear in the new mathematical framework as the idea of using an internal model to infer a hidden cause or as the notion of a "hierarchical generative model", as PP posits a hierarchy of estimators which operate at different spatio-temporal timescales in order to track features at different scales (Wiese, 2017b). But there are differences and refinements too: For example, the hierarchy does not necessarily have a top level, but it might have a centre, with representational levels as rings on a disc or a sphere (Wiese & Metzinger, 2017, p. 14). "Beliefs" are no longer seen as relations between persons and propositions, but rather as probabilistic models appearing in self-organizing biological systems, conditional expectations implemented in a dynamic, subsymbolic format. In the new framework, the vehicle-content distinction is still a highly useful

[3] For the first peer-reviewed, edited collection of texts on this topic, see Metzinger and Wiese 2017; for an accessible introduction see Wiese and Metzinger 2017, all material is freely available at predictive-mind.net. See also open-mind.net for further relevant contributions.

conceptual instrument, but it applies to complex dynamical processes extended in time. In classical analytic philosophy of mind this distinction often contains subtle residues of Cartesian dualism because it tempts us to reify the vehicle and the content, conceiving of them as ontologically distinct, independent entities. A more empirically plausible model of representational content now describes "content" as an aspect of an ongoing process, an embodied generative model, and not as some kind of abstract object.

For reasons of space I cannot go deeper into any of these issues and instead point readers to the relevant introductory literature (Clark, 2015; Hohwy, 2013; Metzinger & Wiese, 2017a). Instead, I will use the remainder of this chapter to very briefly offer three examples of more recent treatments of the concept of "phenomenal transparency", one of them being a general model while the two others are specific applications of the PP approach to two distinct phenomenological domains.

6.1 Model 1: Unavailability of earlier processing stages

Transparency is a special form of darkness. Taking the phenomenology of visually experienced transparency as an example reveals that we are unable to see something *because* it is transparent. We don't look *at* our glasses, but *through* our glasses. We don't see the window, but only the bird flying by. Importantly, the negative fact that we don't see the medium, the window, is itself not *explicitly* represented in the seeing process itself.

Transparency is a property of phenomenal representations. Importantly, phenomenal representations are not things, but processes. They have a beginning and an end. As they are chains of information-processing events, they also have *stages* – there are earlier and later stages of processing – model evidence changes over time, and if all goes well prediction error is eventually minimized and ambiguity is resolved into a disambiguated final solution. Some representational processes are faster than the temporal resolution of attention, some are slower.

Phenomenal transparency in general, however, also means that something particular is not accessible for subjective experience, namely the *representational* character of the contents of conscious experience. This analysis refers not only to all sensory modalities and to our integrated phenomenal model of the world as a whole in particular, but also to large parts of our self-model. The *instruments* of representation themselves cannot be represented as such anymore, and hence the system making the experience, on this level and by conceptual necessity, is entangled into a robust form of naïve realism. This happens, because, necessarily, it must now experience itself as being in direct contact with the current contents of its own consciousness. What precisely is it that the system cannot experience? What is inaccessible to conscious experience is the simple fact of this experience taking place in a *medium*. Therefore, transparency of phenomenal content leads to a further characteristic of conscious experience, namely the subjective impression of immediacy and mind-independence. The possibility of misrepresentation is not given.

On the functional level of analysis, what makes phenomenal models transparent is the *attentional unavailability for introspection of earlier processing stages in the brain*. The more earlier processing stages, the more aspects of the internal construction process leading to the final, explicit and disambiguated phenomenal content are available for introspective attention, the more the system will be able to recognize these phenomenal states *as* internal, self-generated constructs. Full transparency means full attentional unavailability of earlier processing stages. Degrees of opacity therefore come as degrees of attentional availability. In many processes of phenomenal representation, the degree of opacity vs. transparency varies and fluctuates over time (cf. Extension 1, as explained in sections 4.1 and 4.2 above). I have explained Model 1 elsewhere (Metzinger, 2003a, 2003b, 2014) and will not go into any further detail here. Let us rather look at two other, more recent and more specific applications of an information-processing approach based on physical statistics.

6.2 Model 2: Counterfactual richness

For the domain of perceptual objects, the problem of transparency can be reformulated as the problem of accounting for the phenomenal property of *subjective veridicality* characterizing perceptual content, or, more simply, the fact that objects of perception are experienced as *real*. "Subjective veridicality" has nothing to do with doxastic veridicality; it can occur in the complete absence of states traditionally described as "beliefs" or conceptually analysed as propositional attitudes. Synesthetic concurrents (cf. section 4.1) are one example of phenomenal opacity, i.e., of often highly vivid perceptual content where the relevant phenomenal property of "realness" is selectively missing (Grossenbacher & Lovelace, 2001; Mroczko, Metzinger, Singer, & Nikolić, 2009). One much-discussed philosophical theory of subjective veridicality or "perceptual presence" explains it as fundamentally action-related, namely as the practical mastery of complex dependency-relations between sensory and motor content activated in skilful interactions between body and world; it results from the mastery of so-called "sensorimotor contingencies" (Noë, 2004, 2004; O'Regan & Noë, 2001).

Anil Seth has developed an alternative model of "subjective veridicality" under PP:

> Here, I describe a theory of predictive perception of sensorimotor contingencies which (i) accounts for perceptual presence in normal perception, as well as its absence in synaesthesia, and (ii) operationalizes the notion of sensorimotor contingencies and their mastery. The core idea is that generative models underlying perception incorporate explicitly counterfactual elements related to how sensory inputs would change on the basis of a broad repertoire of possible actions, even if those actions are not performed. These "counterfactually-rich" generative models encode sensorimotor contingencies related to repertoires of sensorimotor dependencies, with counterfactual richness determining the degree of perceptual presence associated with a stimulus. While the generative models underlying normal perception are typically counterfactually rich (reflecting a large repertoire of possible sensorimotor dependencies), those underlying synaesthetic concurrents are hypothesized to be counterfactually poor. (Seth, 2014, p. 97)

If we follow Seth, conscious representations often possess a critical property which I call "counterfactual depth". The idea behind Model 2 is that exactly this property is what determines phenomenal transparency or the experienced “realness” of a representation’s content. The representations themselves are hierarchical generative models (HGMs), predicting sensory input and resolving prediction errors by continuous Bayesian updating:

> The key phenomenological property of the subjective non-veridicality of synaesthetic concurrents can now be explained by appealing to differences between the counterfactual richness of the HGMs associated with inducers and concurrents. For inducers (and in normal perception), perceptual content depends on counterfactually-rich HGMs of the behaviour of hidden causes of fictive sensory signals in response to hidden controls relating to possible actions. In contrast, for concurrents, the corresponding HGMs are hypothesized to be counterfactually-poor because the hidden causes giving rise to concurrent-related sensory-signals do not embed a rich and deep statistical structure for the brain to learn (i.e., to encode into counterfactual representations). (Seth, 2014, p. 108)

If we want to take human phenomenology as seriously as possible and understand why synesthetic concurrents, although vivid and rich in phenomenal character, lack the relevant property of “perceptual presence”, then this gives us a new conceptual model. This model

> proposes that the generative models underlying perceptual content incorporate counterfactual probabilistic representations of hidden causes linking fictive sensory signals (and their expected precisions) to possible actions. This gives mechanistic focus to the notion of mastery of sensorimotor contingencies as described within sensorimotor theory, and explains subjective veridicality or “perceptual presence” in terms of the counterfactual richness of the underlying generative models. Applied to synaesthesia, the theory accounts for the subjective non-veridicality of concurrents by counterfactual poverty: concurrents are experienced as lacking in perceptual presence because of counterfactually impoverished generative models of how hidden causes would modify sensory signals (and associated precisions) in response to actions. This is because, as compared to inducers (or the world generally), there is no corresponding rich world-related statistical structure for any such model to learn. (Seth, 2014, p. 115)

One may certainly ask if the property explained by this approach is not “objecthood” rather than realness *per se* (Froese, 2014), or if not just counterfactual depth, but the hierarchical depth of generative models *itself* may account for phenomenal transparency and the “subjective veridicality” of perceptual presence (where hierarchically deep models represent objects as being highly invariant hidden causes of sensory signals, see Hohwy, 2014b, p. 128). Perhaps the phenomenal quality of perceptual “realness” is an expression of higher-order invariance? However, the relevant point is that there is now an empirically grounded mathematical framework in existence (Friston, 2010; Limanowski & Friston, 2018) providing philosophers of mind with a more fine-grained conceptual scheme under which Fred Dretske’s original line of research can be pursued further, and without employing folk-psychological notions like “belief”, “perception”, “action”, or “feelings”. This framework originates exactly from the two additional elements Dretske introduced into analytic philosophy of mind: statistics and information theory.

6.3 Model 3: The example of transparent beliefs about action policies in temporally thick generative models

Jakub Limanowski and Karl Friston have recently offered a more detailed model of what introspective attention really is (Limanowski & Friston, 2018). However, their proposal does not aim at the phenomenal "presence" of perceptual objects discussed above. Instead, it concentrates on the role of self-consciousness in agency. They focus on a specific domain: the phenomenology of agency, in particular of action *initiation*. Under the PP approach, actions are "self-fulfilling movement prophecies": To enable overt bodily movement, precision estimates associated with sensory prediction errors must be cancelled out by top-down modulation. One could describe this as a process of attending away from somatosensory signals by selectively increasing precision estimates by introspective attention (Limanowski, 2017). In action generation, it is a deployment of 'precision' that may render the perceptual evidence (for action) opaque, while turning transparency into a necessary aspect of beliefs about action. Interestingly, the same principle could hold for mental actions like the control of introspective attention (Metzinger, 2017). It is therefore tempting to ask if this could help us reformulate Dretske's idea of a "transparent introspective scanner" on a mathematical level. Here is what Jakub Limanowski and Karl Friston say about the general computational background, action generation as a form of active inference, and the relationship between attention and consciousness:

> Active inference can be situated within a larger 'free-energy principle,' along which any living system will actively try to remain in a set of unsurprising states by performing inference; i.e., model selection and inversion (Friston, 2010; cf. Hohwy, 2013; Clark, 2015). The models in play here are *generative models* – probabilistic (predictive) mappings from causes (e.g., latent or hidden states of the world) to consequences (e.g., sensory observations, Friston et al., 2017a). If stacked, they yield a deep or hierarchical generative model (HGM), in which higher levels contextualize lower levels, and lower levels provide evidence for higher levels. In this scheme, free energy minimization corresponds to maximizing Bayesian model evidence, which implies a notion of 'self-evidencing' (i.e., a Bayes-optimal model – a free energy minimizing agent – will always try to maximize evidence for itself, Hohwy, 2016; Kirchhoff et al., 2018). [...]
>
> Under a popular algorithmic scheme known as predictive coding (Srinivasan et al., 1982), free energy ('surprise') is approximated in the form of precision-weighted prediction error signals, which are passed from lower to higher levels to update the model's 'beliefs' about its environment, which in turn issue predictions to suppress prediction errors in lower levels (Friston, 2010). *Note that we use the term 'belief' to refer to a conditional expectation (i.e., probabilistic representation) encoded by neuronal activity, rather than in the folk-psychological sense* [emphasis added]. This hierarchical scheme of recurrent message passing notably implies that increasingly higher-level beliefs represent increasingly abstract states of the environment at increasingly broad time scales. In such deep architectures, balancing the relative dominance of prior beliefs or sensory evidence (i.e., prediction errors) across the entire hierarchy is accomplished by weighting the ascending prediction errors by their precision. This means precision has to be estimated and deployed (at each level of the hierarchy): a process that is equated with attention (Feldman and Friston, 2010).

> Perceptual inference, under these principles, associates conscious experience with the 'explanation' for the sensorium that minimizes prediction error throughout the hierarchy (Hohwy, 2013; Seth et al., 2016). (Limanowski & Friston, 2018, p. 2)

Can anything be said about phenomenality as such? According to Karl Friston the level of conscious processing is marked out by "temporal thickness", i.e. an explicit representation of distant, fictive temporal horizons (for an in-depth philosophical treatment, see Wiese, 2017a). Temporally thick representations are also counterfactually deep and – if viewed as dynamic system states playing a crucial causal role in action generation – they continuously connect the present with the organism's proximate future:

> A living system cannot infer the consequences of its action unless it embodies a model of the future. This follows from the simple fact that the arrow of time requires the consequences of action to postdate action ***per se***. This is important because it means that the (generative) models capable of inferring the consequences of action must necessarily endow inference with **temporal thickness**. (Friston, 2018, p. 5)

Friston thinks that there may be a mapping between levels of consciousness and the temporal thickness of our representation of the proximate past and future. He points out:

> Technically, in hierarchical generative models there is usually a one-to-one mapping between the temporal thickness or extent and hierarchical depth. In other words, higher levels of a hierarchical model generally represent sequences or trajectories with a greater temporal span. (Friston, 2018, p. 7)

This would mean that phenomenality as such is graded, that a perceptual object or multimodal scene becomes more "present" as the system manages to increase its counterfactual richness. Let us now recall the epistemological perspective on attention introduced in section 6.1, where attention is a form of second-order statistics – the optimization of precision expectations. Computationally as well as epistemologically, this will be true for the deployment of attention that is subjectively experienced as being directed at an external object (e.g., in perception or intospection$_1$), as well as for the process of attending to something subjectively experienced as mental (introspection$_3$) (Limanowski & Friston, 2018).

> Recall that according to SMT [i.e., the self-model theory of subjectivity], the defining characteristic that disambiguates phenomenally transparent and opaque representations is that the construction of opaque representations is amenable to *introspective attention*, an inward-directed form of resource allocation onto specific parts within my internal reality-and-self-model (Metzinger, 2003). This distinction affords a simple formulation in terms of active inference, where attention is mediated by assigning greater or lesser precision to prediction errors at various levels of hierarchical processing. Importantly, this precision itself has to be predicted; implying that we have (first-order) representations of the (second-order) precision of hierarchically subordinate prediction errors. From the perspective of predictive coding, this means that we also have to infer the deployment of precision which, in a hierarchical setting, starts to look like attention (Feldman and Friston, 2010). In

> enactive formulations of predictive coding (i.e., active inference), descending predictions prescribe action. In terms of descending (first-order) predictions of content, this is usually cast as controlling motor (and autonomic reflexes) through descending proprioceptive or interoceptive predictions, respectively (Adams et al., 2013a; Seth and Friston, 2016). However, we can apply exactly the same principles to *descending predictions of precision* and thereby understand the active deployment of precision weighting as a form of 'mental action' that has exactly the look and feel of introspective attention. This argument rests on an assumed similarity of introspective and 'perceptual' attention (as implied by most transparency accounts, e.g., Harman, 1990; Metzinger, 2003); consequently, introspective attention is seen as a special case of the general mechanism of precision estimation applied to conscious mental representations. (Limanowski & Friston, 2018, p. 3)

Limanowski and Friston suggest that transparency is a necessary aspect of beliefs about action that entail introspective attention, while the precision expectations that "underwrite" introspective attention have the capacity to render the perceptual evidence for action opaque (Limanowski & Friston, 2018, p. 6). If they are right, this would entail an important conclusion on a conceptual level: There is a minimal phenomenal experience of selfhood in agency which, in neurotypical human beings, is phenomenally transparent *by necessity*. In other words, there is a class of beliefs (with "belief" here used in the evidence-based sense of "probabilistic neural representation") that cannot be rendered opaque due to the architecture of the human mind.

> [T]here is one fundamentally important set of posterior beliefs that are both privileged and impoverished – in the sense that they can never be subject to introspective attention. These are the beliefs about policies or sequences of (overt and mental) action that gather evidence from lower (perceptual) levels of hierarchical processing, because expectations about precision are an inherent part of a policy. Heuristically, this means placing confidence in the consequences of action is an inherent part of the policies from which we select our actions. This implies that we cannot, literally, *place* confidence in our policies. In other words, if the genesis of expectations about precision is, in and of itself, entailed in a (mental) action, *beliefs about action cannot be subject to introspective attention*. In other words, posterior beliefs about action are causes, not consequences, of introspective attention (and other actions). This suggests that beliefs about 'what I am' doing are unique, in the sense that they are necessarily transparent. This fits comfortably with SMT – in that these beliefs are inherently about the self and how the self is acting on the world.
>
> One can unpack this argument further and identify examples of transparency that, in virtue of being prescribed by overt or covert action – can never become opaque. A nice example of this is the deployment of sensory precision during sensory attenuation (Brown et al., 2013; Limanowski, 2017; Wiese, 2017c). This is the converse of attention and an important aspect (on the current account) of mental action. [. . .] In short, beliefs about action whether overt or covert (attentional) are necessarily transparent – and are realized by active sampling of the sensorium that has transparency 'written into it.' (Limanowski & Friston, 2018, p. 4)

Self-consciousness is not a necessary condition for consciousness to occur (pace Friston, 2018), but self-*representation* may certainly be a conceptually necessary aspect of biological self-organization. Accordingly, there may be naturally evolved, conscious forms of biological self-representation (e.g., Introspection$_1$), which do not instantiate the phenomenal property of selfhood or any form of self-consciousness. Genuine self-consciousness only appears together with a transparent phenomenal self-model. Self-modeling (Metzinger, 2003a) is an important and highly successful biological process, but it must never lead a system operating with limited neurocomputational resources into infinite regressions and endless internal loops – this would endanger the survival of the system. Having an internal model of yourself as an agent is a relevant special case. As I have pointed out (Metzinger, 2003a, 34, n. 19; 583), one possible solution is that the brain has developed a functional architecture which stops iterative but computationally necessary processes – like the potentially infinite recurrent mental representation of agentive self-control – by *object formation*, i.e., by creating a transparent self-representation which "reifies" a potential infinity of computational steps. If Limanowski and Friston are right, in the special case of active inference, action initiation, and the representation of one's own policies, this representational object may well be what – as phenomenologists – we later (falsely) call a "self" (cf. Wiese, submitted).

Conclusion

What can we say about Fred Dretske's theory of introspective self-knowledge and phenomenal transparency? While his specific proposal of an inferential model did not stand, the idea that the *mechanism* of introspection itself might not instantiate any phenomenal properties is still valid. And while the criticism of many existing higher-order perception models (for example those of Locke, James, Brentano, Armstrong, Carruthers, Churchland and Lycan) and of a too coarse-grained and overly modularistic strategy for transparent meta-representation (Aydede, 2003; Güzeldere, 1995; Kemmerling, 1999) remains convincing, we should not overlook the heuristic fecundity and sustained impact Fred Dretske's theorizing has had on current work in philosophy of mind.

Fred Dretske introduced information theory and a rigorous statistical perspective into modern philosophy of mind. These approaches remain at the centre of contemporary philosophy of mind (Clark, 2015; Hohwy, 2013; Metzinger & Wiese, 2017b). Intentionalism about phenomenal states remains a viable option, because we can instrumentally describe the epistemic states generated by precision-controlled hierarchical Bayesian updating in the human brain as probabilistic *models* (or even "beliefs") (Gładziejewski, 2016; Kiefer & Hohwy, 2018; Metzinger, 2017; Wiese, 2017b, 2018; Williams, 2018). In addition, refined metarepresentational approaches abound today, but the perception/cognition/attention/action divide has been conceptually dissolved on a formal level (Friston, 2010; Wiese, 2018, pp. 205–207) and most theoreticians now work with a subpersonal notion of "inference". Unfortunately, the epistemological dimension has only just begun to receive the attention it deserves (see for example Hohwy, 2014a, 2017). Fred Dretske's interdisciplinary approach to philosophy of mind, along with large portions of his pioneering metatheoretical work, is now gradually coming to fruition, and it is exciting to await the insights his work may yet inspire.

Bio

Thomas Metzinger (*1958 in Frankfurt am Main, Germany) is currently Professor of Theoretical Philosophy at the Johannes Gutenberg-Universität Mainz, Adjunct Fellow and Director of the MIND Group at the Frankfurt Institute for Advanced Study (FIAS). Metzinger is past president of the German Cognitive Science Society (2005-2007) and of the Association for the Scientific Study of Consciousness (2009-2011), and a member of the European Commission's High-Level Group on Artificial Intelligence. In English, he has edited two collections on consciousness *("Conscious Experience"*, Imprint Academic, 1995; *"Neural Correlates of Consciousness"*, 2000) and published one monograph (*"Being No One – The Self-Model Theory of Subjectivity"*, MIT Press, 2003. Important recent Open Access collections (2015) are *Open MIND* at http://www.open-mind.net and *Philosophy and Predictive Processing* at http://predictive-mind.net (2017).

Abstract

This contribution investigates Fred Dretske's conception of phenomenal transparency is still valid, and whether it can be refined and extended into the present debate. The main claim is that his model of introspective self-knowledge is untenable as it stands, but that the conceptual foundations provided by his information-theoretic and representationalist approach to the conscious human mind still possess considerable relevance and fecundity. It also offers two specific examples of how the problem of transparency can be treated using more recent conceptual tools in the specific domains of perceptual presence and action generation.

References

Anton, G. (1898). Hans Förstl. *Wiener Klinische Wochenschrift*, *11*, 227–229.

Armstrong, D. M. (1968). *A materialist theory of the mind*: Routledge.

Aydede, M. (2003). Is introspection inferential? In B. Gertler (Ed.), *Privileged access: Philosophical accounts of self-knowledge.* Aldershot: Ashgate.

Baker, L. R. (1998). The first-person perspective: A test for naturalism. *American Philosophical Quarterly*, *35*(4), 327–348.

Benson, D. F., & Greenberg, J. P. (1969). Visual form agnosia: A specific defect in visual discrimination. *Archives of Neurology*, *20*(1), 82–89.

Carruthers, P. (2017). Mindreading in adults: Evaluating two-systems views. *Synthese*, *194*(3), 673–688. https://doi.org/10.1007/s11229-015-0792-3

Churchland, P. M. (1989). *A neurocomputational perspective: The nature of mind and the structure of science*. Cambridge, MA: MIT Press.

Clark, A. (1989). *Microcognition*. Cambridge, MA: MIT Press.

Clark, A. (1998). *Being there: Putting brain, body, and world together again*: MIT Press.

Clark, A. (2015). *Surfing uncertainty: Prediction, action, and the embodied mind*: Oxford University Press.

Dretske, F. (1981). *Knowledge and the Flow of Information*. Stanford, CA: CSLI.

Dretske, F. (1986). Misrepresentation. In R. Bogdan (Ed.), *Belief: Form, content, and function.* Oxford: Clarendon Press.

Dretske, F. (1988). *Explaining behavior: Reasons in a world of causes* (Sirsi) i9780262040945): Cambridge Univ Press.

Dretske, F. (1995). *Naturalizing the mind*. Cambridge, MA: MIT Press.

Dretske, F. (1999). The mind's awareness of itself. *Philosophical Studies*, *95*(1-2), 103–124.

Fox, K.C.R., & Christoff, K. (Eds.). (2018). *The Oxford Handbook of Spontaneous Thought: Mind-wandering, Creativity, Dreaming, and Clinical Conditions*. New York: Oxford University Press.

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138. https://doi.org/10.1038/nrn2787

Friston, K. (2018). Am I Self-Conscious? (Or Does Self-Organization Entail Self-Consciousness?). *Frontiers in Psychology*, *9*, 579. https://doi.org/10.3389/fpsyg.2018.00579

Froese, T. (2014). Steps toward an enactive account of synesthesia. *Cognitive Neuroscience*, *5*(2), 126–127.

Gładziejewski, P. (2016). Predictive coding and representationalism. *Synthese*, *193*(2), 559–582. https://doi.org/10.1007/s11229-015-0762-9

Grossenbacher, P. G., & Lovelace, C. T. (2001). Mechanisms of synesthesia: Cognitive and physiological constraints. *Trends in Cognitive Sciences*, *5*(1), 36–41.

Güzeldere, G. (1995). Is Consciousness the Perception of What Passes in One's Own Mind? In T. Metzinger (Ed.), *Conscious experience.* Thorverton, UK: Imprint Academic.

Harman, G. (1990). The intrinsic quality of experience. In W. C. Lycan & J. Tomberlin (Eds.), *Philosophical perspectives, vol. 4: Action theory and philosophy of mind* (pp. 31–52).

Hasenkamp, W., Wilson-Mendenhall, C. D., Duncan, E., & Barsalou, L. W. (2012). Mind wandering and attention during focused meditation: A fine-grained temporal analysis of fluctuating cognitive states. *NeuroImage*, *59*(1), 750–760.

Hellie, B. (2007). That Which Makes the Sensation of Blue a Mental Fact: Moore on Phenomenal Relationism. *European Journal of Philosophy*, *15*(3), 334–366. https://doi.org/10.1111/j.1468-0378.2007.00274.x

Hippel, W., & Trivers, R. (2011). The evolution and psychology of self-deception. *Behavioral and Brain Sciences*, *34*(1), 1.

Hofmann, F. (2018). How to know one’s experiences transparently. *Philosophical Studies*, 1–20.

Hohwy, J. (2013). *The predictive mind*. Oxford: Oxford University Press.

Hohwy, J. (2014a). The Self-Evidencing Brain. *Noûs*.

Hohwy, J. (2014b). Elusive phenomenology, counterfactual awareness, and presence without mastery. *Cognitive Neuroscience*, *5*(2), 127–128. https://doi.org/10.1080/17588928.2014.906399

Hohwy, J. (2017). How to entrain your evil demon. In T. Metzinger & W. Wiese (Eds.), *Philosophy and Predictive Processing.* Frankfurt am Main: MIND Group. https://doi.org/10.15502/9783958573048

Hölzel, B. K., Lazar, S. W., Gard, T., Schuman-Olivier, Z., Vago, D. R., & Ott, U. (2011). How does mindfulness meditation work? Proposing mechanisms of action from a conceptual and neural perspective. *Perspectives on Psychological Science*, *6*(6), 537–559.

Kemmerling, A. (1999). How Self-Knowledge Can't be Naturalized (Some Remarks on a Proposal by Dretske). *Philosophical Studies*, *95*(3), 311–328.

Kiefer, A., & Hohwy, J. (2018). Content and misrepresentation in hierarchical generative models. *Synthese*, *195*(6), 2387–2415. https://doi.org/10.1007/s11229-017-1435-7

Limanowski, J. (2017). (Dis-)Attending to the Body. In T. Metzinger & W. Wiese (Eds.), *Philosophy and Predictive Processing.* Frankfurt am Main: MIND Group. https://doi.org/10.15502/9783958573192

Limanowski, J., & Friston, K. (2018). 'Seeing the Dark': Grounding Phenomenal Transparency and Opacity in Precision Estimation for Active Inference. *Frontiers in Psychology*, *9*, 643. https://doi.org/10.3389/fpsyg.2018.00643

Lutz, A., Slagter, H. A., Dunne, J. D., & Davidson, R. J. (2008). Attention regulation and monitoring in meditation. *Trends in Cognitive Sciences*, *12*(4), 163–169.

Metzinger, T. (2003a). *Being No One: The Self-model Theory of Subjectivity*: MIT Press.

Metzinger, T. (2013a). The myth of cognitive agency: subpersonal thinking as a cyclically recurring loss of mental autonomy. *Frontiers in Psychology*, *4*, 931. https://doi.org/10.3389/fpsyg.2013.00931

Metzinger, T. (2013b). Why are dreams interesting for philosophers? The example of minimal phenomenal selfhood, plus an agenda for future research1. *Frontiers in Psychology*, *4.* https://doi.org/10.3389/fpsyg.2013.00746

Metzinger, T. (2015a). M-Autonomy. *Journal of Consciousness Studies*, *22*(11-12), 270–302.

Metzinger, T. (Ed.). (1995). *Conscious experience*. Thorverton, UK: Imprint Academic.

Metzinger, T. (2003a). *Being No One: The Self-model Theory of Subjectivity*: MIT Press.

Metzinger, T. (2003b). Phenomenal transparency and cognitive self-reference. *Phenomenology and the Cognitive Sciences*, *2*(4), 353–393.

Metzinger, T. (2008). Empirical perspectives from the self-model theory of subjectivity: a brief summary with examples. *Progress in Brain Research*, *168*, 215–278.

Metzinger, T. (2010). Bewusstsein [Consciousness]. In H. J. Sandkühler (Ed.), *Enzyklopädie Philosophie.* Hamburg: Meiner Verlag.

Metzinger, T. (2014). How does the brain encode epistemic reliability? Perceptual presence, phenomenal transparency, and counterfactual richness. *Cognitive Neuroscience*, *5*(2), 122–124. https://doi.org/10.1080/17588928.2014.905519

Metzinger, T. (2017). The Problem of Mental Action. In T. Metzinger & W. Wiese (Eds.), *Philosophy and Predictive Processing.* Frankfurt am Main: MIND Group. https://doi.org/10.15502/9783958573208

Metzinger, T. (2018). Why is mind wandering interesting for philosophers? In K.C.R. Fox & K. Christoff (Eds.), *The Oxford Handbook of Spontaneous Thought: Mind-wandering, Creativity, Dreaming, and Clinical Conditions.* New York: Oxford University Press.

Metzinger, T., & Walde, B. (2000). Commentary on Jakab's "Ineffability of qualia". *Consciousness and Cognition*, *9*(3), 352-62; discussion 363-9. https://doi.org/10.1006/ccog.2000.0463

Metzinger, T., & Wiese, W. (Eds.). (2017a). *Philosophy and Predictive Processing*. Frankfurt am Main: MIND Group.

Metzinger, T., & Wiese, W. (Eds.). (2017b). *Philosophy and Predictive Processing*. Frankfurt am Main: MIND Group.

Metzinger, T., & Wiese, W. (Eds.). (2017c). *Philosophy and Predictive Processing*. Frankfurt am Main: MIND Group.

Metzinger, T., & Windt, J. M. (2014). Die phänomenale Signatur des Wissens: Experimentelle Philosophie des Geistes mit oder ohne Intuitionen? In T. Grundmann, J. Horvath & J. Kipper (Ed.), *Die Experimentelle Philosophie in der Diskussion* (pp. 279–321). Suhrkamp.

Metzinger, T., & Windt, J. M. (2015). *What Does it Mean to Have an Open MIND?* Open MIND. Frankfurt am Main: MIND Group. Retrieved from https://open-mind.net/papers/general-introduction-what-does-it-mean-to-have-an-open-mind/at_download/paperPDF

Metzinger, T. (2003b). Phenomenal transparency and cognitive self-reference. *Phenomenology and the Cognitive Sciences*, *2*(4), 353–393.

Moore, G. E. (1903). The refutation of idealism. *Mind*, *12*(48), 433–453.

Mroczko, A., Metzinger, T., Singer, W., & Nikolić, D. (2009). Immediate transfer of synesthesia to a novel inducer. *Journal of Vision*, *9*(12), 25.

Noë, A. (2004). *Action in perception*: MIT Press.

O'Regan, J. K., & Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences*, *24*(05), 939–973. https://doi.org/10.1017/S0140525X01000115

Raffman, D. (1995). On the persistence of phenomenology. In T. Metzinger (Ed.), *Conscious experience* (pp. 293–308). Thorverton, UK: Imprint Academic.

Ratcliffe, M. (2008). *Feelings of being: Phenomenology, psychiatry and the sense of reality*. Oxford: Oxford University Press.

Reid, T., & Brookes, D. R. (1858). An inquiry into the human mind on the principles of common sense: A critical edition. In T. Reid, W. Hamilton, & D. Stewart (Eds.), *The Works of Thomas Reid, DD, Now Fully Collected, with Selections from His Unpublished Letters.* Maclachlan and Stewart.

Seth, A. K. (2014). A predictive processing theory of sensorimotor contingencies: Explaining the puzzle of perceptual presence and its absence in synesthesia. *Cognitive Neuroscience*, *5*(2), 97–118.

Shapiro, L. (Ed.). (2014). *The Routledge handbook of embodied cognition*: Routledge.

Shoemaker, S. (1990). Qualities and Qualia: What's in the Mind? *Philosophy and Phenomenological Research*, *50*, 109–131.

Van de Cruys, S. (2017). Affective Value in the Predictive Mind. In T. Metzinger & W. Wiese (Eds.), *Philosophy and Predictive Processing.* Frankfurt am Main: MIND Group. https://doi.org/10.15502/9783958573253

Vogeley, K., Bussfeld, P., Newen, A., Herrmann, S., Happé, F., Falkai, P., . . . Zilles, K. (2001). Mind reading: Neural mechanisms of theory of mind and self-perspective. *NeuroImage*, *14*(1 Pt 1), 170–181. https://doi.org/10.1006/nimg.2001.0789

Wiese, W. (2017a). Predictive Processing and the Phenomenology of Time Consciousness. In T. Metzinger & W. Wiese (Eds.), *Philosophy and Predictive Processing.* Frankfurt am Main: MIND Group. https://doi.org/10.15502/9783958573277

Wiese, W. (2017b). What are the contents of representations in predictive processing? *Phenomenology and the Cognitive Sciences*, *16*(4), 715–736. https://doi.org/10.1007/s11097-016-9472-0

Wiese, W. (2018). *Experienced Wholeness. Integrating Insights from Gestalt Theory, Cognitive Neuroscience, and Predictive Processing*. Cambridge, MA: MIT Press.

Wiese, W., & Metzinger, T. (2017). Vanilla PP for Philosophers: A Primer on Predictive Processing. In T. Metzinger & W. Wiese (Eds.), *Philosophy and Predictive Processing.* Frankfurt am Main: MIND Group. https://doi.org/10.15502/9783958573024

Williams, D. (2018). Predictive Processing and the Representation Wars. *Minds and Machines*, *28*(1), 141–172. https://doi.org/10.1007/s11023-017-9441-6

Windt, J. M., & Metzinger, T. (2007). The philosophy of dreaming and self-consciousness: What happens to the experiential subject during the dream state? In D. Barrett & P. McNamara (Eds.), *The new* *science of dreaming: Volume 3: Cultural and Theoretical Perspectives* (Vol. 3). Westport, CT & London: Praeger.

Wittgenstein, L. (1971[1958]). Philosophische Untersuchungen. *Frankfurt: Suhrkamp*.