Notes accompanying talk on

CONSCIOUSNESS AND META-CAUSATION

Work in Progress seminar,

Philosophy Department, University of Birmingham, 22 May 2019.

John Barnden, Emeritus Prof. of AI, University of Birmingham

[jabarnden@btinternet.com](mailto:jabarnden@btinternet.com), https://www.cs.bham.ac.uk/~jab/

(NB: These notes are not a self-contained handout, and should be considered in conjunction with the slides, which I can send to anyone interested, and which will be on my website.)

**More on: Background Assumptions and Motivations**

* I assume throughout that being-occurrently-conscious in the phenomenal sense is in the first instance a property of processes, and only derivatively a property of creatures, computers, programs, or mental states (except insofar as these are cashed out as processes).
* Although I am (obviously) dissatisfied with existing theories of consciousness, some may provide important *facilitatory* conditions/mechanisms for consciousness, at least in its richer aspects/cases.

Example conditions/mechanisms: global workspaces (Baars 1988), attentional schemas (Graziano 2013), complex causal networks (Oizumi, Albantakis & Tononi 2014), long-distance binding mechanisms (Yates 2018), looping architectures (Edelman & Tononi 2000), cognitive poised-ness (Tye 2000), higher-order thought frameworks (Carruthers 2011, Rosenthal 1993), …

* Such theories, in my view, don’t provide much that’s persuasive even in terms of *necessary* conditions for consciousness *in general* (within a physicalist framework), although they may provide necessary conditions for some specific types of consciousness (e.g. particularly elaborate or introspective or action-oriented types).
* BUT neither am I persuaded by arguments (e.g. from Explanation, Knowledge, Conceivability – Chalmers 2010) towards types of dualism or (non-physical) monism incl. panpsychism or panprotopsychism (Brüntrup & Jaskolla 2016).
* For a contemporary strong view on the reality of consciousness—that it is in fact the only certain reality--see Strawson (2016), though I don’t adopt his panpsychism. Strawson also has other stances congenial to my approach, such as a process-based view of the universe and the view that the experiencer in an episode of experiencing can be just the experiencing itself.

**More on: The Huge Promissory Note to Physics**

We shouldn’t be embarrassed to point the way (in a responsible, considered way) towards new physics, *especially* in the consciousness area, in the light of:

* + Physicists suggesting things like dark energy, dark matter, modified gravity, …, just to address ordinary observable phenomena ( as opposed to something as mysterious as consciousness).
  + The weird current status of quantum physics, including the continued existence of ideas such as consciousness causing wave-function collapse (what I’m proposing is less radical and fantastic than this! … or on the other hand might conceivably help explicate it)
  + Lack/difficulty of the type of experimentation on brains that might be required to reveal parts of physics that apply only within conscious processes.

**More on: (Causation and) Meta-Causation in General**

**[i.e. not necessarily at basics physics level, and not necessarily as dynamism]**

NB: I tend towards heuristically-motivated projectivism (see, e.g., Kutach 2014: 31–32) about causation in senses different from the dynamism of basic physics: In those senses: i.e., causation in those senses is a heuristic notion or variegated bundle of notions, that we apply, perhaps not completely coherently, to the world as imposed construals of it, for a variety of not necessarily compatible purposes. So, of course, these issues are only magnified if we consider meta-causation.

* While meta-causation in my sense appears to be a minority concern, and receives only a brief mention in an authoritative survey (Ehring 2009), it has been viewed as important: e.g. in the talk in January 2019 by David Kovacs at a FraMEPhys workshop [Kovacs 2019], and in certain areas such as teleological accounts of intentionality and representation (Koons 1998). It can also be seen in discussions about God creating the whole universe including its causal instances and of free will affecting the normal course of causation.
* Meta-causation in my sense as a part of consciousness has already been at least implicitly raised in discussions of whether people can have direct conscious experience of causation (see, e.g., Beebee 2009).
* Meta-causation in my sense is variably ontologically problematic according to what the “causal relata” are: events, facts, … E.g. If causes and effects are facts, great. If they’re events, presumably a causing will have to be too for it to be a relatum in a meta-causing.
* Meta-causation in my sense is sometimes called **“higher-order” causation** (Koons 1998) or [misleadingly] “iterated” causation (Ehring 2009). The latter term is unfortunate because it smacks of mere chains of causation links.
* Some people (e.g. G. Ellis 2016) use “meta-causation” to mean downward/top-down causation. **I do NOT mean downward causation in my usage.**
* Sometimes the term is used to mean the modification of parameters of causal models (Fernández 2010).
* Caution: “meta-causal powers” (B. Ellis 2013) have been mooted in the neo-Aristotelean “powers” area (see, e.g., Groff & Greco 2013, Jacobs 2017, Mumford & Anjum 2011). However, meta-causal powers are ones whose exercisings create/destroy/modify other *powers* (which can be thought of as *dispositions* to cause things): they don’t *directly* affect (or respond to) causings as such**, and so are NOT to do with meta-causation in my sense.** (In effect, “meta-causal powers” should be parsed as meta-<causal-powers> not <meta-causal> powers.)
* Having said that, there’s no reason why the effect of a left-handed meta-causing in my sense should not be some such effect on a power, in a powers-based rendering of my approach.
* Relatedly, there are Vetter’s (2015) iterated potentialities: excluding the base case of an ordinary potentiality, these are potentialities for there to be a potentiality, and might be called meta-potentialities (but Vetter reserves the latter term for potentialities possessed by potentialities).
* Meta-causing of order N > 1 (i.e., meta-meta-…-meta-causing with N metas) works on causation at any level below N but including some at N-1, and on ordinary causation.

**Thought-Experiments suggesting Genuine Processuality is Needed for Consciousness**

NB: These thought-experiments are mostly very similar, to the manipulations they imagine, to thought-experiments throughout the consciousness literature, e.g. in Bishop (2009) and Kirk (2005). However, my particular goals are different. CAUTION: this section is only for arguing that a conscious process must be a genuine process, NOT showing that that genuine processuality matters TO the process itself.

1. Suppose we think that consciousness is (or is metaphysically grounded in or lawfully correlated with or …) patterns of activity at some level in a conscious brain – e.g. patterns of neural firing. We now imagine making a sort of “copy” of the changing patterns of firing while eliminating the causation that links them together in the original brain, as follows.

The copy is a discretized one, with at each moment a pattern of (e.g.) electric potentials imposed on completely electrical units of some sort from outside the whole copy. The units are entirely disconnected from each other. The potential are imposed at a discrete sequence of times t1, t2, t3, …, where the pattern at t\_i corresponds to the pattern in the original brain at time t\_i, and the pattern is kept constant in the interval [t\_i, t\_i+1). Then, we have pseudo-process across t1, t2, t3, … that can match the original brain ever more exactly as we make the discretization ever finer. Surely none of the pseudo-processes are conscious. After all, the units in the copy have nothing to do with each other objectively, so why should a changing pattern across all of them constitute anything special, let alone consciousness?

So, unless for some reason just being smoothly-changing (as opposed to having discrete jumping of potential values in the discretization) somehow matters to being-conscious (but why would it, especially if the jumps are within the normal jitter of e.g. neural potential values?), it seems reasonable to suppose that it’s the causal linkage in the original brain that matters.

1. The point comes out more starkly if you think a run of a suitable computer program is conscious, where what mattered was the pattern of values (bits) that arise successively in computer memory. It surely wouldn’t matter if you re-engineered the computer to work with all the actual 1s and 0s in memory written out on an enormous piece of paper, with at each clock cycle in the computer the 1s and 0s are suitably rearranged according to the demands of the computer program. You could literally take a film of the changing pattern of 1s and 0s (let’s say at one film frame per clock cycle for simplicity), thus missing out the mutual causation. Surely the pseudo-process consisting of a running of the film wouldn’t be conscious.

Another relevant thought here: If you thought the running of the film was conscious, why would the *running* of it be at all significant? Wouldn’t just tapping your finger on each successive film frame, together with at each tap the frame tapped upon, define a pseudo-process that involves the same sequence of computer states as the running film? Is that pseudo-process conscious?

1. There is an additional “diagonalizing” version of such arguments that manipulates similar intuitions but in a different way.

Take any conscious process, where for definiteness the process consists of changing values of some sort on a physical substrate such as a brain or computer. Split the process up arbitrarily into sequential temporal chunks. We can now imagine taking N copies of the original substrate and process, each copy cut up into chunks at the same temporal joints. Suppose the internal causation within a process doesn’t matter to consciousness. So we now have N+1 identical instances of consciousness in different places, don’t we, given a natural assumption that copying the whole of a conscious process delivers another conscious process? So far so good, it seems.

But since causation doesn’t matter, and since we need to allow a conscious process to physically move from place to place in an arbitrary way, and since we can surely replace the particular matter in the substrate at any time by identical but different matter, it’s easy to see that *any* sequence formed in the following way should be conscious: chunk 1 from an arbitrary on of the N+1 processes, chunk 2 from another arbitrary one, chunk 3 from another, and so on. This leads to a vast number of possible sequences (N^T where T is the number of chunks in the original process).

Perhaps some readers will have the intuition that we do in fact, inadvertently so to speak, have this number of criss-crossing conscious processes, but to me this conclusion is so weird I’d rather assume that causation within processes does matter for whether they’re conscious or not.

**More on: Pre-Reflective Self-Consciousness (PRSC) / Auto-Sensitivity (PRAS[-GP]),**

**and the Transparency Issue**

* The PRSC assumption doesn’t negate the idea that much consciousness is self-conscious in a reflective, concept-based way. BUT such consciousness still contains the PRSC, and without it the reflective introspection would be unconscious. See especially Kriegel (2009).
* There have been many variations of PRSC, and Guillot (2017) has ably pointed out some conflations/confusions. See also Williford (2015) [though himself criticized to an extent by Guillot]. Other contemporary references are Gallagher & Zahavi (2015), Kriegel (2009), Sebastian (2012), Zahavi (2005). The idea is often traced back to 20th century phenomenologists such as Sartre, and authors such as Kriegel and Zahavi represent a neo-phenomenological turn.
* Although I’m basically sympathetic to the PRSC assumption, there is indeed the *danger* that we can only (consciously!) think up ideas like PRSC when we’re reflecting on consciousness, not just normally engaged in it, so there’s at least a *danger* that proponents of PRSC may be indulging in (e.g.) false memory of their own normal phenomenology.
* That is roughly balanced by the *danger* that opponents of PRSC have wrongly understood what proponents mean but it, because it’s very difficult to describe and it’s often described in a vague, confused and/or over-rich way (Guillot *ibid.,* Williford *ibid.*), and/or are themselves wrongly remembering their own normal phenomenology when philosophizing about it!
* Levine (2018) proposes that we *are* in some way always aware of our experiencing while engaging in it, but the resulting subjectivity and sense of immediacy gets “transferred” in our phenomenology to the perceived red patch or whatever. I find this a very congenial proposal, but I think is more carefully dealt with by something in the vicinity of my PRAS assumption, which replaces conscious awareness of the experiencing by sensitivity to it.
* Indeed, a problem here is that Levine, Williford and other authors have jumped too quickly to the assumption that our own awareness of our own experiencing is, from the start, a form of *conscious* awareness. It may turn out to be but this shouldn’t be baked in by postulation.
* In sympathy with Williford (2015) and others, I want to avoid assuming awareness of a rich or personally distinct “self” in PRSC or PRAS. Hence I prefer “auto” where “self” is normally used as a prefix to “consciousness”, except when consciousness of a “self” worth that name is really intended.

After all, a self-cleaning oven doesn’t clean some personal self, at least not at the current stage of AI technology. And if you called an automobile a selfmobile you wouldn’t then credit it with a “self”, I presume!

* I do want to respect transparency intuitions to some extent, for safety. But I hold that the (e.g.) red patch in the experience is actually experienced as *own-thing-as-being-experienced-now,* and so the subject’s own experiencing, *though detected*, has not been *phenomenally* detected in its own right, and consciousness of it has been “transferred” into that experienced thing (cf. Levine *ibid*). The alleged transparency would in this sense be a result of over-simplified post-hoc reflection that misconstrues a red patch in past non-reflective experience as something not itself experiencing-laden.

I don’t have a strong view as to how often perceiving is transparent (in the above limited sense) versus self-conscious.

* PRSC doesn’t start to cross the consciousness to physics explanatory gap, because it just partially analyses consciousness in terms of specialized consciousness. But PRAS(-GP) does so start.

Relatedly, sufficiency for consciousness is vacuously intrinsic to PRSC. This sufficiency goes away in the step to PRAS[-GP]. But my main arguments are currently only towards a necessary condition anyway.

* Under my PRAS-based account, PRSC-laden experience can still happen when the sensitivity is indeed a matter of *conscious* awareness of the experiencing—i.e. when the proto-acquaintance is in fact acquaintance. I propose that this happens when the auto-sensitive processing includes sensitivity to that auto-sensitivity, i.e. second-order auto-sensitivity. So notice that *first*-order self-consciousness comes from *second*-order auto-sensitivity. (And so on to higher levels.)
* With the equation of sensitivity with proto-acquaintance, sensitivity to own sensitivity is reminiscent of the notion of *acquaintance with acquaintance* (see Raleigh, forthcoming, and references in Williford 2015). However, we need to be careful to distinguish between acquaintance with acquaintance-as-a-universal versus acquaintance with instances of acquaintance. The right version for me is proto-acquaintance with proto-acquaintance instances. I don’t propose proto-acquaintance with proto-acquaintance as a universal.
* Approaches such as mine do not imply separable layers of processing. The (e.g.) visual processing may contain inseparably the aspects of processing constituting first-order auto-sensitivity, when the seeing is conscious, and those aspects may inseparably contain the aspects needed for second-order auto-sensitivity, when it is self-conscious. And so forth.
* The ideas surrounding PRAS are very similar to those found other so-called “same-order” theories of consciousness, but with a focus on processes rather than mental states, and not committing to an idea that the auto-sensitivity is a matter either of representation or something perception-like. The main originality of the approach lies instead in introducing the auto-sensitivity to own genuine processuality and in claiming that meta-causation is crucial.

**More on: The Argument to Meta-Causation from PRAS-GP**

* Some sort of dynamic auto-simulation might be proposed to get over one of the objections to a representational approach, namely that the required representation might have to be too propositional and/or just too complex. But such simulation would just be a form of representation involving structural similarity (Shea 2018), with all the problems of non-naturalness and mis/failed-representation as usual.
* The non-naturalness and mis/failed-representation problems with auto-representation are what make Willford (2015) propose auto-acquaintance, not auto-representation.
* The effects of the meta-causation (the meta-causation that the argument supports) are conceivably quite simple in form: e.g, there is no currently evident need to take them to be anything as complex as a representation of prior states or causal linkage. But this is a big issue for further research …
* … and indeed a lot else about the meta-c is left “to be determined” by further research, such as:
  + How long an interval before the current moment does the meta-causation need to respond to?
  + Which causings within that interval does it need to respond to?
  + What effects does the current meta-causing have?
  + How intertwined/integrated is the meta-causation with base-level causation?

**Further Support for Meta-Causation: A Thought Experiment resting on Auto-Indexicality**

In Barnden (2018) I present a thought-experiment that seeks to show that supposing that a supposition that the auto-sensitivity demanded by PRAS-GP is fully accounted for by causal reaction to own prior state trajectory—with no direct reaction to its inner causal linking—leads to deep weirdness. This weirdness is in the form that either it’s no longer an objective matter what conscious processes are present or we’re pressured to make an unpalatable postulation about that issue.

This *reductio-ad-weirdum* is accompanied by the observation that the problem goes away if auto-sensitivity by meta-causation is assumed to be crucial in a conscious process.

Essentially the point is one of indexical *de-se-*ness: how a conscious process is directly and non-conceptually to identify itself to itself.

**More on Dynamism and Meta-Dynamism**

* As regards Humean versus anti-Humean views on the universe’s regularities, see, e.g., Kutach (2014), Maudlin (2007) and Schaffer (2016).
* I find Maudlin’s (2007) views and arguments especially congenial, but instead of tending in the direction of regarding physical laws needing to be something real over and above the sheer regularities of the universe, I would say that, in a law-based view of the universe, what primarily needs to be a real *constituent* of the universe is the ongoingly-being-law-governed of its unfolding. Whether laws as such are also, separately, real *constituents* is another matter and sounds (yet) more mysterious to me.
* Meta-Dynamism of order N+1 would be expressed in equations that mention (features of) dynamism at levels up to and including N, plus perhaps ordinary physical quantities.
* NB: I include in dynamism/causation things that might normally be regarded as causal, such effects transmitted through quantum entanglement.
* The dynamism exists everywhere and at the lowest spatiotemporal scales, in that sense is at the level of “micro-physics”. But it can be considered also in a macro way, just as fields are. The special meta-causation for consciousness may possibly be well analysable at a macro level.

**A Precursor Paper**

The current state of the project is a major development from preliminary ideas and rather rough arguments in Barnden (2014).

Because of my reservations (see above) about the objective reality of “causation” other than as overall physical dynamism or “oomph” [see, e.g., Demarest (2017), Kutach (2014) and Schaffer (2016) for the latter term] I mostly tried to avoid the term in that paper. Instead, I used “running[ness]” for the dynamism or oomph. (In fact I mean a restricted version of dynamism, namely law-governed, as I had not yet considered casting the ideas in a powers-based framework.)

**Relationship of this Project and Talk to My AI Background**

Though my career discipline is AI, this project and talk are not specifically focussed on the question of conscious AI systems.

Nevertheless:

* The project has a major implication for the possibility of conscious AI programs: the physical implementation of the program must be implemented in such a way that meta-causation is suitably involved: the causal interactions between successive machine states must suitably meta-cause effects in the computation. An ordinary implementation won’t do, and it’s not enough for the computation itself to be reflexive or to detect states (in the ordinary sense) of its implementational base.
* The project does have its origins in thoughts about how AI systems could be conscious, right back when I was a graduate student. They’ve been tunnelling hrough my consciousness ever since! But my actual main interest is in consciousness in general, with AI systems as a special case.

My main official research in recent years has been on metaphor understanding and the use of metaphor in thought itself. In particular, I’m interested not just in the metaphors we use in talking and thinking about conscious states at large, but particular in metaphor we use (not necessarily consciously) in thinking about and otherwise manipulating our own conscious states. I believe this phenomenon may affect the intrinsic nature of phenomenality at least in human beings (Barnden 2005; updated in a recent talk—slides available on request).

**References**

Barnden, J.A. (2005). Metaphor, self-reflection, and the nature of mind. In D. N. Davis (Ed.), *Visions of Mind: Architectures for Cognition and Affect*, pp.45--65.

Barnden, J.A. (2014). Running into consciousness. *J. Consciousness Studies*, *21*(5–6): 33–56.

Barnden, J.A. (2018). Consciousness, causal oomph and meta-oomph. Poster presented at the *26th Conference of the European Society for Philosophy and Psychology,* 10--13 Sep. 2018, Rijeka, Croatia.

Baars, B. (1988). *A Cognitive Theory of Consciousness.* Cambridge: Cambridge University Press.

Beebee, H. (2009). Causation and observation. In H. Beebee, C. Hitchcock & P. Menzies (Eds), *The Oxford Handbook of Causation.* pp.471–497. Oxford Scholarship Online ( Oxford University Press).

Bishop, J.M. (2009). A cognitive computation fallacy? Cognition, computations and panpsychism, *Cognitive Computation, 1*, pp.221--233.

Brüntrup, G. & Jaskolla, L. (Eds) (2016). Panpsychism: contemporary Perspectives. Oxford Scholarship Online, October 2016.

Carruthers, P. (2011) Higher-order theories of consciousness, in Zalta, E.N. (ed.) *The Stanford Encyclopedia* *of Philosophy* (Fall 2011 ed.), online at ttp://plato.stanford.edu/archives/fall2011/entries/consciousnesshigher/.

Chalmers, D.J. (2010). *The character of consciousness.* New York: Oxford University Press..

Demarest, H. (2017). Powerful properties, powerless laws. In Jacobs (2017), Ch.4.

Dowe, P. (2009). Causal processes. In *Stanford Encyclopedia of Philo*sophy (Spring 2009 ed.), online at http://plato.stanford.edu/archives/spr2009/entries/causation-process/.

Edelman, G.M. & Tononi, G. (2000). *A Universe of consciousness: How matter becomes imagination.* Basic Books.

Ehring, D. (2009). Causal relata. In H. Beebee, C. Hitchcock & P. Menzies (Eds), *The Oxford Handbook of Causation.* pp.387–413. Oxford: Oxford University Press.

Ellis, B. (2013). The power of agency. In Groff & Greco (2013), pp.186–206.

Ellis, G. (2016). *How can physics underlie the mind? Top-down causation in the human context.* Dordrecht: Springer.

Fernández, E. (2010). Living is surviving: causation, reproduction and semiosis. Manuscript presented at *Tenth Annual Conference on Biosemiotics*, Gatherings in Biosemiotics, Braga, Portugal.

Gallagher, S. & Zahavi, D. (2015). Phenomenological approaches to self-consciousness. In Edward N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2015 Edition), URL = http://plato.stanford.edu/archives/spr2015/entries/self-consciousness-phenomenological/.

Graziano, M.S.A. (2013). *Consciousness and the Social Brain*. Oxford: Oxford University Press.

Groff, R. (2013). Whose powers? Which agency? In Groff & Greco (2013), pp.207–227.

Groff, R. & Greco, J. (Eds) (2013). *Powers and capacities in philosophy: the New Aristotelianism.* New York and London: Routledge.

Guillot, Marie (2017). *I me mine:* on a confusion concerning the subjective character of experience. *Review of Philosophy and Psychology, 8*, pp.23--53.

Jacobs, J.D. (Ed.) (2017). *Causal powers.* Oxford: Oxford University Press.

Kirk, R. (2005). *Zombies and consciousness.* Oxford: Clarendon Press (Oxford University Press).

Koons, R.C. (1998). Teleology as higher-order causation: A situation-theoretic account. *Minds and Machines, 8***:** 559–585.

Kovacs, D.M. (2019). The question of meta-causation. Talk at *FraMEPhys/Gothenburg Conference on Metaphysical Explanation in Science*, co-organized by Philosophy department, University of Birmingham, at Ikon Gallery, Birmingham, 10–11 Jan 2019.

Kriegel, U. (2009). *Subjective consciousness: A self-representational theory.* Oxford: Oxford University Press.

Kutach, D. (2014). *Causation.* Cambridge, UK: Polity Press.

Levine, J. (2018). *Quality and Content: Essays on Consciousness, Representation, and Modality*. Scholarship Online: April 2018.

Maudlin, T. (2007). *The metaphysics within physics.* Oxford: Oxford University Press.

Mumford, S. & Anjum, R.L. (2011). *Getting causes from powers.* Oxford: Oxford University Press.

Oizumi, M., Albantakis, L. & Tononi, G. (2014). From the phenomenology to the mechanisms of consciousness: Integrated Information Theory 3.0. *PLoS Computational Biology, 10* (5): e1003588.

Raleigh, T. (forthcoming). Introduction: The recent renaissance of acquaintance. Introduction chapter in J. Knowles & T. Raleigh, *Acquaintance: New Essays.* Oxford University Press.

Rosenthal, D.M. (1993). State consciousness and transitive consciousness. *Consciousness and Cognition, 2* (4), pp.355–363.

Schaffer, J. (2016). The metaphysics of causation. In E.N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2016 Edition), online at <http://plato.stanford.edu/archives/fall2016/entries/causation-metaphysics/>

Sebastián, M.A. (2012). Experiential awareness: Do you *it* to *me*? *Philosophical Topics, 40*(2): pp.155–177.

Shea, N. (2018). *Representation in cognitive science.* Oxford: Oxford University Press.

Strawson, G. (2016). Mind and being: The primacy of panpsychism. In Brüntrup & Jaskolla (2016).

Tye,M. (2000). *Consciousness, color, and content.* Cambridge, MA: MIT Press.

Vetter, B.(2015). *Potentiality: From dispositions to modality.* Oxford: Oxford University Press.

Williford, K. (2015). Representationalisms, subjective character, and self-acquaintance. In T. Metzinger & J.M. Windt (Eds), *Open MIND*: 39(T). Frankfurt am Mein: MIND Group. <https://fias.uni-frankfurt.de/mindgroup/>

Yates, D. (2018). Neural synchrony and the causal efficacy of consciousness. *Topoi,* published online 22 Sep 2018. https://doi.org/10.1007/s11245-018-9596-7

Zahavi, D. (2005). *Subjectivity and selfhood: Investigating the first-person perspective.* Cambridge, Mass. and London, UK: MIT Press (a Bradford Book).