Response to Reviewers (Revision 1)

April 24, 2020

Reply to Editor

Q1: In addition, I have a question. if we view molecules as coarse grains of elementary particles, would the phenomenon of molecular recognition have high NTIC and thus be conscious?

A: Hi Prof. Hoffman, thank you for your attention to our manuscript and your interesting question. If we understand correctly, molecular recognition refers to specific types of interactions among molecules that form molecular complexes. There is no question that molecular complexes are coarse-grainings of molecules via molecular recognition. Based on our hypothesis, if the dynamics of a molecular complex are an NTIC process, the information closure theory of consciousness (ICT) would hold that the molecular complex has a certain degree of consciousness.

Reply to Reviewer 1

Q1: Regarding the NTCI measure:

a) Bertschinger et al. 2006 define informationally closed as $J_t(E \to Y) = 0$, but this is not a requirement for NTIC. Meaning, a positive NTIC measure does not entail that $J_t = 0$. $J_t = 0$ is only true if the NTIC is equal to $I(Y_{t+1}; E_t)$. In what sense does it correspond to non-trivial information closure then? (see also my concern #2, maybe that is related) As defined in eq. 6 a video camera might have very high NTIC if it records a natural scene, even though it is in no way informationally closed from the environment.

b) Is $NTIC_t$ state dependent or not? As capital letters are used I assume it is based on (conditional) mutual information measures across one time step but averaged over the possible states of the system at time t and t+1, so state distributions. But which distributions are used to compute $NTIC_t$? The stationary joint distribution of states at t and t+1? It cannot be just some observed measure, because then the level of consciousness would depend on how long the observation is. Moreover, if $NTIC_t$ is not state dependent, then the Reflex example on p. 10 is not intuitive. Because the system (brain) is the same and awake whether it behaves reflexively or not, so while the content may depend on the type of action, the $NTIC_t$ (level) would stay the same.

A:
a) It is correct that the information flow term $J_t(E \to Y)$ does not have to be zero for $NTIC > 0$. However, in Definition 2, a C-process is informationally closed to the microscopic universe process X. This implies that the C-process is also informationally closed to any other coarse-graining of X, including E (see also our reply to Q2). Currently, we only consider that closed NTIC processes are C-processes. Regarding the video camera example, please see more discussion in our reply to your Q6.

b) It is correct that NTIC is not state-dependent. We agree that state-dependent measurements are essential to describing the dynamics of conscious experience. We are working on state-dependent formulations for the next version of the ICT. In contrast to the present version, this work involves point-wise informational measures and requires more space and discussion. We would therefore prefer to address the state-dependent version of ICT in future studies and keep the first version of our theory simple.

Nevertheless, we appreciate that the reviewer's raising of this issue, and have added a short paragraph about it:

> **Line 610:**
> *" In this article, we do not use a state-dependent formulation of NTIC. However, we believe that state-dependent NTIC is essential to describing the dynamics of conscious experience. The next version of ICT therefore requires further research using point-wise informational measures to construct state-dependent NTIC."*

Q2: Hypothesis and Implication 3: Why is it crucial that there is an underlying X with respect to which Y is a C-process?

A: The C-processes are constructed to answer questions about the scale problem of consciousness. Their definition therefore states their relation to the microscopic scale X. At the same time it would be possible to remove X from the definition and instead require that Y is informationally closed with respect to all other processes. In some cases such a definition would be more general than the present one for example when there is no final microscale but instead an infinite number of more and more microscopic scales. However, we find the use of an explicit microscale X easier to understand and sufficient to communicate the idea. Note that closure with respect to X (if it exists) implies closure with respect to every coarse-graining of X. To grasp this, note that the information transfer from any coarse-graining (including E and S) of X to Y must be lower than that from X to Y, i.e. $I(Y_{t+1}; X_t|Y_t) \geq I(Y_{t+1}; S_t|Y_t)$ and $I(Y_{t+1}; X_t|Y_t) \geq I(Y_{t+1}; E_t|Y_t)$. Accordingly, Y is informationally closed with respect to X implies that Y is also informationally closed with respect to S, E, and all other coarse-grainings of X.

Q3: Does Y being informationally closed with respect to X imply that Y is also informationally closed with respect to E, i.e. $J_t(E \to Y) = 0$? I don't think that could be true for the brain, or part of it really. My brain's next state, at whatever level, certainly depends to some degree on unpredictable sensory inputs from the environment. In the end, is $I(Y_{t+1}; E_t|Y_t) = 0$ required for consciousness or not? (see point 1a). In other points in the manuscript this does not seem required, e.g. p. 10 "If a process is not informationally closed, the degree of NTIC is low resulting in low or no consciousness".

A: Yes. As mentioned in our answer to Q2 Y being informationally closed with respect to X implies that Y is also informationally closed with respect to E, i.e. $J_t(E \to Y) = 0$. As explained in Question 1a, we currently only consider informationally closed processes as C-processes. The reviewer is also correct that our internal states depends to some degree on unpredictable sensory inputs from the environment. This results in some interesting predictions. For example, unpredicted sensory inputs may temporally break informational closure and ICT predicts that conscious agents may then lose consciousness for a short time. Unfortunately, referring to your Q1b, the current version of ICT does not include state-dependent IC and NTIC. Therefore, the current formulation of ICT cannot well describe the dynamics of conscious experience under unpredictable states. We expect that our next version of ICT will be better able to answer your question.

Q4: Exclusion or no exclusion? The authors state multiple times that "every process with a positive non-trivial information closure (NTCI) has consciousness." (p.3, and also p.12). a) Yet, Figure 4 is ambiguous in the sense that the maximum somehow seems important, while actually all levels should give rise to separate experiences. I don't see how ICT without something like exclusion (IIT style) would indicate that the maximum should correspond to human consciousness. It could just be any one out of many consciousnesses. b) p. 14: comparison to IIT: what does ICT say about different sets of variables at the same level. There might well be multiple ways to partition X into S and E, with many overlapping S's. Would those all be conscious? They would not be informationally closed from each other but they could all fulfill the requirements in the Hypothesis. It seems like finding the maximum within a level over the possible sets of elements/variables is a necessary step

to identify boundaries (I think also Krakauer does that across sets of variables, but not 100% sure)

A: We can see how Figure 4 was ambiguous with respect to this question. We do not impose the exclusion criterion in ICT and have changed Figure 4 to remove the impression that it does.

Instead of imposing the exclusion criterion, the concept of informational closure itself already involves the notion of individuality and defining the boundary of a system. (Bertschinger *et al.*, 2006). This allows ICT to solve the individuality problem of consciousness in some cases (but not all; please see the paragraph in Limitations from Line 589) without imposing an exclusion criterion. Therefore, it is true that Krakauer's boundary detection procedure is potentially helpful and can be adapted for ICT in the future.

This is also why ICT allows the existence of multiple consciousnesses across the scales of coarse-graining (which IIT does not). Because conscious processes are informationally closed from each other across the scale of coarse-graining, they are not able to know the existence of other consciousnesses within the same system (information flow from other processes is zero). This is in line with using informational closure in level identification (Pfante *et al.*, 2014b).

We understand that not solving individuality completely is a major weakness of the current version of ICT (and in fact of all IC- and NTIC-based theories). More work on this issue is needed in our future research.

Q5: Feedback and memory: p. 11. If $NTIC > 0$ is sufficient and true information closure is not required, then feedfoward networks can be conscious according to ICT. Also, in that case, memory is not necessary, as the video camera would have $NTIC > 0$ as long as it records from a stable environment. In that case $I(Y_{t+1}, Y_t) > 0$ and $I(Y_{t+1}, Y_t|E_t)$ is small. This issue here is that having information about the next state does not imply that there is any causation, so if Y_t is highly correlated with E_t and E_t causes Y_{t+1} then NTIC is high. (Thus the IIT emphasis on causation). If somehow the fact that Y must be informationally closed wrt X does the heavy-lifting here that should be made more explicit (see above).

A: The reviewer is correct. We have reconsidered the case of the feed-forward networks and the video camera example (which can be considered as a single-layer feed-forward network). According to ICT, feed-forward networks can be conscious for deterministic sensor processes. We thank the reviewer for correcting this point. In this case also, a video camera can achieve positive levels of NTIC and, therefore, be conscious. We have worked extensively on Sec. 2.1 and Sec. 5.1 to make this clear. In summary, we have mainly revised our manuscript as follows:

In summary, we mainly revised our manuscript as follows:

- Added Section 2.1 to illustrate two NTIC scenarios about causality.
- Added sensory channels in Fig. 2
- Extensively modified the paragraphs about feed-forward networks and memory in Section 5
- Added a session in the Appendix to prove that NTIC achieved by modelling the environment can potentially achieve higher NTIC than the one which copies sensory values (e.g., the video camera).

Q6: Does ICT imply that one is unconscious while dreaming? In that case $I(Y_{t+1}; Y_t|E)$ should almost be identical to $I(Y_{t+1}; I_t)$ and thus lead to NTIC approx. 0.

A: We want to emphasise that not all processes in the neural system are NTIC processes, since some processes in the neural system are not informationally closed. To the conscious (NTIC) process, the rest of the neural system is considered as part of the environment. This notion is briefly indicated at line 212 in our first version of the manuscript. These processes are still active during sleep and dreaming. We speculate that, during dreaming, the neural system can stably form the NTIC process with respect to its environment, i.e. other parts of the neural system. At present, however, this is mere speculation so we restrained ourself from making the statement in our manuscript. However, we refrained the reviewer for bringing up this important question. This is also an empirical question

that can and should be tested in future studies. We believe that this is still worth mentioning. 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."*

Q7: Minor:

1. "contributions to the field" This section is just a copy paste from the abstract and thus not necessarily helpful. Not sure though what the purpose of this section is meant to be by Frontiers.
2. Introduction: "We currently lack a theory... " IIT and the geometric theory of consciousness (Fekete et al.) have proposed solutions. So there are theories. The following paper should be of interest and should probably be cited: Fekete T, van Leeuwen C, Edelman S (2016) System, subsystem, hive: Boundary problems in computational theories of consciousness. Front Psychol 7:1041.
3. y_t corresponds to the content. I would argue that an activation state, without taking the system or relation between elements etc into account is meaningless and cannot capture/explain the structure of an experience. Though this issue can be dealt with at a later time.

A:

1. In fact we were also confused about this during submission. We have modified this as follows: *"We propose a new theory of consciousness and discuss some preliminary implications."*
2. We thank the reviewer for the reminder. We have modified our manuscript and cited the relevant references.

   Line 75:
   Original:
   " We currently lack a theory to identify the scale which conscious processes correspond to. "

   Revised:
   " We currently have only few theories (e.g., Integrated Information Theory (Hoel et al., 2016) and Geometric Theory of consciousness (Fekete & Edelman, 2011, 2012)) to identify the scale which conscious processes correspond to (also see the discussion in Fekete et al. (2016)). "

3. We agree with the reviewer's comment. In fact, we do not differentiate states of processes between activation and relation. To take neural networks as an example; : the relations between elements are is determined by the topology of a the network (structure) and its hyperparameters (e.g. activation functions). We agree that the activation and relation of a neural network should be both be considered. After all, they are all of them merely the parameters for the process of the neural network. The only difference between the two groups of the parameters is the temporal scales of changes. States of activation may have more rapid dynamics than states of relation. In ICT, the two sets of parameters do not have any qualitative differences.

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