Forthcoming in the *Journal of Consciousness Studies*

François Kammerer

The meta-problem of consciousness and the evidential approach[1]

Abstract:

I present and I implement what I take to be the best approach to solve the meta-problem: the evidential approach. The main tenet of this approach is to explain our problematic phenomenal intuitions by putting our representations of phenomenal states in perspective within the larger frame of the cognitive processes we use to conceive of evidence.

1. Solutions to the meta-problem and their limit

In "The Meta-Problem of Consciousness", David Chalmers presents the meta-problem of consciousness: the problem of explaining phenomenal intuitions in a topic-neutral way. Here I present what I think is the best approach to answer the meta-problem: the *evidential approach*.

The most promising solutions to the meta-problem reviewed by Chalmers all have something in common. First, one notices that the puzzling character of consciousness is grasped through introspection. Then, one tries to understand why we should expect introspection – the representation, by a cognitive system, of its own mental states – to generate problem intuitions. Maybe we should expect introspective mechanisms to represent mental states as instantiations of *primitive properties* (Chalmers, 2018, p. 25) and/or as instantiations of *primitive relations* to properties (Chalmers, 2018, p. 27), and/or as states about which we have *immediate* (i.e. non inferentially mediated) *knowledge* (Chalmers, 2018, p. 23).

[1] Thanks to David Chalmers, Keith Frankish and Wolfgang Schwarz for their useful comments, and to Sonia Paz-Higgins for her help.

These solutions all encounter a problem, mixing what Chalmers calls the "resistance problem" (similar to what I called the "illusion meta-problem" earlier (Kammerer, 2018)) and what he calls the "belief problem" (Chalmers, 2018, p. 25, 27). For most representations (including representations of primitive properties and relations, or representations providing immediate knowledge), we can easily contemplate that the represented entities could appear to us to be a certain way and yet still be different. I can represent that there is a white object in spatial contact with a red object in front of me (primitive properties, primitive relation), or that my name is François (immediate knowledge), or that I hope that no one will wish me a happy birthday (immediate knowledge). However, I can also easily represent to myself how all of that could *appear to me to be the case* without being the case. I could be *hallucinating* colored objects, *dreaming* that my name is François (while in fact it really is Jean-Pierre), or I could be *self-blind* and hiding from myself the fact that I actually long for birthday wishes. In comparison, phenomenal experiences stand out. I do not think things could *appear to me* as if I had an experience (say) of a red object without really having such an experience. If an experience *appears to me*, I have it.

Following Chalmers, we might explain this by the fact that we have a *sense of acquaintance* with experiences but not with other things. We take experiences to be directly and concretely presented to us, in a way that *reveals* them – while we do not think the same of colored objects, hopes and first names. The exact nature of this sense of acquaintance might be hard to pin down, but one can start by noticing that this sense of acquaintance is *more than a mere sense of immediate knowledge*: we think we know immediately (without inferential mediation) our own names, our (standing) hopes and beliefs, etc., but not that we are *acquainted* with these things. Indeed, only experiences seem concretely presented and revealed. Crucially, this sense of acquaintance is *also* different from a mere sense of *certainty* (at least in standard senses of "certainty"): arguably, I am certain that $(P\&Q) \rightarrow P$ or that 2+2=4, but the corresponding entities do not seem to be concretely presented ("right there") to me in the same way.

Our sense of acquaintance with experiences might explain why we think that they never appear other than they are. However, appealing to our sense of acquaintance *describes* the problem rather than solving it. Indeed, that we have such a sense of acquaintance stands itself in need of an explanation – at least on two levels. First, there is the *mechanism question*: what kind of mechanism generates our sense of acquaintance with our experiences? Second, the *design question*: why does phenomenal introspection (and *only* phenomenal introspection) rely on a similar mechanism? The "limitative" part of the question is important: we introspect some mental states (hopes, beliefs, etc.) with a mechanism that *does not* generate any sense of acquaintance. Why on earth is *phenomenal introspection* not just like *belief introspection*?

If we want to explain our sense of acquaintance, introspection, narrowly understood, is a red-herring – the wrong starting point. Instead of asking what we should expect of a cognitive system representing its own internal mental states, I suggest that we ask what we should expect from a cognitive system representing *some* mental states as grounding its (or other's) *evidential situation*. This is the evidential approach: trying to explain problem intuitions by understanding our grasp of consciousness within the larger frame of the cognitive processes we use to conceive of evidential situations.[2]

Why follow the evidential approach, on the face of it? As I said, our sense of acquaintance is not a mere sense of *immediate knowledge*, nor a mere sense of *certainty*. It is the sense that experiences are always *presented* to the experiencer in a peculiar way, which reveals them. But the concepts of *presentation* and *revelation* used to articulate our sense of acquaintance are arguably *epistemological*, *evidential* concepts. Something is presented to us when it *appears* to us, or is *cognitively given* to us – it features in the *evidence* we have at our disposal regarding

[2] The approach called "Immediate Knowledge" is, in Chalmers' list, the one that is closest to this approach – although their starting points are different (introspective representations vs representations of evidence).

reality, it is part of our *data*. Something is *revealed* to us when *all of it* appears to us – when its whole nature is systematically part of our *data*. It therefore seems natural, to understand our sense of acquaintance – this peculiar evidential status we give to experiences – to investigate the relation between the way we grasp experiences and the way we intuitively represent *evidence*, or *data*. The idea is that these two grasps might be intertwined in a way that explains our sense of acquaintance.

2. Evidential systems

The evidential approach, broadly understood, has been followed in various ways (Kammerer, 2016b, 2016a, 2019; Schwarz, 2018; Sturgeon, 1994) and might also be indirectly suggested by Chalmers himself when he stresses the importance of our sense of acquaintance (p. 25, p. 39). Here I would like to present one version of it.

First, I ask: what should we reasonably expect of a sophisticated enough cognitive system? Answer: we should expect such a system to represent *evidential situations*, using a representational mechanism which satisfies certain constraints (and notably represents a class of evidential states as *independent from beliefs* and *self-presenting*). This gives a partial answer to the *design question*. Second, I ask: what mechanism *do we use* to represent this class of evidential states? I speculatively suggest an evolutionarily plausible mechanism, which does the job, but also generates *a sense of acquaintance* (as well as other problem intuitions) with the states it represents. This gives a speculative answer to the *mechanism question*. Together, these considerations constitute an attempt at solving the meta-problem of consciousness by following the evidential approach.

Let us start with the point of view of design.

A. We should expect sophisticated enough cognitive systems to be evidential systems. Imagine a sophisticated cognitive system, with access to the world (sensory subsystems, behaving more or less like modules), cognitive states (beliefs, etc.), motivational states (desires, etc.), and meta-cognitive representations. Arguably, its meta-cognitive representational repertoire should not only feature representations of cognitive states and motivational states, but also of evidential states. It would represent *some* mental states as constituting the *evidential basis* for *some* propositions ("I am/it is in mental state X, in virtue of which I have/it has evidence for proposition P"). Let us call a system able to do this an *evidential system*. Here I understand *evidence* as follows: (a) having evidence for P gives a reason to believe that P; (b) knowing the total evidence possessed by a system is necessary (and arguably sufficient) to know what it should believe (and, if one assumes rationality, what it will believe). Note on the side: sometimes I will talk of evidence for *beliefs,* or for *states of affairs*, which is just a quick and rough way to say: evidence for a proposition which is the complement of a *belief*, or for a proposition made true by a given *state of affairs*.

Why should we expect sophisticated cognitive systems to be evidential systems? Because an ability to represent evidential situations (on top of a mere ability to represent cognitive and motivational states) allows for the following. (1) A more flexible and reflective process of belief fixation (improving answers to questions such as "What to believe now?") (2) Sharing evidence with others, leading to better epistemic cooperation ("I have evidence for P" – which is not the same as "I believe that P"). (3) Rational assessment of the beliefs of others ("Given their evidence, they should believe that P"). (4) Conditional predictions of others' beliefs ("Given their evidence, *if they are rational*, they will believe that P").

B. We should expect evidential systems to differently represent evidence that depends on beliefs, and evidence that holds independently of beliefs ("passive evidence"). One thing that an evidential system could do is represent that evidence for a belief B is provided by a knowledge of a fact F, partially constituted by an underlying belief B'; evidence for B' is then represented as partially constituted by an underlying belief B'', etc. Some beliefs could be represented as having no further support. There is no particular reason to think that the representation of such *unsupported beliefs* should create any difficulty. In my own case, I arguably represent my beliefs that *I hope no one wishes me a happy birthday* or that *my name is François* in this way. These are *unsupported beliefs*: beliefs for which I do not have further justification – which I *just happen to have and to maintain*.

Our system also has access to the world through sensory subsystems. Arguably, the output of these systems (sensory judgments, constituted by sensory representations, which, we can suppose so here, represent the instantiation of sensible properties, characterized as primitive and basic qualities of objects) is robust, and relatively independent of the states which are *under the direct control of the system* (typically, beliefs and intentions). In a fodorian vocabulary, these subsystems behave like *modules* (this is also the case of other systems dedicated to memory, language comprehension, form recognition – but I will set these aside for now and focus entirely on sensory systems). Moreover, the output of these subsystems is the causal basis for many of the system's beliefs.

We should expect all of these *real features* of the system's function to be at least *schematically captured* by the evidential system. Thus, aside from representing evidential relations amongst beliefs, the system should be able to track the output of sensory subsystems and represent this output as *providing evidence to the system, independently of what its beliefs (and other states it controls) are*. In the first-person case, the system can thus judge "Whatever I believe, decide, etc., I have evidence for P" ("P" corresponding to a sensory judgment, that we can express with sentences such as "There is a red thing in front of me"). Let us call such evidence, independent of the states controlled by the system, "passive evidence".

C. We should expect evidential systems to represent passive evidential states as *presenting* (as providing evidence *for other states*) but also always at the same time as *self-presenting* (as providing evidence *for themselves*). Consider a first-person belief about a given passive evidential state E ("I am in E": for example, "I have evidence, independently of what I believe, in favor of there being a red thing in front of me"). How will an evidential system represent the evidential support for such a belief? There are four obvious possibilities: it could represent this belief (i) as supported by *other beliefs*; (ii) as an *unsupported belief*; (iii) as supported by *another passive evidential state F*; (iv) as supported by E itself.

The third possibility requires the representation of potentially infinite and/or circular chains of passive evidential states, which seems uselessly cumbersome. The first two possibilities give rise to *unstable* self-ascriptions of passive evidential states: a system could believe that it has a piece of evidence for P *independently* of what it believes, but also believes that whether or not it *should believe* that it has such evidence *depends* on what it believes. This leads to the recognition of unstable situations: one sometimes think that one should believe that P but should also believe that one should *not* believe that P.

The fourth possibility, by contrast, provides a simpler and more convenient way to represent the evidential support of first-person beliefs about passive evidential states. It is reasonable to expect an evidential system to use it. So, we should expect passive evidential states to be represented as self-presenting, in the precise sense that we should expect them to be represented as such that, whenever one is in such state, one has evidence for one's being in such state.

Chalmers makes a similar point when stating that introspective systems will tend to give a foundational role to directly introspected states, and see them as providing evidence for other things as well as for themselves (Chalmers, 2018, p. 24). But Chalmers thinks that these states are seen as self-presenting simply because the system needs stopping points in the chains of justifications. I disagree: unsupported beliefs would be good stopping points too! The reason why passive evidential states must be seen as self-presenting comes from the need to have a coherent and stable conception of their *passive* nature.

So, from the point of view of design, we should expect a sophisticated cognitive system: (A) to be an *evidential system*; (B) to differently represent *evidence that depends on beliefs* and *unsupported beliefs* on the one hand, and *passive evidence* on the other hand; (C) to represent passive evidence as self-presenting (in the sense I gave to the expression). *Caveat:* (A-C) express what we should naturally expect, not what is necessary. There might be ways to build a sophisticated cognitive system (though maybe not an efficient one) without satisfying (A-C).

3. The evidence-by-resemblance mechanism (ERM)

A system satisfying (A-C) will not necessarily develop a full sense of acquaintance regarding its passive evidential states: self-presentation is not enough for acquaintance! However, I think that *we humans* are sophisticated cognitive systems, satisfying (A-C), and that the mechanism by which we represent passive evidential states (usually called "conscious experiences") *does* generate a sense of acquaintance about them. There might have been *other possible ways* to satisfy (A-C), but evolution, in our case, came up with a certain mechanism, which generates this sense of acquaintance.[3]

I call this mechanism "**ERM**". It presupposes that we already are *evidential systems*, endowed with a concept of evidence (i.e. we already satisfy A).

Evidence by Resemblance Mechanism (**ERM**):

[3] This mechanism is the one I described in (Kammerer, 2016a, 2019). The one described in (Kammerer, 2016b) was more specific and its presentation somewhat *ad hoc*; its explanatory power is captured and extended by the new one. See (Kammerer, 2019, n. 23) for a comparison.

(i) ERM is an innate, modular mechanism. It forms and applies representations at a subpersonal, implicit level. These representations **track** (notably) the output of our sensory subsystems (there is no need to imagine a dedicated complex tracking mechanism, as ERM can directly "take up" the output of such subsystems, which is arguably *already broadcasted* in the system)

(ii) These representations characterize their referents as *passive – constitutively independent* of internal states under the control of the system, such as beliefs and intentions (arguably, a matter of the functional relations of these representations with the representations of beliefs, intentions, etc.)

(iii) These representations are composed by recruiting our *sensory representations* (representing sensible properties, characterized as primitive and basic qualities of objects), our concept of *evidence*, and a primitive and basic concept of *resemblance*. They represent their referents as states **which *resemble* the external states of affairs represented by sensory representations**[4], and which provide evidence to the subject who has them in virtue of that resemblance, according to this rule**: a passive state provides evidence for whatever state of affairs it maximally resembles.** *Maximal resemblance* here means the following: a given passive state maximally resembles a state of affairs X if and only if (a) it resembles X; (b) it is, amongst all possible passive states of the subject, the one that *resembles X the most*.[5]

An evidential system (satisfying A) using ERM represents passive evidential states in a way that distinguishes them from the evidence provided by beliefs (it satisfies B). Crucially, if we

[4] Given that these representations have such content, and given that they *track* sensory states, one can say they *characterize* sensory states as resembling (in a specific way) the external states of affairs they detect. If (as I think) sensory states *do not really thus resemble* these states of affairs, this characterization is a *mis*characterization.

[5] Given this conception of maximal resemblance, there is nothing logically problematic with one given passive state maximally resembling *various different states* at the same time.

consider the rule used by this system to determine the evidential power of its passive evidential states (see feature iii), we see that these states will be represented as always *self-presenting*. Indeed, any passive evidential state will naturally be represented as *maximally resembling itself* (it is arguably a constitutive rule of any concept of resemblance that all things perfectly resemble themselves), which means any passive evidential state will be represented as *providing evidence for itself* (on top of other things). So, a system using ERM represents passive evidential states as self-presenting (it satisfies C).

ERM is a simple mechanism, following simple rules. The representations it recruits (representations of evidence, of resemblance, and sensory representations) are arguably representations that any complex cognitive system has. By following the rule of resemblance, it gives the output that is required from the point of view of design: passive evidential states are represented not only as *presenting*, but as *self-presenting*. That makes ERM a simple, evolutionarily plausible mechanism, which naturally satisfies C. It is therefore not entirely surprising that (as I speculate) evolution has gifted us with ERM to represent passive evidence.

4. Solving the meta-problem

Suppose we do use ERM. It represents (in a modular way) that we enter passive evidential states which provide evidence for certain external states of affairs (for the proposition 'there is a red thing there'), as well as for themselves (for the proposition 'I have passive evidence that there is a red thing there'). When seen as *presenting* other states, such states are also seen as capable of providing *misleading evidence*. We can be in a state that *resembles* a given state of affairs, for which we thereby have evidence, even if that state of affairs is not the case (e.g. if I am in a state resembling the presence of a red thing, although there is no red thing). But a crucial by-product of the functioning of ERM is that it represents that it is impossible to have misleading evidence *for our own passive evidential states*.

Indeed, what does ERM state about the *evidence* we have for our own passive evidential states? ERM cannot represent such evidence as dependent on *beliefs* (supported or unsupported), as this would create the kind of unstable self-ascriptions mentioned in 2.C. So, this evidence must be *passive*. But what happens when ERM represents that I have passive evidence for me being in passive evidential state E? ERM outputs that I am in a passive state which maximally resembles E – in a passive state that, amongst all possible passive states, is the one that resembles E the most. So, ERM outputs that I am in a state type-identical with E. Therefore, ERM represents that I cannot have misleading evidence for myself being in a passive evidential state of a certain kind, and that, more generally, a system's passive evidential states must be *exactly like* what the system's evidence for them presents them to be: ERM represents passive evidential states as *revealed*. Note that this only happens in the *first-person*: nothing prevents my evidence for *your* passive evidential states from not being *passive*, but from depending on my *beliefs* (as this does not lead to unstable ascriptions), which is why I easily accept that I can have misleading evidence about *your* passive evidential states.

Let us take stock. ERM represents passive evidential states as *presenting* the instantiation of sensible properties represented by sensory representations, as *self-presenting*, and as *revealed*. So, a system using ERM has a *sense of acquaintance* regarding its passive evidential states. The key to understand the generation of this sense of acquaintance is the use by ERM of the rule of maximal resemblance: this is a simple, rough rule, which gives the output required from the point of view of design (it generates a sense of *presentation* and *self-presentation*). However, it has a crucial by-product (not explained at the level of design): it generates a sense of *revelation*.

Passive states are represented as *resembling* instantiations of sensible properties, represented by sensory representations. If, as I supposed earlier, our sensory representations characterize sensible properties as *primitive* qualities, *physicalism* regarding passive evidential states will be judged inacceptable – if one has a conception of the physical which excludes anything resembling a primitive quality. Of course, the system can always integrate the primitive qualities *represented by sensory representations* in the physical world, *by judging that they are not really as they are presented to be* (they are not really primitive, but complex physical properties). Crucially however, such a move is not available for passive evidential states – because they are represented as *revealed*. Therefore, a system using ERM will develop persistent and ineliminable *primitivist* and *anti-physicalist* intuitions about passive evidential states. It think it would also develop other problem intuitions.

I think that human phenomenal introspection uses ERM: our representations of phenomenal states are nothing but representations of passive evidential states. This explains, *at the level of mechanism*, why we have a sense of acquaintance regarding phenomenal states (thus solving the "resistance problem"), as well as other problem intuitions. Why we use ERM is not itself explained at the level of design; but it is nevertheless *not entirely arbitrary*, as ERM is an evolutionarily plausible mechanism that satisfies the design constraints that a sophisticated cognitive system is expected to satisfy. Moreover, this view explains, at the level of design, why we differently represent evidence depending on beliefs, and passive evidence. This solves the "belief problem" (i.e. explains why phenomenal introspection is not akin to belief introspection) at the level of design.

My view, contrary to other views taking the evidential approach (Schwarz, 2018), does not predict that we will have absolute, unshakable certainty about experiences. Indeed, we *do not*; as *some* people (illusionists) deny that they really have experiences. What my view predicts as intuitively unacceptable is *not* that we can make mistakes about experiences, but that we can make mistakes about experiences based *on misleading evidence* about them (at least when we use our innate, intuitive representation of evidence). The only mistakes about experiences we intuitively accept are mistakes made *in spite of the available evidence* – due to irrationality. We intuitively think that only systematically irrational persons – *mad* persons – can be *systematically* wrong about their experiences.[6] This is why so many realists about consciousness jokingly suppose (wrongly) that illusionists must be zombies (the realist assuming here that they are themselves *not* zombies), lunatics – or insincere.

The evidential approach is the best way to tackle the meta-problem. My own version hopefully gives an approximately correct solution to this problem. The hypothesis that we use ERM for phenomenal introspection is, of course, highly speculative. The evidential approach has been (and will be) pursued in different directions, leading to different hypotheses. I would very much like to see more such hypotheses developed in the future.[7]

The view I presented is silent on the *existence* of phenomenal states, but I recommend conjoining it with strong illusionism (Frankish, 2016): I think that it seems to us that we are in phenomenal states (in the sense that we tend to represent that we are in passive evidential states, which are self-presenting and revealed), but we are not. Such conjunction gives us an illusionist theory which explains why illusionism regarding consciousness is uniquely difficult to believe and to contemplate. Our deep sense of acquaintance prevents us from making intuitive sense of the idea that consciousness is an illusion, although this idea is coherent (and true) when formulated *without* our innate and intuitive concept of evidence. If this view is true, we should not be surprised if David Chalmers (and many others!) thinks that illusionism is obviously false and denies the most fundamental and immediate data we have. This kind of judgment is

[6] This might be why Descartes, in the *Meditations*, rules out the *madness hypothesis* (“sed amentes sunt isti”) *before* the *Cogito*.

[7] For potential views similar to mine but which do not rely on *resemblance,* see (Kammerer, 2019, n. 24).

naturally explained when we understand the intertwining of our grasp of consciousness and of our representation of evidence – of the *data*.

References:

Chalmers, D. (2018). The Meta-Problem of Consciousness. *Journal of Consciousness Studies*, *25*(9-10), 6-61.

Frankish, K. (2016). Illusionism as a Theory of Consciousness. *Journal of Consciousness Studies*, *23*(11-12), 11-39.

Kammerer, F. (2016a). *Le problème de l'expérience consciente: une tentative de dissolution* (PhD dissertation). Université Paris-Sorbonne, Paris.

Kammerer, F. (2016b). The hardest aspect of the illusion problem - and how to solve it. *Journal of Consciousness Studies*, *23*(11-12), 123-139.

Kammerer, F. (2018). Can you believe it? Illusionism and the illusion meta-problem. *Philosophical Psychology*, *31*(1), 44-67.

Kammerer, F. (2019). The illusion of conscious experience. *Synthese*. https://doi.org/10.1007/s11229-018-02071-y

Schwarz, W. (2018). Imaginary Foundations. *Ergo*, *5*(29).

Sturgeon, S. (1994). The Epistemic View of Subjectivity. *The Journal of Philosophy*, *91*(5), 221-235.