Response to Reviewers (Revision 1)

April 24, 2020

Reply to Reviewer 2

Q1: I had a few problems in reading the paper, which I think should be addressed (especially the first) before the paper is published. For detailed notes on these see below.

1. The coarse-graining idea seems undefined in critical ways, but the paper reads as though it is well-defined. So maybe the authors have compressed too much detail? In the formalism presented it is unclear what coarse-graining (or information closure to other grains) amounts to. With this lack of detail, it is then unclear to me why intermediate maxima in IC are a plausible result (why does IC not just decline with progressive coarse graining).
2. The idea of 'simulation'. This is a more minor point but I would appreciate if the authors could be clearer on this. It is arguable whether or not consciousness simulates anything (e.g. see Hoffman's interface theory); and some of the assertions the authors make re simulation seem unfounded anyways. But it could be that they mean something different or subtle (that consciousness is linked to or represents or etc the environment)
3. There were several other assertions that stood out to me as strange or wrong, but could be a matter of explanation or ignorance on my part (the feedforward system, the cut-apart system). Either way some more explanation would be good.

A: We thank the reviewer for these critical comments. We have replied in detail in the following Q&A.

Q2: The English is not perfect and needs some good proofreading, though it was understandable throughout.

A: We have recruited a professional proofreading service to edit the manuscript. Please inform us if it is still not clear or unreadable.

Q3: On line 182, defining C-processes as cases where 'Y is informationally closed to X'. This is now information closure in the sense of coarse-graining, but no such thing has been defined yet? Is it supposed to be so trivial that it is not required to make it explicit? To this point, closure is defined entirely in terms of system with respect to environment.

Basically, I am not sure whether closure between scales is a matter of same-time (e.g. $I(Y_{t+1}; X_{t+1})$) or across-time interactions ($I(Y_{t+1}; X_t)$). Or could it be both?

A: We agree that this point is not clear. Even though the Bayesian graphs for the system-environment dependency differ from that for the micro-macro level dependency, we applied the same definitions of information flow ($J_t(X \to \mathrm{Y})$) and informational closure ($J_t(X \to \mathrm{Y}) = 0$) to both graphs.

We have added a paragraph to make this point explicit.

> **Line 205:**
> *" Note that, here we applied the same definitions of information flow (Eq. 1)*
>
> $$J_t(E \to Y) = I(Y_{t+1}; E_t | Y_t) \tag{1}$$

to the system-environment dependency and the micro-macro scale dependency

$$J_t(X \rightarrow Y) = I(Y_{t+1}; X_t | Y_t) \tag{2}$$

even though the Bayesian graphs differ in the two scenarios. Both these settings have been previously used in the literature (see Bertschinger et al., 2006; Pfante et al., 2014b).
"

Q4: On line 241: '.. not sufficiently coarse-grained variables have low values of NTIC', this is phrased as though it is necessarily true (and also that 'we saw above', though I do not see above where this is justified), but is it?

At the very finest grain, wouldn't even all the most 'stochastic' dynamics of a system be accounted for by prior states of the system, or by the environment? Why wouldn't we assume that NTIC *only decreases* as the system is coarse-grained (as number of elements/states is decreased), for similar reasons as it must be zero at a 1-element 1-state system as the authors note? This seems an important point to be very clear on…

A: We appreciate that this section in the previous version was unclear. We have therefore added a new paragraph and significantly revised this section. In short, we more formally describe how NTIC may change non-monotonically from the finest microscale to the coarsest macroscale. Please see our revised manuscript from Line 261 to 277.

Q5: on line 263: 'NTIC processes encodes environmental information in its state. This suggests that a NTIC process can be considered as a process that simulates the environmental dynamics.

Why does encoding suggest simulation? An encoding *could* be a simulation, but it could well be nothing like a simulation, if we understand simulation to mean something like an imitation or reconstruction of some structure or dynamics. Two totally different environmental situations could potentially be encoded in exactly the same way (eg. as a string of 1s and 0s).

The 'simulation' notion is brought up again in section 5.2 (line 344), citing Bertschinger et al, though I do not find the idea in that paper.. I think authors need to be clearer on what they mean by 'simulation' to make this point.

A: The notion of simulation is the same as the 'Modeling' case on Page 4 in Bertschinger *et al.* (2006):

> *"B2) Modeling: The system reaches synchronization and internalizes the correlations observed in the environment by building up own structures."*

To avoid reader confusion, we have added a new section, Sec. 2.1, in this revision to describe the modelling scenario. Further, we have replaced "simulation" and "encoding" with "modelling" in many parts of the text. We hope that this change clarifies our point.

Q6: line 306: 'Blindsight patients… make above chancel-level visual judgments without having any conscious perception about visual stimuli' I am not sure this characterization of blindsight is correct, it may be a matter of the phrasing. It could be - or probably is - the case that 'blindsight' patients always have some conscious experience of what drives their correct responses, but that the 'visual character' of these experiences is degraded or missing. (Overgaard, Experimental Brain Research 2011) for a specific example (Mazzi, Bagattini, Savazzi; Frontiers in Psychology, 2016)

A: Yes, we specifically wanted to address the visual character of their experience. We agree that some patients still preserve some forms of conscious experience. We have changed our sentence and cited the relevant references as follows:

Line 335:
Original:
" Blindsight patients are able to track objects, avoid obstacles, and make above chance-level visual judgements without having any conscious perception about visual stimuli. "

Revised:
" Blindsight patients are able to track objects, avoid obstacles, and make above chance-level visual judgements with degraded or missing visual experience.(However, in some cases, they may still preserve some forms of conscious experience; See Mazzi et al. *(2016); Overgaard (2011)) "*

Q7: on Line 316: On the issue of a feedforward network, the current state of a feedforward network with more than one layer is certainly driven by its own past states! So, mutual information of a feedforward network over a time lag should not be zero, unless I am misunderstanding something here? At the same time, I can see a version of the authors' argument here - for a unit in a feedforward network with depth (distance-from-input) of N, a time lag of N time steps would always account fully for the states of the element. Is this the idea?

A: A: The reviewer is correct. As we have now clarified that in the case of a deterministic observation process a feedforward network (or a copying process) can be informationally closed and yet also maintain mutual information such that it is conscious according to ICT. We call such processes passive adaptation processes and have introduced this in the new Sec 2.1. We mention the possibility of conscious feedforward networks now in Sec 5.1 and have added a footnote specific for n-layer feedforward networks at Line 348. Note that most observation processes in the real world are not deterministic so that most passive adaptation processes are not conscious.

Q8: on line 441: under the 'Prediction after system damaged', it is suggested that ICT predicts cutting a system in half would render both halves unconscious. But this would only be the case if neither half contains its own C-processes, yes? Since ICT allows that many C-processes can exist at the same time, it would have to be some special case for this prediction to hold true. So it seems to me the prediction is actually similar to that of IIT.

A: We assume that we have one conscious (NTIC) process which involves information in both brain hemispheres. This process is informationally closed only when we consider the information in both hemispheres. We agree that if both hemispheres each have their own NTIC process ICT should predict no change in conscious experience before and after cutting. Cutting should not make any difference because they are informationally closed with respect to each other. We also agree that this prediction is relatively premature. Systematic comparisons between model predictions can be done by rigorous modelling studies in the future. We have weakened our statement by added the following sentence:

> **Line 482:**
> *"Nevertheless, this prediction is relatively premature. In the future, rigorous modelling studies will allow systematic comparisons between model predictions."*

Q9: Line 540: the theory doesn't really seem to intend to solve the hard problem(s) at all, much less 'completely solve'. I was expecting the problem of dreaming to come up in the last section (maybe along with SMC for which it is a serious problem). When dreaming the information between environment and the system is virtually zero; is this a problem for ICT? Also, more specific phenomena like the eigengrau - if you take away all visual input, for long enough, I do not lose my visual experiences - rather they take on a special state. Does ITC need to accept the possibility that even *trivial* IC can be a conscious process?

A: We appreciate the reviewer for bringing up the hard problem of consciousness and dreaming. This is a very good question, and dreaming clearly poses a challenge for ICT. At the moment we can only speculate about possible solutions. One point we wish to emphasise is that not all the processes in the neural system are NTIC processes. To the conscious (NTIC) process, the rest of the neural system should be considered as part of the environment. This notion is briefly indicated at Line 212 in the original manuscript. We speculate that, during dreaming, the neural system stably forms the NTIC process with respect to its environment, i.e. other parts of the neural system. The same idea can be also applied to phenomena like Eigengrau. At present, however, this is mere speculation so we refrained from making the statement in our last version of the manuscript. However, we again thank the reviewer for raising this important question and believe that it is still worth mentioning. Accordingly, we have added a short paragraph as follows:

> **Line 614:**
> *"Explaining conscious experience during dreaming is always a challenge to theories of consciousness. ICT currently does not have a specific answer to dreaming. However, we wish to emphasize that not all processes in the neural system are NTIC since some processes are not informationally closed. They mainly passively react to sensory inputs or other processes in the neural system. To the conscious (NTIC) process, the rest of the neural system and the body should also be considered as part of the environment. They retain some degree of activity during sleep and dreaming. We speculate that, during dreaming, the neural system stably forms an NTIC process with respect to its environment, i.e. the other parts of the neural system. At present, however, this remains mere speculation. Identification of the NTIC process(es) during dreaming is an important milestone in extending the scope of ICT."*

Finally, we believe that this is a good empirical question that can and should be tested in future studies.