Filosofický ústav AV ČR
Institute of Philosophy

Tomáš Marvan
Institute of Philosophy of the Czech Academy of Sciences
Jilská 1, Praha 1
110 00

Examiner's report on Martin Vraný's PhD thesis submission:
"Naturalizing the Unity of Consciousness. Can Neuroscience Explain a Fundamental feature of Subjectivity?"

(Thesis supervisor: Professor Jaroslav Peregrin)

The PhD thesis of Martin Vraný addresses the topic of the unity of consciousness. Out of a number of distinct senses of this unity, Vraný mostly restricts his attention to "subject unity" and "integration unity", both taken only in their synchronic aspects. The first, to put it roughly, concerns the unification of various contents into contents of one consciousness at a certain point in time. The latter concerns the neural mechanisms that are causally responsible for the paralell unification of the vehicles of experienced contents. Vraný takes these two aspects of the unity of consciousness to be two sides of the same coin. His claim is that while we need traditional philosophical tools such as Kant's theory of self-consciousness to address the first type of the unity, contemporary cognitive neuroscience is equipped to adress the second type.

For a PhD thesis, the text is in many places relatively unaccessible and difficult to comprehend. It is quite possible that I repeatedly had this impression because I am not an expert on the unity of consciousness. Another obstacle might be that my conception of consciousness and of the goals of the theory of consciousness to some extent differ from Vraný's. However, I believe that a PhD thesis should aim for maximum clarity. This thesis falls somewhat short of this desideratum. I was repeatedly asking myself whether

the text is so difficult to read because Vraný very condensely expresses a clear idea, or because the idea expressed, and the broader contexts of thoughts in which the idea is embedded, is not entirely clear. Moreover, I believe the submitted thesis is somewhat over-ambitious in the scope of the difficult problems it tries to tackle. A necessary consequence of this is that some of the issues are just skimmed though without proper philosophical care.

On the other hand, the virtue of the thesis is that it comes to grips with some *very* difficult and fundamental philosophical problems and attempts to think anew the problem of the unity of consciousness from the viewpoint of contemporary cognitive neuroscience. I applaud Vraný's bold attempt to extract from some of the leading neuroscientific theories of mind the materials for a comprehensive account of the "integration" unity of consciousness. The theories he focuses on are the global workspace theory, predictive coding theory and the information integration theory. Vraný considers each of these in turn, assessing its potential towards contributing to the empirically grounded theory of the integration unity of consciousness. He holds that each of these theories is suitable for addressing different aspect(s) of the integration unity. His patchwork approach seems to me to be correct: take what's valuable in each of these accounts and put it together. I learned new things from these parts of his thesis and consider his pioneering attempt in this field to be a valuable addition to international consciousness studies.

In the following, I will comment on a couple of points in the thesis that captured my attention. Some of them can be taken up in the discussion at the defense of the thesis.

(1) On Dennett's criticism of "Cartesian Materialism": "Even when dualism is openly discarded, this obviously problematic picture is often tacitly substituted by the no less problematic picture of a material place or a process responsible for conscious seeming (over and above the process of discriminating the represented content itself)." (pp. 2--3) The irony is that most empirically grounded theories of consciousness -- including those that both Dennett and Vraný favour -- explicitly posit a set of localizable neural processes responsible for conscious perception. Global Neuronal Workspace theory is

an example of such a theory. The process of "ignition" of the Workspace is *the* neural mechanism that makes contents conscious. According to the advocates of the GNW theory, it is possible to precisely locate this process both in time and in space. Therefore, it is not clear to me what exactly is supposed to be wrong with the picture of "a material place or a process responsible for conscious seeming" according to Dennett (or Vraný). Without the assumption that particular brain processes in particular brain areas are responsible for the existence of conscious states, most of the cognitive neuroscience of consciousness as actually practised today would simply not exist. Does Dennett just want to say that there is no "central" place where "it all comes together" and becomes conscious? If yes, than let me note that there are theories of consciousness that are not committed to this picture and see mechanisms of consciousness as ditributed in various places in the brain. The "Cartesian Materialism" objection should therefore not target them. However, the Global Neuronal Workspace is *precisely* the sort of place in the brain where it "all comes together" according to the advocates of this theory such as Stanislas Dehaene and Lionel Naccache. Despite this, Dennett thinks that the GNW theory is the best cognitive-neuroscientific theory of consciousness on the market (Dennett, "Are We Explaining Consciousness Yet?", *Cognition*, 2001). This makes his case against "Cartesian Materialism" even more baffling.

(2) On the adaptive value of consciousness: "The project of devising a naturalistic account of consciousness can be convincing only under the assumption that consciousness, like other biological adaptations, has some adaptive function(s)." (p. 29) This is controversial. A case can be made for the claim that consciousness is a *spandrel*, not a biological adaptation (see Robinson, Maley and Piccinini, "Is Consciousness a Spandrel"? *Journal of the American Philosophical Association*, 2015). Suppose that this is true: consciousness is not a biological adaptation. Even so, it could well be that it still has some *causal* functions, because even evolutionary spandrels can have various causal functions. We thus need to distinguish between having an adaptive value (enhancing inclusive fitness of an organism) and having a causal function. I guess that what Vraný says above could be reformulated in these weaker terms: consciousness seems to have at least some causal functions such as, crucially, guiding flexible, non-automatic action. But even this weaker claim can be disputed. Benjamin Kozuch recently argued, with the

help of experimental studies of motion, that (visual) consciousness is an *epiphenomenon* **relative to some types of behaviours that are typically thought to be consciousness-dependent (Kozuch, Consciousness and Mental Causation: Contemporary Empirical Arguments for Epiphenomenalism, forthcoming in Kriegel (ed.),** *Oxford Handbook of the Philosophy of Consciousness***). In particular, Kozuch argues that visual consciousness does not directly drive action (roughly: because action is directly governed by neural activations in the dorsal stream and these need not, perhaps even** *cannot* **become visually conscious). Whether we endrose this argument or not, it shows that to place such a heavy emphasis as Vraný does on consciousness' evolutionary or causal functions while providing an explanation of how consciousness arises in nature is strategically misguided. It is simply not the case that, eg., localization of the neural correlates of consciousness cannot constitute a successful explanation of consciousness unless we "demonstrate that a particular set of neural processes constitute consciousness by fulfilling the adaptive functions in question". (p. 29) It might turn out that the there are** *no* **functions of this kind.**

(3) "Type physicalism is false" (p. 42). This is another controversial claim and I for one would not be prepared to endorse it. Theory of mind-brain type identity (type-TI) is currently undergoing a revival, with authors such as William Bechtel, Thomas Polger or Lawrence Shapiro leading the opposition against multiple-realization-style functionalism and providing new arguments for type-TI. Careful assessment of the functionalist literature, these authors claim, leads one to conclude that as far as the actual empirical evidence is concerned, the case for multiple realization is heavily overstated. However, since Vraný himself argues that a rejection of type-TI does not significantly endanger the project of assembling a neuroscientific account of the unity of consciousness, we do not need to discuss this issue at the viva.

(4) On the basic principles of Predictive Coding (PC) theory: Vraný follows the main exponents of the PC theory (notably Andy Clark) in characterising one of its crucial principles in two ways. Formulation #1: The brain is organized as a hierarchy of areas in such a way that higher-level areas try to predict the activation at lower-level areas (p. 109). Formulation #2: higher level areas model the *causes* **of the lower-level neural**

activity (p. 137) (A variant of this latter formulation is that our conscious perceptions construct the external world by modelling the hidden external causes of our surface sensory stimulations.) It was never clear to me how these two formulations could be taken to constitute a single story. It seems that formulation#1 is putting forward a less ambitious claim than formulation#2. The brain can, in principle, implement the less ambitous "predictions" without taking into account the causes of the lower level activations. All it takes to do that is simply to anticipate, thanks to previous learning, what the lower-level activation will be like. Rodolfo Llinás, in his *I of the Vortex* **(The MIT Press, 2001), canvassess how the brain can achieve this sort of prediction thanks to a relatively simple mechanism drawing on variations in the minute voltage across the cell's enveloping membrane. The second, causes-involving sort of prediction seems to be a more complex process. My hunch is that a relatively robust notion of contentful representation would need to be recruited for this second sort of account. This would be congenial to those versions of the PC theory that trade in representations, such as those of Jakob Hohwy or Paweł Gładziejewski, both cited by Vraný. However, there are other versions of the PC theory that drop representations (see, eg., Downey, "Predictive Processing and the Representation Wars: A Victory for the Eliminativist (via Fictionalism)",** *Synthese*, **2017, or Hutto, "Getting into Predictive Processing's Great Guessing Game: Bootstrap Heaven or Hell?",** *Synthese*, **2017). Any light on this conundrum from Vraný would be most welcome.**

****

Despite my minor reservations and critical comments, I recommend the submitted dissertation with the tentative grade of *pass*. I fully recommend that the title "PhD" is granted to Martin Vraný on the basis of this PhD submission, which is a valuable contribution to the field of consciousness studies.

Tomáš Marvan **Prague, 25/01/2018**