The Color Phi Phenomenon and Predictive Coding: Challenging Dennett's Indistinguishability Argument

Sahaj Singh Maini

May 2024

Abstract

In the book "Consciousness Explained" by Dennett (1993), the author proposes his theory of consciousness called the multiple drafts model. He utilizes the color phi experiment to support his theory and argue against the existing Cartesian theater model of consciousness. The color phi phenomenon experiment exhibits a perceptual illusion in the subject. It involves watching a screen where two frames are shown in rapid succession. A red dot appears in the first frame, and a green dot appears in the second frame. When the dots are shown with appropriate spacing and timing, the subjects report a sensation of motion and change of color in the middle of the illusory path of the dot. Dennett suggests that if there was a "time and place" for conscious experience (as proposed by the cartesian theater model), it would be possible to distinguish, in principle, whether the phenomenon is generated before or after the conscious experience. However, he argues that it is not possible to distinguish between them as they account equally for the available data (including the color phi experiment). In this essay, we will discuss the predictive hypothesis for perceptual illusions and reflect on the respective argument proposed by Dennett.

The Multiple Drafts Model vs. The Cartesian Theater Model

Cartesian theater is the defining aspect of what the author calls cartesian materialism, which he claims is the remnant of cartesian dualism. René Descartes proposed cartesian dualism, where the central claim is that the immaterial mind and the material body, while being two distinct entities, causally interact. This causal interaction was proposed to occur through a central location in the brain, the pineal gland. The author suggests that when the idea of the immaterial mind is removed from cartesian dualism, we are left with a model where the mental phenomena is still assumed to occur in a centralized region in the brain. In other words, this central location is a singular point in the brain where sensory information is integrated and consciousness arises. This central region is analogous to imagining a theater where a homunculus, as audience, observes the sensory information. The author suggests that although the notion of immaterial mind has been discarded, the residual idea of a central seat of consciousness is a lingering influence of cartesian dualism.

According to the cartesian theater view of consciousness, there exists a boundary somewhere in the brain, where conscious experience occurs. The order of arrival of the stimuli to this finish line equals the order of presentation in conscious experience. As

a consequence, this view suggests that there is a single, unified stream of consciousness that represents our experience from one moment to another.

On the other hand, the multiple drafts model suggests that there is no fixed time and place where conscious experience occurs. Instead, the contents of consciousness are distributed across the brain, in both space and time. As a result of this distributedness, probing a subject's conscious stream at different places and times elicits different effects and narratives.

According to this model, all the information entering the brain gets processed in a parallel, distributed manner with elaboration and interpretation occurring in the brain, in multiple areas, all at the same time. Additionally, information only has to be discriminated once and once it is discriminated, it gets passed around undergoing continuous editorial revision. As a result of being passed around and continuously revised, at any given point in time, there are fragments of multiple narratives at various stages of editing in various parts of the brain. These narratives are considered to be the brain's interpretations or explanations of the sensory input and internal states.

As a consequence, there is no stand-alone defining narrative that represents our stream of conscious experience. Instead, our stream of conscious experience is a result of multiple narratives (that the author calls drafts) undergoing continuous editing and revisions. This view of consciousness emerging from multiple, parallel and distributed processes in the brain challenges the single, unified stream of processing. It suggests that our subjective experience is complex and ever-changing subject to the interplay of various neural mechanisms.

Color Phi Experiment

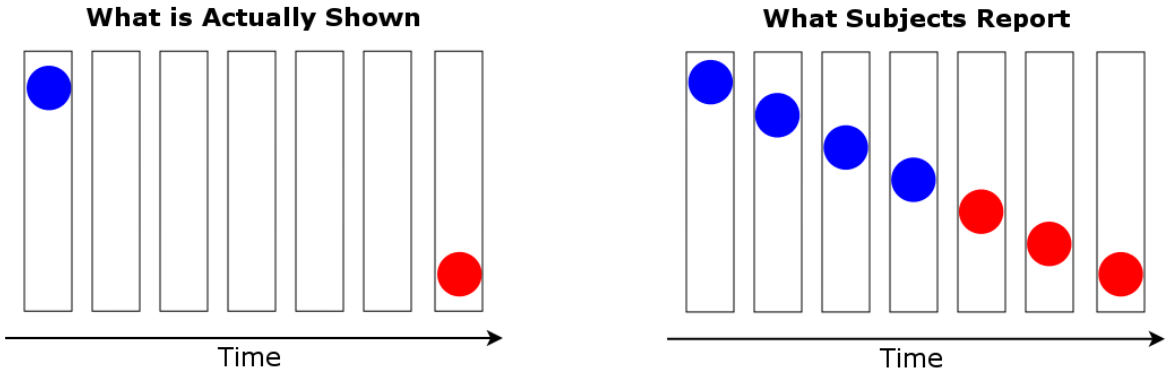
The color phi phenomenon is a perceptual illusion that was first described by Kolers and von Grünau (1976). It is a complex version of the phi phenomenon that was described by Wertheimer (1912), which refers to the observation that the presentation of a sequence of static images at high frequency can be perceived as continuous movement (also known as "apparent motion").

In the classic color phi phenomenon experiment, a subject watches a screen where two images are projected in succession. The first image contains a red dot at the top of the frame and the second image contains a green dot at the bottom of the frame. The images are presented in succession for a specified interval of time with specified spacing between the red and green dot. At certain combinations of spacing and timing, the subject perceives apparent motion. What distinguishes the color phi experiment from the phi experiment is the fact that along with apparent motion, the subject also perceives a change in the color of the dot, abruptly, along the illusory path as shown in 1.

The color phi experiment shows that the brain actively "fills in" the missing information to create a coherent perceptual experience. Dennett uses the color phi experiment to argue against the cartesian theater model and to support his alternative multiple drafts model as we will see in the next section.

Stalinesque vs. Orwellian Hypothesis

In the context of the color phi phenomenon, Dennett illustrates two possible explanations from the cartesian materialist point of view, for how the brain constructs the perceptual



(a) What is actually shown

(b) What is actually seen

Figure 1: Illustration of color phi phenomenon (from Wikipedia contributors (2023)). As per the example in the book, the blue and red dot need to be replaced with red and green dot respectively.

illusion of motion and color change. The terms "Orwellian" and "Stalinesque" are borrowed from the novel "1984" by George Orwell.

In the Orwellian hypothesis, the red and green dots are processed separately and what follows is a post-experiential revision. In other words, after the subject becomes conscious of the green spot, both these experiences (red and green dot) get wiped from the memory. They get replaced by a revisionist record of the red dot moving to the location of the green dot and changing the color on the way. Under this hypothesis, the conscious experience of the color phi illusion occurs after the actual events (the flashing of the spots) have been processed and the memory misrepresents what actually occurred in the experience after the fact.

Alternatively, in the Stalinesque hypothesis, similar to how Stalin's regime would censor information after being released to the public, the information gets edited or pre-processed to "fill in" the missing information before passing the edited version up to consciousness for a "viewing". This hypothesis suggests that the brain actively manipulates and edits incoming sensory information in order to create a coherent narrative that can then be fed to consciousness. This form of manipulation occurs unconsciously, and by the time it reaches our consciousness, the information seems seamless and complete.

It is now important to understand the condition under which the above two hypotheses undermine the cartesian theater model, which, according to the author, is the inability to differentiate between them. This inability to distinguish between them challenges the idea of a clear boundary between pre-conscious and post-conscious processing. It challenges the idea of a location or moment of conscious experience. If conscious experience occurs in a definitive time and place, then, consequently, we should be able to determine the sequence of events. The following text from the book expresses the author's argument -

"Both models can deftly account for all the data - not just the data we already have, but the data we can imagine getting in the future. They both account for the verbal reports: One theory says they are innocently mistaken, while the other says they are accurate reports of experienced mistakes. Moreover, we can suppose, both theorists have exactly the same theory of what happens in your brain; they agree about just where and when in the brain the mistaken content enters the causal pathways; they just disagree about whether that location is to be deemed pre-experiential or post-experiential."

The author assumes that both theories are indistinguishable, not just for the existing

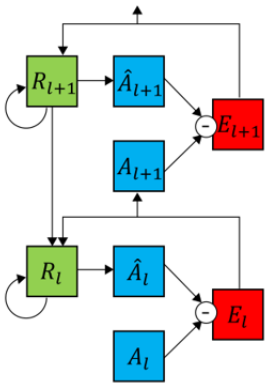


Figure 2: Â represents prediction, A represents incoming sensory input, R represents recurrence, and E represents the error calculation between predicted and incoming input. (from Lotter et al. (2016))

data but also for the data that may be collected in the future. In the context of the color phi experiment, however, it may now be possible to distinguish between the hypothesis that explains the phenomena. In order to rule out one of the hypotheses, namely the Orwellian hypothesis, all we would need to show is that the Stalinesque hypothesis is true. In other words, in order to show that the Orwellian hypothesis is false, all one is required to do is to show that the discrimination of the color phi illusion occurs in the early layers of the visual stream. A compelling case for this proposal will be provided in detail, in the following section.

Perceptual Illusions in Connectionist Models

In recent years, connectionist learning models that utilize the backpropagation learning rule have gained acceptance and have been extensively used for modeling neural responses in the sensory cortex (Yamins and DiCarlo (2016)). There has also been a consistent effort in behaviorally aligning these vision-based deep learning models with human vision (Ahlert et al., 2024)(Wichmann and Geirhos, 2023)(Geirhos et al., 2021). Given the recent advancements as evidence for the success of deep learning models for modeling biological vision, it is a rather obvious question to ask, if these models can help us in building a better understanding of the color phi phenomenon.

Many recent advancements in the field of computer vision take inspiration from the theory of predictive coding. Predictive coding postulates that the brain functions through continuous generation and update of the mental model of the environment. The model minimizes the difference between the sensory signal predicted by it and the actual sensory signal received as input. Lotter et al. (2016) proposed a deep learning model for visual processing called PredNet inspired by a previously proposed hierarchical predictive coding model by Rao and Ballard (1999). The model has top-down and bottom-up connections with recurrence at each layer as seen in Figure 2, where the error between the incoming signal and predicted signal is minimized using backpropagation. The authors trained the model to perform a simple task of predicting the next frame given the current frame, in natural videos. Unbeknownst to the authors, as shown by Watanabe et al. (2018), the model predicted rotational motion in illusion images (shown in Figure 3) that were static, similar to visual perception in humans.

Keuninckx and Cleeremans (2021) show that the color phi phenomenon can be ex-

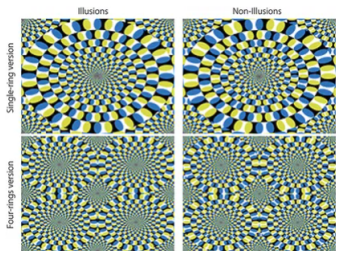


Figure 3: Rotating snake illusions (from Watanabe et al. (2018))

plained using recurrent connectionist networks built with dynamical toy model neurons. The model does not require training using backpropagation, instead, they perform linear regression using the output of the recurrent model to analytically compute the read-out layer that predicts the color and the position of the dot. The approach to explain the perceptual illusion provided in this work is different from that of Watanabe et al. (2018), as the solution is hand-engineered by providing a supervising signal to the model. Whereas in the previous approach, the model had never seen the illusion during the training phase. Instead, the motion, changes in the natural video, and the architecture inspired by predictive coding turn out to be a good enough prior to predict the rotational motion in illusion images. Nonetheless, these approaches go on to show that such simple illusions can be captured by rather simple statistical models when equipped with suitable priors.

These findings from connectionist models provide compelling evidence that the color phi illusion arises in the early stages of visual processing, without requiring higher-level cognitive explanations. The fact that simple recurrent neural networks can replicate the illusion, either through unsupervised learning on natural videos or with a hand-engineered solution, suggests that the apparent motion and color mixing are a result of the predictive and integrative dynamics occurring in early visual areas. This aligns with the Stalinesque hypothesis that the illusion is genuinely perceived rather than being a memory confabulation as posited by the Orwellian view. Of course, the neural networks used in these studies are simplified models, and more work is needed to conclusively map their architectures and behaviors onto the human visual system. Nevertheless, they demonstrate how perceptual illusions can naturally arise from basic mechanisms of prediction and inference in hierarchical recurrent sensory processing streams.

Now, if neural responses that correlated with the illusion, similar to the activations in the above models, were to be found in the visual cortex on performing the color phi experiment, one could conclude that the only hypothesis (from the cartesian materialist point of view) that explains the color phi phenomenon would be the Stalinesque hypothesis as the illusion occurs early in the brain before the conscious experience occurs. I suggest so, because I, like many others, assume the models that mimic the functionality required to elicit the relevant activations for the perceptual illusions mentioned above, do not undergo the conscious experience.

Conclusion

Given the above-provided argument, the color phi experiment is not the most favorable phenomenon to evaluate conscious experience as it does not require high-level cognitive functions. The findings from connectionist models suggest that the illusion can be explained by predictive and integrative processes occurring in early visual areas, without

the need for higher-level cognitive explanations. If neural responses correlating with the illusion were found in the visual cortex during the color phi experiment, it would provide evidence that the illusion occurs early in the brain, before conscious experience. This would challenge the conclusion made by Dennett that both, the Stalinesque and Orwellian hypotheses, account for all the data. While this might weaken Dennett's specific argument based on the color phi phenomenon, it may not necessarily invalidate his broader multiple drafts model.

References

Ahlert, J., Klein, T., Wichmann, F. A., and Geirhos, R. (2024). How aligned are different alignment metrics? In *ICLR 2024 Workshop on Representational Alignment*.

Dennett, D. C. (1993). *Consciousness explained*. Penguin uk.

Geirhos, R., Narayanappa, K., Mitzkus, B., Thieringer, T., Bethge, M., Wichmann, F. A., and Brendel, W. (2021). Partial success in closing the gap between human and machine vision. *Advances in Neural Information Processing Systems*, 34:23885–23899.

Keuninckx, L. and Cleeremans, A. (2021). The color phi phenomenon: Not so special, after all? *PLoS computational biology*, 17(9):e1009344.

Kolers, P. A. and von Grünau, M. (1976). Shape and color in apparent motion. *Vision research*, 16(4):329–335.

Lotter, W., Kreiman, G., and Cox, D. (2016). Deep predictive coding networks for video prediction and unsupervised learning. *arXiv preprint arXiv:1605.08104*.

Rao, R. P. and Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, 2(1):79–87.

Watanabe, E., Kitaoka, A., Sakamoto, K., Yasugi, M., and Tanaka, K. (2018). Illusory motion reproduced by deep neural networks trained for prediction. *Frontiers in psychology*, 9:340023.

Wertheimer, M. (1912). Experimentelle studien uber das sehen von bewegung. *Zeitschrift fur psychologie*, 61:161–165.

Wichmann, F. A. and Geirhos, R. (2023). Are deep neural networks adequate behavioral models of human visual perception? *Annual Review of Vision Science*, 9:501–524.

Wikipedia contributors (2023). Color phi phenomenon — Wikipedia, the free encyclopedia.

Yamins, D. L. and DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3):356–365.