Olivia Y. Lee
SYMSYS 205 – Final Paper

Predictive Processing: Efficiently processing high-dimensional, multimodal inputs

Humans regularly engage multiple sensory modalities when interacting with the environment, each of which are complex and involve high-dimensional information. Furthermore, given that many core human perceptual processes involve multiple modalities, an active field of research is dedicated to understanding how humans quickly and efficiently process multimodal sensory input, especially when each modality provides different resolutions of information and have unique influences on our perceptual experience. Several theories have been put forward about how raw, high-dimensional sensory inputs are processed to form higher level representations and enable us to act. One of such frameworks is the predictive processing theory (Clark, 2013), which posits that the brain is a hierarchical generative system that constantly models the environment according to Bayesian principles to predict incoming sensory information. This view emphasizes the tight integration between cognition and perception, since the generative system producing predictions is fundamentally influenced by cognitive factors. This paper argues for the plausibility of the predictive processing framework over the standard bottom-up model of perception, especially in the context of efficiently processing high-dimensional multimodal inputs, where the qualitative space of each modality has unique dimensionality and structure.

The outline of this paper is as follows: Section I begins with a comparative analysis of vision, sound, and smell, in terms of the dimensionality of their quality spaces and attentional selectivity. This sets the stage for understanding the influences of inputs from different perceptual modalities on our conception of the environment. Section II explores the perceptual processing pipeline and the influence of cognitive penetration, drawing connections between the discussion of multimodal perception in Section I with the hierarchical four-stage model of perception (Vetter & Newen, 2014). It also introduces the predictive processing framework, which contends that the brain constantly generates models of the environment according to Bayesian principles to predict incoming information. Section III then outlines arguments for the plausibility of the predictive processing framework and concludes with open questions about predictive processing and its relation to perception and cognition.

I. A Comparative Analysis of Vision, Sound, and Smell

Humans engage in multiple sensory modalities as we perceive and interact with the environment. However, there are remarkable differences in the influence of each modality on our perceptual experience, and some have even proposed a hierarchy of sense modalities. Locke (1979) asserts that vision is "the most comprehensive of all our senses", as it portrays a wide range of "Ideas of Light and Colors, which are peculiar only to that Sense; and also the far different Ideas of Space, Figure, and Motion, the several varieties whereof change the appearances of its proper

Object, viz. Light and Colours". The idea that visual perception can be represented by a set of well-defined properties, such as illumination and color (which themselves are similarly well-defined quality spaces), while that of sound and smell are comparatively more ambiguous, warrants efforts to develop a more concrete comparison framework to make sense of these differences, for instance analyzing the dimensionality of the quality spaces representing each modality. This paper will pay closer attention to vision, sound, and smell, though analogous arguments can be extended to taste and touch.

Beginning with vision, there is the sense that the visual space is complete, in that it seems possible to determine a basis from which a quality space with clear dimensions can be defined. An example of an attempt to do so is Siegel and Byrne (2017), which presents a set of thin properties of vision: color, illumination, shape, texture, size, spatial relations, and motion. Neuroimaging studies have identified specialized areas in the brain's visual system dedicated to recovering information about these properties early in the visual processing hierarchy. Therefore, it is plausible that the thin properties form a basis that other visible properties supervene on. This provides an explanation for the perception of rich properties (e.g., object recognition and emotion detection) through changes in thin properties. It has been debated whether the set of thin properties forms a minimal basis for the visual information space; for example, color and spatial relations could potentially give information on shape, illumination, and texture. While it may be possible that there is redundancy in the set of thin properties perceived by the visual system, the argument that the visual information space can be clearly defined with clear dimensions arising from a set of visual properties still stands. In addition, visual perception is selective and requires the agent to attend to certain elements of the visual field. For these reasons, visual perception is seen to be the primary sense that we rely on for high resolution information, seeing as with vision alone we are less likely to rely on other senses to confirm visual input. On the other hand, with sound or smell, we typically require either prior contextual information or visual information to reduce uncertainty about (i.e., confirm) our inferences from sensory input.

With this intuition, it suggests that the quality space of audition is less clearly defined than that of vision, and therefore less complete. While it is possible to dispute the "correctness" of color perception (Tye, 2006a), such similarity judgements are more difficult to do with sound, suggesting the quality space of sound is less definitive. However, it is still possible to identify a set of properties that characterize sound, such as pitch, loudness, and timbre. This naturally leads to the question of the relative ability of vision and audition to convey spatiotemporal information. In terms of space (i.e., distance) information between a subject and object, audition requires information to be collected over time for the subject to determine the scale of correlation between sound intensity emitted from the object and actual distance. The estimate of this scale of correlation may also be less accurate since sound can be distorted by external factors like obstacles or acoustics of the environment. In contrast, vision allows for fast, high-resolution determination of distance provided the target object is within the subject's field of view. In terms of time (i.e., duration)

information, both vision and audition can convey information about whether an event has occurred or the duration of an event. However, the interesting phenomena observed in the double flash illusion (Shams et al., 2002) suggests that audition may be more informative in conveying temporal information. In the cross-modal double flash illusion, one dot flashes on the screen while one and two beeps are sounded, however when there are two beeps, people frequently report experiencing two flashes. This suggests that the brain is wired to premise duration information from audition over that from vision. The fact that sensory audio inputs can overwrite visual inputs if they conflict suggests that we may rely more on audition for temporal information.

The difference in primary modality for distance or space perception versus time or duration perception is that the former is relative to the perceiver (e.g., the perceivers proximity to an object), whereas the latter is absolute from the perceiver's perspective (e.g., whether or how many times an instantaneous event like a flash occurred). Based on O'Callaghan's (2007) theory of sounds, sounds are *events* located in the environment near their sources and are consequently temporal in nature. The sound waves produced by such events are the objects of auditory perception. Therefore, while visual perception allows subjects to experience ordinary objects directly (by ascribing physical properties to objects seen in visual experiences), auditory perception removes the subject from directly perceiving the source. This is evident in the McGurk effect (McGurk & MacDonald, 1976), where subjects rely on high-resolution visual information to confirm low-resolution or ambiguous auditory information. This suggests that audition conveys crucial elements of spatiotemporal information but conveys lower resolution information than vision, seeing as we usually turn our attention (i.e., our visual field) to detect and reduce uncertainty about the source of the sound. While audition is a narrower information channel, it is less selective than vision in that we can hear sounds without intentionally attending to the object emitting the sound and can eventually turn our attention to the source by directing our field of vision towards it.

Like sound perception, olfactory experiences similarly remove the subject from directly experiencing ordinary objects. The invisible gaseous emissions from their sources are the primary objects of olfactory perception, and we secondarily experience the ordinary objects emitting the odors. However, while sounds convey spatiotemporal information and can unfold over time, Batty (2009) asserts that smells are typically immediate and static over time, hence there is no localization in olfactory experience. This is an extreme stance, though it does seem true that the information conveyed by olfactory perception is of a much lower resolution than audition or vision. Vision allows for direct experience of ordinary objects, and because sounds are located near their sources, we can estimate a proxy for distance based on sound intensity. However, smells can freely diffuse and permeate through spaces, and do not interact with objects in the environments based on the laws of physics like sounds do (through reflection, diffraction etc.). Therefore, olfactory perceptual inputs are usually combined with contextual knowledge to derive spatiotemporal information. For example, the knowledge that something smells good in a kitchen nearby allows us to estimate distance from the kitchen (spatial information) or whether food is ready (temporal

information). Familiarity with food smells from prior experience also play a significant role in conveying spatiotemporal information. However, if one were to detect an arbitrary, unfamiliar smell without prior knowledge of what or where the ordinary object is, it would be difficult for the isolated olfactory experience to convey high-resolution spatiotemporal information. A cognitive or contextual leap is needed for us to derive information from olfactory perception, though with familiarity or training through prior experience, this cognitive leap becomes so automatic that it feels perceptual. In a similar way to estimating distances with sound intensity, if one were familiar with a smell and knows *a priori* how intensity relates to distance, deriving spatiotemporal information from the olfactory experience it can seem perceptual. However, as with sound, external environmental factors like temperature and humidity can affect the accuracy of the correlation estimate.

Compared to vision and sound, the dimensionality and completeness of the olfactory perceptual space is less well-defined. Smells seem highly categorical rather than continuous and are very tightly associated with ordinary objects rather than the qualitative properties of the smell, like "floral" (with the ordinary object being flower) or "minty" (with the ordinary object being mint). For instance, one can easily describe an object-agnostic sound (e.g., the sound resulting from a sinusoidal waveform) or an arbitrary object based on its visual properties (e.g., a red sphere), however it is difficult to name an object-agnostic smell other than with adjectives like "sweet", and the ambiguity of such adjectives (which are not quantifiable properties) make it difficult to constitute a properly defined, exhaustive space. Specifically, in the well-defined space of color vision, a sense of completeness arises because whatever one's visual system reports about color will always be a point located in color space. Even in the problem from Tye (2006b) where Jane underrepresents yellowishness in detecting orange, she still claims to see red, which is an identifiable point in color space. The clear dimensionality of color space formalizes a structure allowing for similarity and difference comparisons, which is how debates about the puzzle of True Blue (Tye, 2006a) can arise in the first place. On the other hand, similarity and difference relationships for smells seem much less definitive. Given the tight association between smells and ordinary objects, people often describe smells based on whether certain scents are present or not (as a commonplace example, consider the descriptions people give for chocolate or wine tasting). The qualitative space for smells therefore seems less structured and more difficult to articulate objectively, especially with the tight connections between smell and emotion or experience.

Overall, from the comparative analysis between vision, audition, and olfactory perception, visual perception seems to provide high-resolution information, given the clear dimensionality and completeness of its qualitative space. However, because vision allows an agent to directly experience ordinary objects, it is selective and requires one the attend to certain elements within one's field of vision. Other modalities like sound and smell seem to be less informative, at least for most relevant spatiotemporal judgements. Relative to vision, the qualitative spaces of sound and smell are less formalized, especially that of olfactory perception given its highly categorical

nature and strong associations with ordinary objects rather than qualitative properties. Because they allow the subject to secondarily experience ordinary objects, through the sound waves or gaseous emissions respectively produced by the object, they are non-selective and we can choose to turn our attention to the objects secondarily experienced (e.g., by moving our field of vision towards the source). Tactile or touch perception is another important modality used by humans that was not analyzed here, though similarities can be drawn with sound perception in that there are certain properties that could conceivably define a broad qualitative space (e.g., pressure, warmth etc.). We are more likely to rely on our visual perception for spatiotemporal judgements without requiring confirmatory inputs from other modalities, whereas for modalities like sound or smell, we are likely to use vision to further reduce uncertainty about the environment. These comparisons demonstrate that different modalities have unique influences on how we perceive our environment. Furthermore, sensory information from these modalities is often combined with amodal contextual information, blurring the lines between direct perception and the influence of cognition. With this in mind, it is worthwhile to consider the interplay between perception and cognition in the perceptual processing pipeline.

II. Predictive Processing: The Influence of Cognitive Penetration on Perceptual Processing

A strong body of evidence points to the fact that higher-level cognitive states affect the way we perceive the world, from Russian speakers being able to differentiate between more shades of blue than English speakers to cultural differences between Western and Eastern subjects affecting line estimation. In addition, the McGurk effect and the double flash illusion both demonstrate lateral connections between perceptual modules. The diverging responses from subjects of these intermodal illusion tests demonstrate that cognition influences visual and auditory inputs, especially when ambiguous stimuli and conflicts between sensory inputs must be resolved, introducing biases through beliefs, cultural attitudes, experience, and a myriad of other factors. Considering that the variation in receptor sensitivity is within a reasonable range for normal human perceivers, any perceptual variation that arises can be attributed to cognitive processes. Hence, a proposed pipeline for standard perceptual processing, and the resulting perceptual variation that arises, is as follows: an objective sensory input (color, sound, smell etc.) stimulates sensory receptors, where there is a slight variation though within a reasonable range for normal human perceivers. The most significant variation is introduced in cognitive processing, due to influences such as personal experience, familiarity, affective variation, and other factors, resulting in individuals forming different representations from the same objective sensory input.

Several frameworks have been proposed to account for cognitive penetration embedded into perceptual processing. One of such frameworks, put forward by Vetter and Newen (2014), is a hierarchical four-stage model focusing on visual perception, where cognitive penetration arises from both bottom-up and top-down connections between perceptual hierarchies. A summary of the framework is as shown in the figure below:

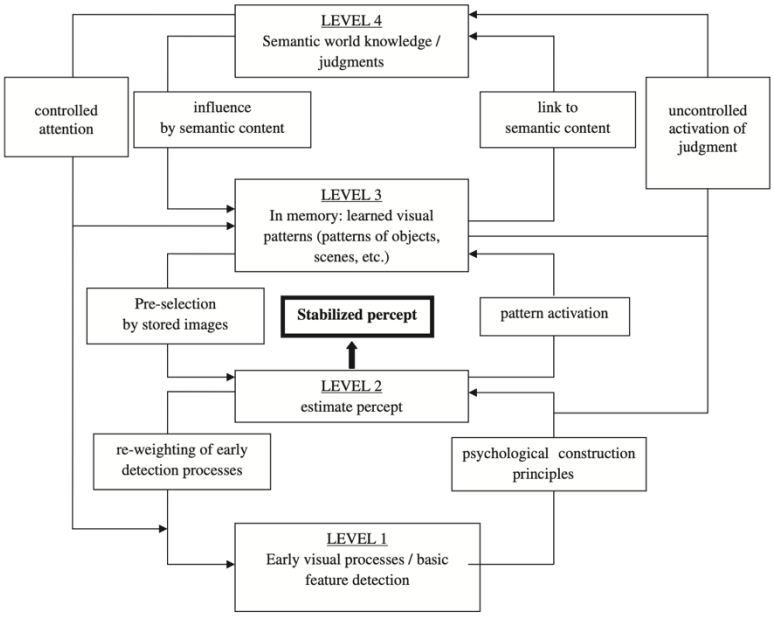

Fig. 1. Schematic diagram of a four-stage hierarchy of perceptual processing (Vetter & Newen, 2014)

In the standard model of perception without any influence from cognitive penetration, sensory input is processed through a purely feedforward procedure. This model emphasizes the influence of cognitive penetration through bidirectional feedback connections between all pairwise combinations of hierarchical levels that mutually influence one another. It also suggests that it is difficult to compartmentalize or separate cognition from perception due to the presence of both upward and downward connections.

Predictive processing (also referred to as predictive coding) is a closely related theory that flips the order of explanation provided by the four-stage model on its head, emphasizing the top-down connections that carry predictions about how sensory information is likely composed. The core idea of the predictive processing theory is that the brain is a prediction machine, supporting perception and action by constantly attempting to minimize prediction error between incoming sensory inputs and top-down predictions created by a hierarchical generative model (Clark, 2013). The predictive processing framework and the four-stage model both rely on bidirectional cortical processing, however, predictive processing suggests that the top-down processes in the four-stage model precede the bottom-up processes. According to predictive processing theory, the brain constantly generates models about the environment based on Bayesian principles to predict incoming sensory inputs, selecting models that corroborate with observations from perception. Relating predictive processing to the hierarchical levels in the four-stage model, the brain builds predictions about the environment based on contextual knowledge about a scenario, corroborates this with experiences and learned associations stored in memory, creates an estimate percept, and verifies this estimation against the basic features detected. If there is an unexpected mismatch between the prediction and incoming sensory input in this top-down process, the brain's predictive model is updated through the bottom-up connections, until this process stabilizes.

The standard perceptual processing pipeline takes stimulation to one's sense organs as input, which is processed by the brain to generate a representation of the environment given that particular stimulation. Predictive processing reverses the order of explanation, postulating that the brain constructs a representation of the environment, and checks that the perceptual input received corresponds with those predictions. A concrete example of this theory applied to visual processing is as follows: consider a normal human perceiver sitting in a chair, looking out in front of them, and begins to stand up. Their brain considers the contextual information of sending signals to the relevant muscles to stand up, and hence predicts the scene in front of them will move downwards as they stand up. The brain then checks its prediction to match the stimuli to their visual system. An inconsistency between the prediction and incoming stimuli triggers a Bayesian update where the brain attempts to generate a new scenario consistent with the stimuli being received.

In relation to the four-stage model, the top-down and bottom-up processes between Level 1 and Level 2 are especially pertinent to the predictive processing theory. The top-down process describes the effect of the estimate percept generated in Level 2 on one's basic feature detection in Level 1. Predictive processing contends that the estimate percept is verified against the actual percept, and a mismatch beyond a certain threshold would then trigger a backpropagation through the bottom-up process, updating the internal model that generates estimate percepts in Level 2 that reduces the prediction error between the predicted and actual percepts. In the visual cortex, backward connections from V2 to V1 carry a prediction of expected activity in V1, while forward connections from V1 to V2 carry the error signal indicating unpredicted or incorrectly predicted activity. This feedback loop repeats to minimize the error signal carried in the forward connections and occurs very quickly; as soon as the earlier stages of the visual system are starting to process stimuli to the sensory organs, the brain attempts to corroborate this input to form a coherent picture. Feedback from this predictive processing paradigm has been found to cause perceived illusory contours (Pang et al., 2021), for instance with Kanisza square and triangle illusions where the subjects' retinotopic edge detectors fired as through a square or triangle was present. Such illusions hence suggest the brain is always predicting future states of the environment. Instead of processing all the information at the next timestep from scratch, it verifies pre-formed or predicted representations against incoming sensory information, attempting to construct a coherent interpretation between predicted and actual percepts.

The predictive processing model thus presents perception as an active, generative process, and by Bayesian principles, the brain always has predictions on how sensory information should be composed. It is an extreme form of cognitive penetration, seeing as the brain's predictive model (which is fundamental for generating the estimated percept) is significantly influenced by factors like experience, affect, and familiarity. Having introduced the predictive processing framework in relation to Vetter and Newen's (2014) four-stage model, we shall turn to arguments for the plausibility of predictive processing as a framework for understanding perception, as well as open questions about the predictive processing theory and cognitive penetration more broadly.

III. Plausibility Arguments and Open Questions

It can be argued that predictive processing is more plausible than the standard model of perception on the grounds that it is computationally intensive to evaluate sensory information from scratch at every timestep. Based on our analysis of vision, audition, and olfactory perception in Section I, sensory information from one modality is already complex and high-dimensional, let alone aggregated multimodal information. It is not plausible that we evaluate our perceptual information at every timestep with no representations *a priori*, since there is essentially an infinite number of potential states the environment can adopt. Priors from contextual knowledge, past experiences, and learned associations stored in memory are thus essential in constraining the state space of the environment. This allows normal human perceivers to be capable of processing multiple high-dimensional inputs in short time frames, sometimes almost instantaneously in the case of reflex actions, to interact with the environment.

The primary argument for the plausibility of predictive processing is therefore in the significant reduction in processing power required to interpret the vast amount of high-dimensional sensory information constantly presented to one's sensory organs and brain. Generating predicted percepts based on experience is hence an efficient way of allocating processing resources to novel information (where the prediction error is high) or important information demanding our attention, whereas situations with low prediction error can safely rely on the brain's predictive model instead of re-processing those sensory inputs from scratch. As a concrete example, consider a human driving a car down a street. An unpredicted and suddenly appearing stimulus, like a deer jumping into the car's path, creates a high error signal with the predicted stimulus (i.e., the expected movement of the road and objects in the visual field as the car moves forward). This guides the driver's attention towards the novel stimulus (the deer) while other stimuli (e.g., road, pavement, trees, etc.) are dedicated less resources for processing.

This model of perception has particularly interesting implications for the distinction between perception and cognition. It suggests that we begin to effortfully think about what our basic sensations entail, but over time, the brain's generative models based on Bayesian principles are continually updated and eventually achieve low prediction error on certain sensory inputs. Those states consequently have high likelihood and high prior probability, thus requiring less resources be devoted to processing them. As the generative model improves in accuracy and stabilizes, the bottom-up connections are engaged much less frequently, thus we become better and faster at processing these inputs, so much so that it seemingly becomes perceptual.

Another argument for the plausibility of predictive processing is that certain reflexive human behaviors suggest that humans do not process information in this bottom-up fashion from basic feature detection to semantic world knowledge. In the above example of a deer jumping in front of a car, the driver instinctively slams on the brakes. Often, the driver does so before they process that the animal is a deer, that it is brown in color, and other semantic details beyond the

fact that an unexpected stimulus was presented to their visual receptors. This supports the plausibility of predictive processing, because in the standard perception model, sensory inputs must be processed through the hierarchy before an action is taken. In contrast, with the predictive processing model, a conflict between predicted and actual percept is sufficient to trigger an action, and the driver need not process all the fine-grained details about the sensory input at the time of action. Therefore, because certain associations that trigger reflex actions cannot be processed quickly enough in the standard model of perception, such phenomena lend support to the predictive processing theory where actions arise from conflicts between predicted and actual percepts.

While the above arguments suggests that predictive processing offers a strong proposal for a unified science of mind and action, some open questions remain about the approach. One question pertains to the initialization of the brain's hierarchical generative model; it is not entirely clear how initial predictions are formed, and we have yet to determine how human babies begin forming, verifying, and correcting predicted percepts. One possibility is that models of the environment are generated with uniform probabilistic models or random noise, and eventually this noise is distilled down into a coherent structure after several updates. This process could be analogous to that involved in seeing a Dalmation dog in a picture of seemingly random noise after one's attention is guided towards the target, where the predicted percept is the original picture which is updated after receiving the actual percept of a Dalmation dog. Furthermore, as we observe with artificial neural networks, randomly initialized networks can converge on very different stabilized models, even when given the same inputs. This may have interesting implications on explaining perceptual variations among humans. In addition, it is unclear whether the architecture for verifying predicted percepts against actual percepts is built into our perceptual systems or developed through experience, and if the latter, how this architecture is developed and results in the predictive model stably converging, given the high degree of freedom when both verification and prediction architectures are allowed to freely vary.

Another open question about predictive processing involves conflict resolution, both within a single modality and across modalities. In predictive processing, the brain forms a probabilistic representation of the environment, and compares this prediction with stimuli to one's perceptual organs, only updating the generative model if there is a conflict between predicted and actual percepts. First, we consider conflict resolution within a modality. Such conflicts are typically much quicker to resolve, as with the example of a deer darting out in front of a driver's car. They have been argued to be more automatically resolved, as demonstrated by the phenomenon of binocular rivalry which occurs when each eye is presented with a different image and subjective perception alternates between them, whereas similar images are overlaid on one another. Hohwy (2008) hypothesizes that the alternation between stimuli in rivalry occurs when (i) there is no single model or hypothesis about the causes in the environment that has both high likelihood and high prior probability and (ii) when one stimulus dominates, the bottom-up signal for that stimulus is explained away while that of the suppressed stimulus is not and remains an

unexplained but explainable prediction error signal. This induced instability results in perceptual transitions or alternations during binocular rivalry. In other words, if there is a low degree of conflict between the two images, the predictive processing framework merges the images in the brain's attempt to construct a coherent model of the environment, however a sufficient degree of conflict between stimuli is not sustained within the visual system. Yet, signals for both images are being processed in the visual cortex, and it is not the case that the signals are processed in an alternating fashion (like two interfering sound waves). The brain still forms representations from both stimuli, but it cannot make a unified percept out of both. Based on Hohwy's (2008) hypotheses, it is interesting to consider the latter case where one stimulus dominates. It remains an open question how these findings apply to experiments demonstrating the dominance of faces over objects (Persike et al., 2014), as well as cross-modal influences on binocular rivalry, for instance the increased dominance of faces in visual consciousness when paired with negative gossip (Anderson et al., 2011).

Conflict resolution between modalities is more complex and often less automatically resolved. Conflicting sensory information results in the brain maintaining multiple different models of the environment, which can be costly. If all humans do is take information in and process it, as in the standard model of perception, there is no *prima facie* problem with receiving conflicting cross-modal information, other than potentially developing erroneous associations between multimodal information. With the predictive processing framework, a coherent prediction is made but it is partially confirmed and partially violated. As a concrete example, consider the common phenomenon of motion sickness when using virtual reality (VR) headsets. Motion sickness is often reported when VR users are immersed in virtual reality environments displaying visuals that do not corroborate with the individual's actions or movements, for instance visuals of riding a rollercoaster while the individual is mostly physically stationary. In this case, there is significant conflict between the individual's visual and proprioception experiences which is not resolved in the way binocular rivalry within a single modality is resolved (e.g., "alternating" stimuli). There is a mismatch between current simultaneous visual and proprioception experiences and previously learned associations between visual and proprioception experiences. However, though such cross-modal conflict lasts longer and is not automatically resolved, there is an obvious cost in withstanding this conflict, which is the discomfort when using VR. Therefore, it is unclear how the brain attempts to build a coherent model of the environment from this information, and how the predictive model updates when the brain receives conflicting information. This prompts further inquiry into properties that can be represented cross-modally and properties that can only be represented in a single modality.

In conclusion, the predictive processing theory offers a plausible framework for the efficient processing of complex, high-dimensional multimodal sensory input. The remarkable ability of humans to process a wealth of sensory information in a short span of time lends credence

to the hypothesis that the brain is a predictive engine constantly modeling the environment and verifying its predictions with incoming sensory input. Many core human perceptual processes involve multiple modalities, where each modality has different structural dimensionality, provides different resolutions of information, and uniquely influences our broader perceptual experience. Predictive processing is a unified framework through which we can begin to make sense of how humans quickly and efficiently process multimodal sensory inputs. Ultimately, predictive processing demonstrates that perception and cognition are inherently interconnected and mutually influence each other through bidirectional feedback loops, which provides important insights into potential sources of perceptual variation among humans.

Bibliography

Anderson, E., Siegel, E. H., Bliss-Moreau, E., & Barrett, L. F. (2011). The visual impact of gossip. *Science (New York, N.Y.)* 332(6036):1446–1448. doi: 10.1126/science.1201574.

Batty, C. (2009). What's That Smell? *Southern Journal of Philosophy* 47(4):321-348.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences* 36:181-204.

Hohwy, J., Roepstorff, A., & Friston, K. (2008). Predictive coding explains binocular rivalry: an epistemological review. *Cognition* 108(3):687–701. doi: 10.1016/j.cognition.2008.05.010.

Locke, J., & Nidditch, P. H. (1979). An essay concerning human understanding. Oxford: Clarendon Press.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature* 264:746–748. doi: 10.1038/264746a0.

O'Callaghan, C. (2007). *Sounds: A Philosophical Theory*. Oxford: Oxford University Press.

Pang, Z., O'May, C.B., Choksi, B., & Rullen, R.V. (2021). Predictive coding feedback results in perceived illusory contours in a recurrent neural network. *Neural networks: the official journal of the International Neural Network Society* 144:164-175.

Persike, M., Meinhardt-Injac, B., & Meinhardt, G. (2014). The face inversion effect in opponent-stimulus rivalry. *Frontiers in human neuroscience* 8:295. doi: 10.3389/fnhum.2014.00295.

Siegel, S., & Byrne, A. (2017). Rich or thin? *Current Controversies in Philosophy of Perception*. New York, USA: Routledge. pp. 59-80.

Shams L., Kamitani Y., & Shimojo S. (2000). What you see is what you hear. *Nature* 408:788–788. doi: 10.1038/35048669.

Tye, M. (2006a). The puzzle of true blue. *Analysis* 66: 173-78.

Tye, M. (2006b). The truth about true blue. *Analysis* 66: 340-44.

Vetter, P., & Newen, A. (2014). Varieties of cognitive penetration in visual perception. *Consciousness and Cognition* 27:62-75.