Making progress on the prefrontal debate

Matthias Michel, Center for Mind, Brain, and Consciousness, New York University

Rafael Malach, Department of Brain Sciences, Weizmann Institute of Science, Rehovot, Israel

Authors' version. Published in the *Journal of Consciousness Studies*. Please cite the published version.

The role of prefrontal cortex in conscious perception: a joint proposal

As can be evident from the opposing prefrontalist and localist views presented in the preceding two reviews- the current experimental evidence is, unfortunately, still not conclusive enough to settle the debate concerning the specific role of PFC in conscious experience. To recount briefly the two opposing notions—the localist perspective views the entire cortical mantle as a mosaic of local cortical areas—each underlying a different and unique category of conscious contents. By contrast, the prefrontalist view, at least the one outlined here—proposes a general, enabling function for PFC for all types of conscious experience. Specifically, following higher-order theories of consciousness, this role is derived from the proposal that conscious experience includes an essential metacognitive element—without which perceptual states remain unconscious. The PFC is proposed to underlie this obligatory metacognitive element, and hence must participate in each and every kind of conscious experience.

While these opposing views may seem irreconcilable—we, the authors of these opposing reviews, actually agree on the methodology that is needed to resolve this major debate. Furthermore, we also agree on the kind of future experiments that hold the potential to refute one or both of these theories. While these experiments are far from being applicable to current research in human neuroscience, we feel that it may be fruitfully applied in the near future in animal models, and further down the line also in human patients as part of beneficial clinical procedures.

Agreement: the critical need for a convergence of correlational and causal methods and results.

A central methodological approach that we believe could greatly enhance the power and decisiveness of human experimental data in the search for the neuronal basis of consciousness is the insistence on converging multi-modal and rigorous methodology.

Correlational methods, such as obtaining fMRI data during conscious and non-conscious conditions, have been useful to provide preliminary evidence on the neural bases of consciousness. But we also recognize that these methods are severely limited. In particular, null results—such as failures to find an increase in PFC activity in the conscious condition, or failures to find differences in visual areas between conscious and non-conscious conditions—are difficult to interpret. Future research should

instead strive to look for convergences across methods, using brain imagining but also causal methods, such as the disruptive impact of cortical lesions, the effects of electrical stimulation of cortical regions, and even the specific symptoms associated with pre-epileptic auras.

It should be acknowledged that when considering the human neuroscience methods—the state of the art, for obvious reasons, is dramatically inferior compared to animal models. Thus, the methodological tool kit available to human neuroscience is unfortunately extremely limited in its spatial and temporal resolutions, both when studying cortical activation patterns and particularly when trying to causally manipulate or shut-down such activations. The only cautious strategy that can be adopted, until this methodologically blurred landscape is cleared up, is to insist on convergence of many lines of evidence on the one hand, and on reproducibility, and rich and detailed information of each case on the other hand.

Animal models can of course be considered as providing important supporting evidence to human-derived theories. Pioneering research on binocular rivalry, or research on the phenomenon of blindsight in animal models, for instance, has been highly significant and informative. However, such experimental findings in animal models must be taken with extreme caution since, unlike human participants, it is impossible to gain a rigorous insight into the conscious experiential state of the studied animal. This puts a severe limit on the interpretability of animal models when the question turns to the neuronal correlates of an animal's subjective conscious experience.

Proposed experiment I: Transiently shutting off functionally targeted PFC regions.

The central disagreement between the proposed localist and prefrontalist perspectives discussed here is, unsurprisingly, manifested in the expected consequence of blocking PFC activity. The proposed localist view considers shutting off PFC to be of no major consequence to perceptual experience. By contrast, the prefrontalist view considers PFC activity to be an essential aspect of all conscious experience including conscious perception—hence predicting that blocking PFC activity will abolish conscious perception altogether, without necessarily affecting unconscious perceptual processes.

Following our suggested multi-dimensional approach—stressing the need for converging methodologies, we propose here as a general strategy the combination of functional mapping and the targeted blockage of activity in the identified PFC areas. Our proposed experiment will proceed along three central experimental stages. First, the candidate PFC areas that are expected to play an essential role in conscious perception will be identified. Second, activity in the targeted PFC areas will be transiently blocked while the participants are exposed to vivid visual stimuli. Finally, a recall task will be conducted immediately following the PFC blockage to assess whether the participants were conscious of the stimuli.

We expect this experiment to yield informative results that may distinguish between the localist and prefrontalist perspectives—because these views lead to diametrically opposing predictions. While the localist perspective accepts that blocking PFC activity may suppress the ability of participants to introspect and report on the visual content presented *during* PFC blockage, it predicts that participants will report having consciously perceived the stimulus once given the possibility to report what they experienced during PFC blockage. According to the localist view, this situation is rather

similar to the natural case of absorption—such as during an engaging movie—in which introspective ability is suppressed yet we are perfectly capable of recalling having experienced the engaging movie.

By contrast, prefrontalist theories will predict that all conscious experience should be eliminated due to the PFC suppression—leading to the inability of participants to become conscious of the stimuli and as a consequence their failure to report experiencing the stimulus while PFC suppression is maintained. Critically, according to the prefrontalist view, once PFC suppression is removed, participants should report not having experienced the relevant stimulus. One important caveat should be mentioned as far as the prefrontalist prediction is concerned. Prefrontalists do not necessarily predict that participants should be unable to recall the identity of the stimulus in a forced-choice task, as that information could be encoded unconsciously. Hence, the report should instead be free, subjective and focusing on whether or not the subject had a conscious experience of the stimulus during PFC suppression. We will now outline the experimental design in some more detail.

I. Functional identification of candidate PFC regions. It should be noted that the functional mapping of the candidate PFC regions largely depends on the specific version of the prefrontalist hypothesis that is tested. We suggest starting from correlational measures to identify the relevant PFC regions. Contrasting conscious versus unconscious perception reveals candidate PFC regions. These regions can then be manipulated bilaterally. In particular, re-representationalist theories (see the review 'Consciousness and the Prefrontal Cortex' in this issue) hold that specific contents are re-represented in PFC. Targeting areas where those re-representations can be decoded should thus prevent conscious experience of the relevant stimulus. As for relationalist theories (again, see 'Consciousness and the Prefrontal Cortex' in this issue), we identify two possibilities: one can either target the areas that show an increased activation during contrastive analysis, or directly attempt to prevent feedback from those areas of PFC to the relevant visual areas. It should be noted that if this latter option is preferred, re-representationalist and relationalist prefrontalist views make opposite predictions, as re-representationalists see no constitutive role for feedback in consciousness.

II. Transient inactivation of targeted PFC regions. It should be noted that well-controlled and precisely timed inactivation of targeted cortical areas can be performed at present, only in animal models. However, there are currently inaccurate proxy measures that are conducted for clinical purposes in humans that may be considered- two particularly relevant ones are the WADA test in which the frontal lobe is unilaterally anesthetized in its entirety. The second approach is repetitive TMS conducted routinely for alleviation of depression. However, it should be emphasized that both these methods are far from being optimal in terms of precision and timing. Thus, our proposal is meant more as a suggestion for future potential experiments when the clinical tools will likely achieve the level of precision and sophistication currently enjoyed only in the animal model field.

III. Recall. This stage is a well-established and thoroughly studied procedure in human cognitive neuro-science, so doesn't need much elaboration here. We propose to employ free visual recall—where participants are merely asked to freely come up with the images they saw during the PFC suppression session. The localist prediction is of essentially no degrading effect on such recall due to the prefrontal blockage compared to consciously perceived material. Prefrontalists predict that subjects will report not having consciously perceived anything in the free recall. As noted above,

it is nevertheless possible that subjects might be able to identify what they saw under forced-choice conditions due to unconscious memory signals—a phenomenon akin to blindsight.

Experiment II. Functional mapping and selective blocking of local recurrent connectivity in visual and PFC cortex.

This study complements the first experiment in that it targets the local cortical structures hypothesized to be essential by the localist theories while leaving the global networking intact. Thus, localist theories propose that the critical element that is obligatory for the emergence of conscious content is the local, recurrent, integration of information—while prefrontal theories hypothesize it is the global connections, either between those areas and PFC, or between PFC and back, that are essential.

The experiment will again include three stages, with a similar design as in experiment 1. In the first, functional mapping stage, the relevant visual areas will be identified by presenting the participants with different categories of vivid visual images, such as faces, places and objects, and the posterior cortical regions activated by these images will be identified using e.g. fMRI. The relevant PFC regions will be identified in a similar fashion to that described in the first experiment.

As to the second stage of the experiment, involving blockage of local recurrent connectivity, it should be emphasized that no such procedure is currently available neither in animal models nor for clinical treatment in humans, and hence this experiment can be considered only as a future possibility— conditional on its development in animal models. However, the blockage of local recurrent connections could turn out to be valuable as a less damaging procedure for treatment of epilepsy, so it is conceivable that such procedures may become available in the future.

At present the experiment is suggested more as a futuristic possibility aimed to highlight the different experimental predictions stemming from the prefrontalist vs. localist perspectives.

Specifically, in the second stage, similar to the first experiment, participants will be presented with the visual images and will be asked to describe them. Critically, while performing the task, the local connections within the visual areas will be blocked. Here the extent of blockage will depend on the version of localist theory tested. For example, theories proposing recurrent top-down activation across visual cortical areas will necessitate blockage of such cross area connectivity. By contrast, local theories proposing the dependence of conscious experience on local geometries within each cortical area will predict an effect stemming from more localized inactivations, sufficient to disrupt these local geometries. A similar, complementary, blockage will be applied to the local, recurrent connections of the relevant PFC regions.

Following the inactivation session, and similar to experiment 1, participants will be asked to freely recall whether they consciously experienced the relevant images.

The prediction of the localist view will be of a drastic dissociation between the PFC and visual cortex blockage effects. Thus, blocking of local recurrent connections in posterior visual areas is expected to prevent the emergence of conscious visual percepts, despite the global flow of information carried by feed-forward and feed-back cross-areal connectivity.

By contrast, re-representationalist prefrontalist views predict that, as long as the relevant information is forwarded and re-represented in PFC, subjects should have a conscious experience. Meanwhile, predictions of relationalist prefrontalist views differ depending on the exact modalities of those views. Some of them might agree with the localists that the relevant recurrent processes are *necessary* for conscious experience, while not being sufficient. One way to test these views would be to specifically block feedback connections between PFC and the relevant visual areas—an experiment akin to Experiment I above.

In Summary

We believe that a decisive source for the fundamental disagreements that plague the neuroscience of consciousness is not conceptual or philosophical, but instead experimental and methodological. For obvious reasons the experimental data derived from human studies is still inaccurate and unreliable. We therefore feel that an optimal way forward is by insistence on deriving conclusions only when converging lines of evidence- correlational, such as brain imaging data and causal- such as lesion and stimulation data are considered together, in the context of large, repeatable and carefully analyzed data sets.

We outline here, being fully aware of the futuristic nature of these proposals- two major experiments aimed to test the different predictions of the localist and prefrontalist perspectives. These experiments, we acknowledge, cannot be performed using state of the art methodologies in human clinical neuroscience. They are brought here more as future possibilities and as an illustration of the diametrically, potentially testable, predictions of the two theories. It will be intriguing to see how future research comes out in support of one or the other theories- or, quite likely, which will be perhaps the most exciting option- a totally different theory- completely unanticipated by our current thinking.