5 Predictive Processing and Object Recognition

Berit Brogaard and Thomas Alrik Sørensen

5.1 Introduction

There has been a lot of recent interest in predictive processing (PP) theories of cognition (Bar, 2003; Clark, 2013, 2016, 2020; Friston, 2003, 2009, 2010; Hohwy, 2012, 2013, 2020). While initial PP models focused primarily on visual perception (Friston, 2003, 2005), recent advocates have suggested that the predictive framework can account for all mental processes and indeed all of the brain's operations (Clark, 2013, 2016, 2020; Hohwy, 2013, 2020).

PP theories take issue with classic models of visual perception. On the classical approach, lower cortical areas (V1, V2, V4, etc) in the ventral stream process sensory information that has been filtered through the thalamus (LGN) and then project this information to higher regions [e.g., the inferior temporal (IT) cortex]. Here, the information is further processed in light of feedback from object templates stored in long-term memory. Once the last visual area in the ventral stream has processed the information from earlier areas, a visual perception of the distal object is generated. The traditional model thus focuses primarily on bottom-up processing and to a lesser extent on top-down modulation. On the predictive view, this picture is reversed (Clark, 2013, 2015; Feldman & Friston, 2010; Hohwy, 2012, 2013). The predictive framework posits that the brain deploys internal models, which contain information extracted from past experience, to generate predictions, or hypotheses, about its surroundings. These predictions are then matched to the incoming visual information. Mismatches between predictions and incoming signals – so-called prediction errors – are then projected bottom-up to higher brain areas, where they are used to update the predictions. This process, which occurs hierarchically, continues until the prediction errors are minimized to the greatest extent possible, and the winning prediction determines the visual content. In contrast to traditional models, predictive approaches thus hold that all bottom-up processes are signals conveying prediction errors to higher regions. By only

DOI: 10.4324/9781003084082-7

processing prediction errors rather than all visual stimuli, the brain saves energy (Friston, 2003, 2004, 2009).

Here, we argue that the predictive approach falls short of providing a complete account of visual perception. Specifically, we take issue with the predictive approach's core idea that all bottom-up signals are prediction error signals and that prediction error minimization is "all that the brain ever does," as Jakob Hohwy puts it (2013, p. 7). Although our point is a general one, we will focus on the case of object recognition. As we will see, there is a substantial body of evidence suggesting that there are three stages to object recognition: (i) scene gist processing, (ii) attentional object selection, and (iii) hypothesis testing. We argue that PP theories lack the resources to accommodate the first two stages of object recognition. Ransom, Fazelpour, and Mole (2017) and Ransom et al. (2020) have previously argued that the predictive account of attention is unable to account for voluntary object attention and affect-biased attention. These conclusions challenge one of the predictive account's key claims, viz. that it offers a unified theory of the mind. We will argue that attention during the earliest stages of object recognition presents a further problem for the predictive account of attention.

The chapter is structured as follows. In Section 5.2, we outline the details of the predictive approach, as presented by Karl Friston, Andy Clark, Jakob Hohwy, and others. In the two subsequent sections, we review the empirical evidence for the claim that object recognition begins with gist processing and argue that the PP framework is unable to accommodate gist processing. In Section 5.5, we offer a brief overview of the previous studies showing that predictive models of attention are unable to account for selective attention and affect-biased attention and then argue that attention at the earliest stage of object processing presents a problem for the predictive approach. Finally, in the concluding section, we discuss some ways in which the predictive account may be augmented to provide a unified theory of the mind.

5.2 The Brain as a Hypothesis-Testing Mechanism

The PP approach is often cast as a solution to the problem of how the brain determines the distal cause of an incoming visual signal. This problem arises because any visual input has an infinite number of possible distal causes, which raises the question of how the brain reliably determines which is most probable. Consider this analogy from Hohwy:

> You are like the brain, the house is the skull, and the sound is auditory sensory input. As you are wondering about the cause of the input, you begin to list the possible causes of the input. It could be a woodpecker pecking at the wall, a branch tapping at the wall in

> the wind, a burglar tampering with a lock, heavy roadworks further down the street, a neighbour's loud music, or those kids throwing stones; or it could be something internal such as loose water pipes banging against each other. Let your imagination rip: it could be that your house has been launched into space over night and the sound is produced by a shower of meteorites. There is no end to the possible causes. Call each of these possibilities a *hypothesis*. The problem of perception is how the right hypothesis about the world is shaped and selected. (2013, pp. 15–16)

In Hohwy's analogy, you wonder about the cause of the auditory input and begin to list possible causes of the input. It might be caused by a woodpecker pecking at the wall, a branch tapping at the wall in the wind, a burglar tampering with a lock, heavy roadworks further down the street, a neighbor's loud music, those kids throwing stones, loose water pipes banging against each other, a shower of meteorites, and so on *ad infinitum*. Of course, the brain could not possibly test infinitely many hypotheses. So, it needs to somehow narrow down the infinite set to a more manageable one. PP's popularity is partly due to its advertisement as a solution to this problem (Hohwy, 2013). PP holds that the brain generates predictions, or hypotheses, and then uses Bayes' principle to determine the most probable hypothesis.[1] The competing, or alternative, hypotheses are generated by models that group together patterns, or statistical regularities, derived from past sensory inputs (Friston, 2009). One key concept in Bayes' principle is *likelihood*: how probable it is that the hypothesis accurately predicts the distal cause of the sensory input. The more probable it is that the hypothesis accurately predicts the distal cause of the sensory input, the greater its likelihood. Since mosquitos do not make pecking sounds, the likelihood that the pecking sound is caused by a mosquito buzzing around the ceiling lamp is low. So, this hypothesis' likelihood is low. But as far as you know, there are countless other hypotheses with a high likelihood. A second concept in Bayes' principle is a hypothesis' independent, or *prior, probability*. According PP, the prior is also determined by information about the environment, extracted from past experience. In our example, the prior probability of the hypothesis that the pecking sound is produced by a shower of meteorites is infinitesimal. But if there are a lot of woodpeckers, burglars, and stone-throwing kids in the area, then the prior probabilities of the woodpecker, burglar, and stone-throwing kids hypotheses are high. In the Bayesian framework, the hypothesis with the greatest *posterior probability* determines what you perceive. According to Bayes' principle, the posterior probability is the product of a hypothesis' prior and the likelihood that the hypothesis accurately predicts a distal cause of the sensory signal. If the woodpecker hypothesis has the highest

posterior probability, owing to its higher prior or its higher likelihood (or both), then you perceive the auditory signal as the sound of a woodpecker.

Of course, Hohwy's analogy is just that: an analogy. Here are three key differences between this analogy and the predictive account of perception. First, in the brain, the Bayesian inferences that go into determining the hypothesis with the highest posterior probability occur at the subpersonal level; these inferences are unconscious (at least for the case of perception).[2] So, you do not first hear a pecking sound and then the sound of a woodpecker. You just hear the sound of a woodpecker. Second, in the brain, there is a hierarchy of generative models that produce hypotheses (or predictions). In the brain, these hypotheses are more akin to "This neural activation in the V4/V8 color region is caused by a red object" than "This auditory signal is caused by a woodpecker." Third, at every level in the hierarchy, hypotheses generated at one level are matched to inputs at the level below. If there is a mismatch, or prediction error, between the hypothesis and the input, this prediction error is used to update the hypothesis. Updating a hypothesis effectively means that the hypothesis is revised in light of the information that did not accurately depict the distal cause of the incoming sensory signal. This process then continues until the brain has arrived at the hypothesis with the highest posterior probability. Determining the hypothesis with the highest posterior probability at each level amounts to minimizing the prediction error between the hypothesis and the sensory input. A perception arises as the prediction error is sufficiently minimized. So, prediction error minimization is a key concept in PP (Friston, 2010; Hohwy, 2013). One complication within the PP framework, which we will turn to below, is that prediction error minimization is subject to expectations of noise.

One of PP's boldest conjectures is that *only* prediction error signals, that is, signals that encode information about the prediction error, are propagated up through the system in a bottom-up fashion. In the following sections, we take issue with this claim. We argue that empirical studies of visual object recognition run counter to this conjecture that all bottom-up processing is prediction error signaling. We begin by looking closer at visual object recognition.

5.3 Object Recognition in Natural Visual Scenes

In ordinary life, objects tend to occur as parts of larger scenes, together with other items that are likely to occur in the same scenes (Trapp & Bar, 2015). While it is often difficult to find an object hidden in a crowded scene, context can facilitate the visual recognition of objects that are congruent with it (e.g., a frying pan in a kitchen) (Fiser & Aslin, 2001, 2005; Kondo, van Loon, Kawahara, & Moore, 2017; Oliva & Torralba, 2007). However,

scene context presents an obstacle to the discrimination of objects that are incongruent with it (e.g., a frying pan in a movie theater) (Auckland et al., 2007; Gordon, 2004; Hollingworth & Henderson, 1998; Oliva & Torralba, 2007; Palmer, 1975). Thus, when viewing a kitchen containing a frying pan and a bicycle helmet, the pan is detected faster and with greater ease than the helmet. However, if an object is located in an unusual place (e.g., a microwave hanging from the ceiling), detecting the object is slower in a congruent scene (e.g., kitchen) than an incongruent scene (e.g., living room) (Bar, 2004; Hoffman, 1996; Meyers & Rhoades, 1978). These effects are also known as "scene consistency-inconsistency effects" (Oliva & Torralba, 2007). Various other contextual factors besides object-scene relationships provide cues that can be exploited by the visual system for the identification of objects, including co-variation of objects, spatial and temporal proximity of objects, spatial configuration of objects relative to each other, typical positions of objects in scenes, familiar relative size of object, and pose of objects in scenes (Bar, 2004; Biederman, Mezzanotte, & Rabinowitz, 1982; Green & Hummel, 2006; Hock et al., 1974; Oliva & Torralba, 2007). For example, chairs and tables are expected to co-occur, whereas a frying pan and an elephant are not; fire hydrants are expected be on top of the sidewalk rather floating in the air; dinner plates are expected to be on top of tables, in stacks on shelves or in the sink or dishwasher but not on the floor; chairs are expected to be oriented toward tables rather than away from them; cars are expected to be oriented along the driving directions of a street rather than in the direction of the sky; and pedestrians are expected to be in an upright position rather than lying down.

The effects of scene context can be so strong that altering the background scene while leaving the target object intact can change the perceived identity of the object. In Figure 5.1, for example, the orange Toyota Supra is recognized as a real car in the nature scene but as a toy car in the street scene and the indoor scene. Biederman et al.'s (1982) prediction that relative familiar size (i.e., the scale of an object relative to other objects) influences object recognition is borne out here. In the street scene, relative size trumps statistical co-variation of objects, whereas both relative size and statistical co-variation of objects contribute to the identification of the car in the indoor scene.

People are sometimes capable of recognizing objects embedded in congruent scenes, even when they completely lack perceptible structures or features that can guide object recognition independently of scene context. In the street scene in Figure 5.2, for example, the blob on the right is identical to the blob on the left after 90 degrees rotation (Oliva & Torralba, 2007). So, the blob's intrinsic features do not reveal its identity. Nevertheless, the scene is immediately recognized as a street scene. Recognition of the scene gist activates a scene template (i.e., context frame) that provides

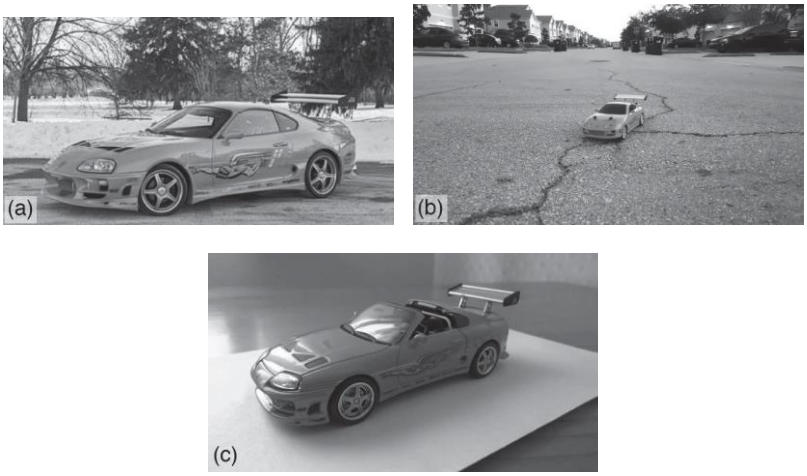

Figure 5.1 The car is immediately recognized as a real car in the nature scene (a) but is recognized as a toy car in both the street scene (b) and the indoor scene (c).

Figure 5.2 The gist of a street scene. The gist of a scene is the scene's low spatial frequency information, such as the global scene configuration and the gross contour of objects. In this image, the "pedestrian" on the right is identical to the "car" on the left after 90 degrees rotation. As cars and pedestrians typically are oriented differently in a street scene, observers recognize one blob as a car and the other as a pedestrian

Source: From Oliva and Torralba (2007)

information about the typical differences in the orientation of cars and pedestrians in a street scene. As a result, the blob on the right is recognized as a pedestrian and the blob on the left as a car.

In dynamic scenes, scene recognition can also facilitate object tracking and trajectory prediction and expectations. In a street scene where a moving bus passes a grocery store on the opposite side of the street and thereby occludes the storefront relative to the vantage point of an observer, she still expects the store to be present once the bus has passed it. However, expectations regarding the trajectory of a pedestrian occluded by the bus are not nearly as strong, as the pedestrian might have gone into the store in the meantime. While scene context facilitates and sometimes is essential to object recognition, people usually recognize a good exemplar (or prototype member) of an object category when presented to us without any scene context in a laboratory setting within 75 ms (see Dall, Wang, Cai, Chan, & Sørensen, 2021; Shibuya & Bundesen, 1988; Sørensen, Vangkilde, & Bundesen, 2015). Even under optimal conditions, however, recognizing an object without scene context is considerably slower than recognizing a familiar scene (36 ms) (Larson, Freeman, Ringer, & Loschky, 2014) (Figure 5.3).

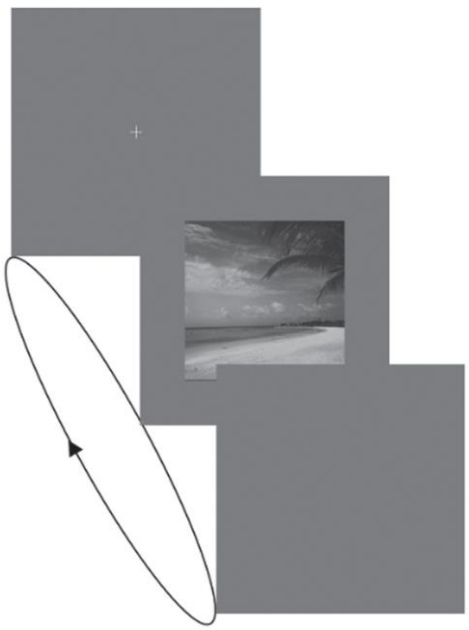

Figure 5.3 Demonstration of scene gist recognition. From KSU, Vision Cognition Laboratory (left, online only).[3] Scene information presented briefly between two blank screens can be extracted rapidly (right)

Studies have shown that the ability to rapidly recognize objects and scenes is due to the collaboration of two distinct visual pathways: a fast pathway that projects the gist of the object directly from the primary visual cortex (V1) to the orbitofrontal cortex (OFC) in the prefrontal cortex, which then generates predictions, or hypotheses, and a slower pathway that processes detailed information in a standard button-up fashion (V1, V4, V5/MT, LOC, IT) (Bar et al., 2006; Torralba, Oliva, Castelhano, & Henderson, 2006). The gist of an object or scene takes the form of low spatial frequency (LSF) information extracted from the sensory signal originating in the object or scene. LSF information encodes gross outlines and object contours, whereas high spatial frequency (HSF) information encodes sharp edges and fine details.

To demonstrate that the recognition of isolated objects takes place via dual visual pathways, Bar et al. (2006) combined functional magnetic resonance imaging (fMRI), magnetoencephalography (MEG), and a behavioral task. In the fMRI study, participants were shown images that were either unmanipulated or manipulated to contain only LSFs or HSFs. The results showed that the LSF image of an object elicited greater activity in OFC than the HSF image of that object, although unmanipulated images resulted in the greatest increase in OFC activity. Greater OFC activation was also observed when an object's LSF image had multiple interpretations compared to just a few, suggesting that the more ambiguous an object's LSF signal is, the greater the workload for OFC (see also Dall et al., 2021). In their 2007 study, Kveraga, Boshyan & Bar found that the fast pathway is a magnocellular pathway, which projects information from V1 to OFC via either subcortical projections or the dorsal "visual for action" pathway. All the brain's magnocellular (M) pathways project information much faster than its parvocellular (P) pathways, but at the expense of detailed information. The brain's visual M pathways process global spatial structure, object contours, depth, and motion.

In the absence of scene context, the inferior part of OFC matches the object gist to object templates in long-term memory to find the best match or best matches. For example, when the object gist with the mushroom contour in Figure 5.4A is matched to object templates in long-term memory in the absence of scene context, there is no single best match but rather several best matches, such as the object templates for a mushroom, a lamp, and an umbrella.

Although the LSFs projected via the M pathway generate predictions, or hypotheses, about the identity of the objects, the HSFs extracted from the object and processed bottom-up via the P pathway are typically required for the visual system to be able to determine the identity of the object. The hypotheses arrive back in the IT cortex temporally prior to the arrival of the finely detailed bottom-up information, and the fast-arriving

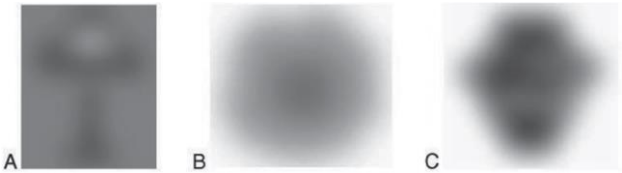

Figure 5.4 Three pictures (256 pixels) of familiar objects (A: lamp, B: flower, and C: vase) filtered to include only the low frequency spatial components (0–4 cycles/picture)

Source: From Bar (2003)

hypotheses are then compared to the slowly arriving fine-grained information. If there is a mismatch between a hypothesis and the fine-grained information, then the fast M pathway sends a prediction error signal to OFC, telling it to update the hypothesis.

Object recognition in real-world scenes proceeds in a similar way. The scene gist (i.e., the LSF information extracted from the scene) is projected directly from V1 to OFC (Bar et al., 2006) (Figure 5.4). But different parts of OFC show selectivity for the gists of objects and scenes. However, the inferior part of the OFC shows selectivity for LSFs extracted from images of isolated objects, the medial part of OFC responds preferentially to LSFs extracted from images of scenes (Aminoff, Kveraga, & Bar, 2013; Bar et al., 2006). The increased activation of the medial areas of OFC recruits a scene template, or what is also sometimes called a *context frame*, *schemata*, a *script*, or a *frame* (Bar, 2004; Friedman, 1979; Palmer, 1975; Schyns & Oliva, 1994). Scene templates are structures in long-term memory, which store statistical scene regularities and derive from past exposure to similar scenes. Once a scene template has been recruited (e.g., a living room scene template), associated object templates are rapidly activated, a process that provides the platform for predictions of which objects are most likely to be found in the scene (e.g., a sofa, a sofa table, a lamp, a television) (Bar, 2009). Activated scene templates constrain expectations with respect to the presence and typical characteristics and location of objects in the scene and provide the ability to direct attention in order to shift gaze to relevant regions of the scene. The scene template serves as a coarse-grained prediction about the distal cause of the scene gist. The scene information is then projected to the IT cortex, where it awaits the later-arriving HSF signal that has been processed bottom-up.

Suppose the task is to determine the distal cause of the mushroom contour in Figure 5.4A in the context of a living room. Although multiple object templates match the object gist with the mushroom contour when processed without a scene gist, one can imagine that only the lamp is a

suitable match when the object gist with the mushroom contour is processed together with the gist of a living room. In the envisaged case, the gist of a living room recruits a living room scene template in long-term memory, which in turn helps narrow down the range of hypotheses about the distal cause of the object gist with the mushroom contour to a single one (viz., the hypothesis that the distal cause of the gist with the mushroom contour is a lamp). Even so, one cannot equate the information projected back to IT with perceptual content, as this signal only consists of LSF information from the mushroom contour and semantic information about a prototypical living room table lamp. But there is obviously more to perceptual content than LSF and semantic prototype information. Perceptual content also contains HSF information, such as information about texture, sharp edges, and colors.

5.4 The Predictive Account and Gist Processing

Let us now turn to one of the problems object recognition presents for the predictive approach, viz., that of accounting for gist processing. Recall that on the predictive account, the only information that gets relayed up through the hierarchy is the prediction error signal, which encodes information about the mismatch between the sensory input and a hypothesis about what caused the signal. A prediction error can also be thought of as information that has not yet been successfully predicted by the hypothesis. Prediction error signals carry information bottom-up in the visual system, eliciting an update of the hypothesis, which is then compared to the lower level sensory signal. This hypothesis-testing process continues until the prediction error is minimized as much as possible. But, as we have just seen, the first step in object recognition is not the generation of a hypothesis but rather the projection of low frequency spatial information – the kind of information that encodes holistic layouts and contours of objects – from V1 to OFC. This gist of a scene or an object recruits a scene or object template encoded in long-term memory. Scene templates and object templates serve as hypotheses about the distal causes of sensory information about scenes and objects, respectively, where the sensory information at this stage is the gist of the object or scene. If a single hypothesis wins out, then this hypothesis is matched against the slower arriving, HSF signal, which has been processed bottom-up. Hypothesis revision continues until the best match has been found. The problem this poses for PP is that the projection of the gist of a scene or object from V1 to OFC cannot be construed as a prediction error signal, as the brain needs to be apprised about its surroundings, at least in broad strokes, before it can generate detailed predictions about its surroundings. If the brain were to start with a randomly chosen hypothesis – a random guess – our vision would fail us far

more often than it actually does, as correcting the potentially gross error of a random guess would be rather time-consuming in most cases.

Andy Clark uses the following analogy to shed light on the idea of a prediction error signal:

> [S]uppose you and I play a game in which I (the "higher, predicting level") try to describe to you (the "lower level") the scene in front of your eyes. I can't see the scene directly, but you can. I do, however, believe that you are in some specific room (the living room in my house, say) that I have seen in the past. Recalling that room as best I can, I say to you "there's a vase of yellow flowers on a table in front of you". The game then continues like this. If you are silent, I take that as your agreeing to my description. But if I get anything that matters wrong, you must tell me what I got wrong. You might say "the flowers are yellow". You thus provide an error signal that invites me to try again in a rather specific fashion—that is, to try again with respect to the colour of the flowers in the vase. The next most probable colour, I conjecture, is red. I now describe the scene in the same way but with red flowers. Silence. We have settled into a mutually agreeable description. (Clark, 2015, p. 5)

The problem with this analogy is that Clark assumes that he (the "higher, predicting level") already believes that he is in a living room, a specific living room indeed. He thus skips right over the first (and the second) stage of object recognition, that is, he ignores that the sensory signal somehow must generate an initial hypothesis, or prediction, about the distal cause of the sensory signal. Otherwise, the initial hypothesis is just a random guess. If we assume that the sensory signal does not initially help shape the formation of a hypothesis, then the game you (the "lower level") play with Clark (the "higher, predicting level") might well run as follows:

Clark: I don't really have any idea where you are. But let me just give it a shot. You are in a living room in Edinburgh.
You: You are wrong about "You are in a living room in Edinburgh."
Clark: How wrong? Is it a room in a house?
You: You are wrong about "You are in a living room in Edinburgh."
Clark: Okay, let me try something more general. You are outside.
You: [Silence]
Clark: Silence means we have settled to a mutually agreeable description. Okay, then. You are outside. You are in your yard.
You: You are wrong about "You are in your yard."
Clark: You are walking your dog.
You: You are wrong about "You are walking your dog."

Clark: You are on a beach.
You: You are wrong about "You are on a beach"
Clark: Bloody hell. I give up …
You: Really! Alright then. I am in the outback of Australia, on a mission to extract venom from the deadly eastern brown snake.
Clark: You're what?

Granted, people often do know at the top predictive level where they are. We bet that you know where you are right now, even without having to open your eyes. This much is true. The problem with this rejoinder is that object recognition does not require knowing where one is. If you are shown slides of different familiar scenes one by one, you can immediately identify the scenes and the objects in them without holding any prior beliefs about what the slides might present. The same goes for object recognition without a scene context. It is possible to recognize familiar objects in isolation of scene context in less than 80 ms without having the slightest hint ahead of time as to what the object on the next slide may be (e.g., Davenport & Potter, 2004).

This example merely serves to drive home the point that you need a bottom-up signal to present at least a general sketch of the scene or object to the prefrontal cortex, so the decision-making part of the brain can generate probable predictions rather than being forced to rely on random guesses. But PP encounters further trouble once we consider how it handles noise, or imprecision, in the sensory input (see also Vance, 2021). Noise can be understood as a meaningless discrepancy between the sensory signal and its distal cause. Externally generated noise in a visual signal may be due to poor viewing conditions, such as morning fog, which makes sensory signals less reliable. Internally generated noise, by contrast, may be due to random deviations in neural firing.

Within the PP framework, updating a hypothesis is supposed to generate a more accurate prediction about the distal cause. Updating a hypothesis on the basis of a noisy, or imprecise, signal, however, is much less likely to give rise to a more accurate prediction. So, PP maintains that noisy, or imprecise, signals have much less influence on the updating of the brain's predictions. To accommodate this idea, advocates of PP assume that in addition to making hypotheses, or predictions, about the distal cause of a sensory signal, the brain also makes predictions about how precise the sensory signal is. The greater precision the brain expects, the greater the gain on the prediction error signal, and the more weight is given to the prediction error in updating the hypothesis. Conversely, if the brain expects a noisy, sensory input, then it attenuates the prediction error signal, thus inhibiting its influence on the update of the hypothesis. But this causes trouble for PP with respect to the gists of scenes and objects.

As we have seen, a substantial body of research shows that in real scenes, the brain first samples LSF information about a scene and propagates it to the prefrontal cortex, where it activates a scene template in long-term memory. The scene template then generates a hypothesis about what sorts of objects are likely to be present in the scene. When dealing with a fixed object, the gist of the object is projected from V1 to the prefrontal cortex, which then activates compatible object templates in long-term memory.

Object gists and scene gists are prime examples of noisy incoming sensory signals. After all, they are encoded in the form of LSF information – information which, by definition, lacks fine details about the object or scene. But PP suggests that the brain attenuates noisy, or imprecise, sensory signals. If, however, the brain attenuated the gists of objects and scenes, then the sensory input would not be able to activate object or scene templates in long-term memory. Accordingly, the brain would not be able to generate an informed prediction about the distal cause of the sensory signal. If, however, sensory signals had not been able to shape predictions about their distal causes through gist signaling, then the brain would have been forced to rely on pure guesswork, and humans and animals would have been unable to perceive objects. The case of object perception thus unveils a flaw in PP's way of dealing with noisy sensory signals.

At this point, advocates of PP may deny that the gists of scenes and objects are noisy sensory signals because they facilitate object recognition. This, however, would be an ad hoc maneuver. The gists of scenes and objects are paradigm examples of noisy signals. For example, when pictures of familiar objects are filtered to include only the LSF components (0–4 cycles/picture), the objects cannot be recognized with high certainty (Bar, 2003) (Figure 5.4). When the distal cause is viewed in the absence of a scene context, the brain ought to expect the object gist with the mushroom contour in Figure 5.4A to lack precision. That is, the brain ought to predict that gists are low-precision signals. After all, in the absence of scene information, the brain has no basis upon which to give priority to the hypothesis that the object gist with the mushroom shape in Figure 5.4A was caused by a living room table lamp rather than the hypothesis that it was caused by a mushroom or an umbrella. So, the brain ought to predict that the object gist in Figure 5.4A is a low-precision signal. But PP holds that low-precision signals are attenuated. So, PP wrongly predicts that gist signals are attenuated.

5.5 Object Recognition Depends on Attention

Given that the gist of a scene can be extracted over the course of a single eye fixation, during which all components of the retinal image have fixed locations, it may be tempting to think that the scene gist is processed

homogeneously across the visual field prior to any involvement of attention. Recent studies, however, have shown that attention plays a pivotal role in scene gist recognition (Larson et al., 2014; see also Berry & Schwartz, 2011)). Although scene gist acquisition occurs within a single fixation, *covert* attention aids in extracting the gist of the scene. Evidence shows that masking central vision during the first 50 ms of eye fixation interferes with visual tasks, such as reading, visual search, scene memory (Võ & Henderson, 2009), and scene gist recognition, whereas masking peripheral vision only interferes with visual tasks when it occurs about 70–100 ms into fixation (Glaholt, Rayner, & Reingold, 2012; Larson et al., 2014). This points to the hypothesis that the type of attention that is operative during a single eye fixation is zoom-out attention, that is, attention that is initially focused in the center of the visual field but then expands diffusely outward into the visual periphery within the first 100 ms of viewing (Figure 5.5). Zoom-out attention is thus a form of (covert) spatial attention.

Once a scene template has been activated, other types of attention determine where we allocate our resources to object recognition. In a visual search task, voluntary attention guides the movement of our eye fixation. However, zoom-out attention is operative during each eye fixation. In the absence of a perceptual task (e.g., visual search), our attentional resources are preferentially allocated to the identification of objects within our peripersonal space – that is, the region immediately surrounding the

Figure 5.5 During a single eye fixation, attention is initially focused in the center of the visual field but then expands diffusely outward into the visual periphery

Source: From Larson et al. (2014)

perceiver, typically within an arm's reach (for a review, see Castelhano & Krzyś, 2020). As a result, objects within peripersonal space are identified more accurately (Fernandes & Castelhano, 2019; Josephs & Konkle, 2019; Man, Krzys, & Castelhano, 2019). This prioritization of information closer to the perceiver is also known as the "foreground bias." As we will explore below, other types of attention can be central to object recognition at the stage of hypothesis testing, including attentional capture, affect-biased, and cued attention. However, the idea that attention facilitates object recognition presents a further problem for the predictive account. Before spelling out the gist of this problem, let us first have a closer look at how the predictive account handles attention. The original proposal, due to Friston, is that attention is the optimization of the expected precision of incoming signals (Feldman & Friston, 2010; Friston, 2009; see also Hohwy, 2013, p. 195):

The Predictive Account of Attention

> To attend to a stimulus just is to turn up the gain on expected high-precision signals while turning down the gain on other signals.

According to PP, to turn up the gain on an expected high precision signal is to enhance the precision of a signal already expected to be highly precise. A prediction error signal with an enhanced gain is given greater weight in the revision of hypotheses about the visual scene. So, PP holds that attention allows a prediction error signal to play a weightier role in hypothesis revision.

However, thus formulated, the PP's account of attention seems to face problems similar to those that fueled the classical debate between early (Broadbent, 1958) and late selection (Deutsch & Deutsch, 1963) in attention, especially to unexpected externally driven salient events like the eruption of a sudden fire or a loud noise in a quiet café. Here, a chicken and egg problem arises: how can the system reliably expect a high precision signal before it knows which object it is processing? One theory that provides a highly effective solution to this problem is the theory of visual attention (Bundesen, 1990). This theory proposes that we change the way we think of the relationship between memory subsystems, so that rather than positing that information flows from short-term memory into long-term memory; this theory takes information from the environment to be matched to mental categories (or templates) in long-term memory (Bundesen & Habekost, 2008). The best match (which could be guided by PP principles)[4] then competes in a stochastic race for active representation in a limited capacity short-term memory, or working memory, store. That is, incoming sensory information is compared to long-term memory and

then attention prioritizes the most relevant categorizations for encoding in short-term memory. Thereby avoiding the problem of how selection can occur before objects have been identified (Dall, Watanabe, & Sørensen, 2016; Brogaard, & Sørensen, 2023).

The details of the predictive account differ for different types of attention, specifically for attentional capture, form of exogenous attention, and cued spatial and voluntary spatial attention, forms of endogenous attention. To a first approximation, we can say that attention is exogenous when it is automatically drawn toward a target (e.g., a spatial region, object, or attribute), whereas attention is endogenous when attention is directed toward a target by an internal state (e.g., volition, expectation, memory, or emotion), sometimes as a result of processing a cue that aids the perceiver in directing her attention to the target.

In attentional capture, the best known form of exogenous attention, attention is grabbed by a stimulus that stands out in some way from its surroundings, such as a scream in a relatively quiet coffee shop, or a red dot in an array of black dots.[5] Attentional capture is thought to be an early visual phenomenon, as perceptual features must be processed early enough in the visual system for them to attract attention and lead to segregation (Beck, 1966; Treisman, 1982). A stimulus captures attention when the associated signal is strong compared to other incoming signals. But now, advocates of PP argue, a signal can reasonably be assumed to be strong due to a higher signal-to-noise ratio compared to weaker incoming signals. Granting this assumption, strong signals are more precise than weaker incoming signals. So, according to PP, the visual system increases the gain on attention-capturing signals. As PP holds that high precision signals have a substantially greater influence on the revision of hypotheses than noisy signals, signals associated with attention-capture thus lead to a significant revision of the brain's existing hypothesis about its surroundings.

Next, let's look at cued spatial attention, one of the most studied forms of endogenous attention. A classic paradigm for studying the effect of cued spatial attention on the speed of detecting a target object is the Posner paradigm (Posner, 1980). Here, volunteers fixate on a central fixation cross. Then a cue appears that points in the direction of the target in 80 percent of trials (valid cue) and in the opposite direction of the target in the remaining 20 percent of trials (invalid cue) (Figure 5.6).

The expected finding in this paradigm is that target stimuli are detected faster when a valid cue directs our attention to the target's spatial location. On the predictive account, the appearance of a cue generates a hypothesis that a target will appear in the cued spatial region, which increases the gain for a high precision signal associated with the target in that region, facilitating detection. If the task involves spotting a particular target (e.g.,

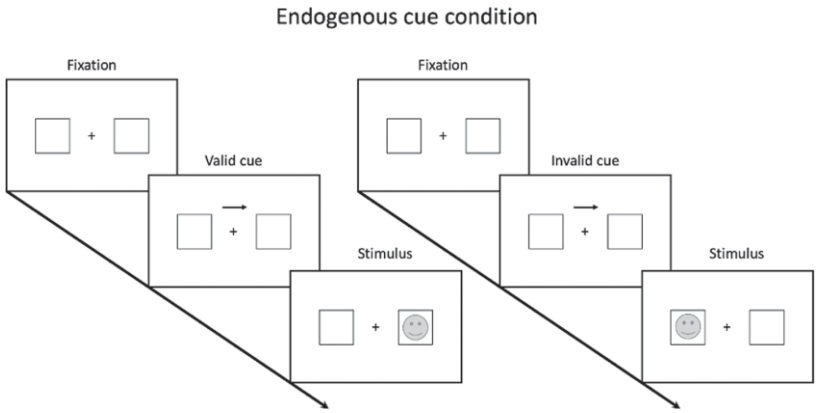

Figure 5.6 Posner paradigm. Valid cues (left) direct attention to the target's spatial location, which allows for faster detection of the target stimulus

Source: Adapted from Posner (1980)

the emotional valence of a smiley, the letter identity, or a Gabor pattern), then cued spatial attention facilitates spotting just those targets.

Voluntary spatial attention, another form of endogenous attention, is not directed by environmental cues (e.g., a pointing finger, a gaze, or an arrow) but rather by the perceiver's internal states (i.e., a desire/belief pair, an intention, or a volition). Hohwy proposes to treat voluntary spatial attention as a kind of action used to test a perceptual hypothesis (Hohwy, 2013, pp. 77–78). The general idea here is that we often try to figure out what the world is like by actively engaging in it, for instance, by walking closer to a target object or inspecting it from different perspectives. Say you expect a wooden construction in front of you to be a real barn but want to rule out that it's a realistic barn facade used in a movie set. You can test your "real barn" hypothesis by making a prediction about what the sensory signal would be like, if your hypothesis were true. For example, unlike a barn facade, a real barn will keep looking like a real barn if you walk around it. So, to test your hypothesis, you can walk around the construction. By doing that, you are sampling more data, thus increasing the "power" of your study. If your "real barn" hypothesis predicts the stimulus well, your walk around the barn will bring about the predicted sensory signal, where the predicted sensory signal here is *the wooden construction still looks like a real barn*. In exploratory perception, then, the prediction error, or mismatch between the hypothesis and your initial sensory signal, is not minimized by revising the hypothesis but rather by acting to bring yourself into a situation where your hypothesis will

match the sensory signal well, if the hypothesis predicts the stimulus well. Hohwy applies this idea of acting to sample additional data to voluntary spatial attention. Voluntary spatial attention, he argues, involves acting for the purpose of testing your hypothesis (Hohwy, 2013, pp. 197–198). But in voluntary attention, the action is not intentional bodily movement, but a mental act, a preparedness, to increase the sensory gain on a high precision signal if one appears in the sampled region of space. If, for example, your hypothesis is that there is a mouse on the kitchen floor, your attention to the kitchen floor increases the sensory gain for a high precision "mouse" signal in that region, resulting in faster detection if a mouse shows up on the floor. Ransom et al. (2017) have recently argued that the predictive approach fails to offer a complete account of voluntary attention. Their test case comes from Ulric Neisser and Robert Becklen's 1975 classical study of "selective looking." Neisser and Becklen used a system of half-silvered mirrors to present two overlapping films of equal quality to participants, appearing in the same segment of their visual field. One depicted actors playing a hand clapping game, filmed from up close, whereas the other depicted actors playing a ball game, filmed from a distance (Figure 5.7).

Neisser and Becklen found that the volunteers were perfectly capable of attending to either film, while ignoring the other, and switching their attention from one to the other, provided that both films were presented to both eyes. Attending both films at once, however, proved to be "demanding" or "impossible." They conclude on the basis of their findings that it is not the distance or clarity of a visual stimulus that enables us to selectively attend to it, nor a "filter" or "gate" created on the spot, but rather the stimulus' intrinsic properties and structure. The type of attention exemplified in the study is voluntary attention, as evidenced by the fact that the volunteers

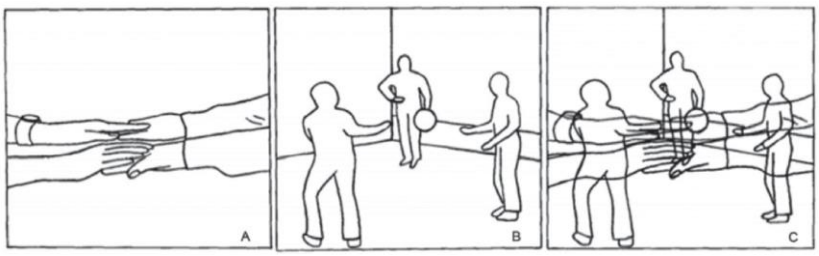

Figure 5.7 Neisser & Becklen's (1975) "Selective Looking" Experiment. Two overlapping films were presented to volunteers, as shown in pane C. One film depicted A hand clapping game, as shown in pane A, and the other depicted A ball game, as shown in pane B

Source: From Ransom et al. (2017)

could choose to attend to either one of the overlapping films and were able to easily switch from one to the other.

According to Ransom et al. (2017), the kind of voluntary attention exemplified in Neisser and Becklen's study presents a problem for the predictive model. The kind of voluntary attention that determines whether a subject is attending to the hand clapping game or the ball game is voluntary *object* attention, where the object is the depicted game. But advocates of the predictive account do not say how we are to understand voluntary object attention. Hohwy (2013), for example, only explains how PP would deal with voluntary *spatial* attention for enhanced target detection. Ransom et al. (2017) thus consider and ultimately reject several ways that advocates of PP could accommodate voluntary target attention. The gist of their argument, however, is this: the participants in Neisser and Becklen's experiment can voluntarily attend to either one of the two overlapping stimuli. But the stimuli differ only in intrinsic features; the signals associated with the stimuli do not differ in context-dependent precision. For example, it is not the case that the hand clapping game is presented in foggy conditions and that the associated sensory signal therefore is less precise; and there is no reason to think the volunteers expect otherwise.

In response to Ransom et al. (2017), it may be argued that rather than being a perceiver-independent matter, the participants' expectations regarding the precision of the sensory signals on which stimulus they choose to attend to. This objection, however, is misguided. If the expected precision of the sensory signal depends on what the volunteers voluntarily attend to, then – on pain of circularity – PP cannot account for voluntary attention in terms of expected precision. Clark (2017) nonetheless insists on something like this response to Ransom et al. (2017). According to Clark, the participants' voluntary switch in attention should be cashed out in terms of their desires, which in turn are to be understood in terms of their predictions about what they will do. As he puts it,

> [D]esires are simply beliefs/predictions that thus guide inference and action (see Friston et al., 2011, p. 157). My desire to *drink a glass of water now* is cast as a prediction that *I am drinking a glass of water now* – a prediction that will yield streams of error signals that may be resolved by bringing the drinking about, thus making the world conform to my prediction. Desires are here re-cast as predictions apt to be made true by action.
>
> Thus consider the prediction (based on some standing or newly emerging belief) that I will now experience, say, the hand-clapping film. This would enslave action, including the 'mental action' of altering the precision-weighting on hand-clappy stuff. In this way

> desires and motivations are revealed as beliefs that enslave action. The apparently non-indicative nature of a thought such as 'let's have a look at the hand-clap film' is now no barrier. For the real content of the thought, as far as the PEM mechanism is concerned, is indicative – it is something like "I am looking at the hand-clap film now." (Clark, 2017, p. 117)

However, Clark's fix simply re-introduces the worry of circularity. Clark's suggestion boils down to this: a participant S attends to a stimulus (the ball game, say) just in case S expects the associated sensory signal to be a high precision signal, but the associated sensory signal is a high precision signal just in case S predicts that S attends to the stimulus. So, S attends to the ball game stimulus just in case S predicts that S attends to the ball game stimulus. The predictive account of attention thus presupposes an account of attention.

In a more recent paper, Ransom et al. (2020) argue that PP also fails to account for affect-biased attention, that is, attention to stimuli that are emotionally, or affectively, salient as a result of their associations with reward or punishment. Their main example of affect-biased attention runs as follows:

> Suppose you walk your dog uneventfully every day past a house on the corner of your block. One morning, however, a large Doberman rushes to the fence, barking and snapping. You jump backwards and for a moment you fear for your life. From this day forward, you give this house a bit of extra attention when you walk past, your eyes always searching the fence for signs of the Doberman, though it is seldom in fact in the yard. (Ransom et al., 2020, p. 1)

Your increased attention to the yard subsequent to your initial encounter with the Doberman is not obviously stimulus-driven, or exogenous. The yard presumably triggers a flashback, an affect-laden memory, which then causes you to pay closer attention to the yard. Affect-biased attention is thus a kind of endogenous attention.

Ransom et al. (2020) argue that affect-biased attention cannot be understood in terms of expectations of a high precision signal, as PP suggests. In the envisaged case, you eventually learn that the Doberman rarely is in the yard. So, your attention to the yard cannot be explained by your expectation that the Doberman will be causing a high precision sensory signal. Rather, it seems that your affect-laden memory of the aggressive dog plays a part in explaining your attention to the yard. Ransom et al. (2020) suggest that it's your desire to shun the yard-associated punishment that drives your attention to the yard, where the yard-associated punishment is the

Doberman lunging at you. The predictive account thus yields the wrong result here, viz. that you attend to the yard because you expect a high precision "Doberman" signal there. Ransom et al.'s (2020) objection points to a more general problem with PP: suppose you step out of the car and jump to the side because it looks like there is a snake under the car. Upon further scrutiny, however, the snake-shaped object is just a stick. Here, the mistaken categorization occurs because if the stick had been a snake, this would be a potentially dangerous situation. This is so, in spite of the fact that you have encountered far more sticks than snakes and therefore should have a higher prior for categorizing the stimulus as a stick.

The problems voluntary object attention and affect-biases attention present for the predictive account are augmented by the role that these types of attention may play in object recognition. While a scene hypothesis is activated within the duration of a single eye fixation, which information is processed bottom-up depends on which object is attended. In a visual search of a complex scene, voluntary attention partially guides eye movements, whereas zoom-out attention covertly scans a spatial region from the visual center to the periphery. The visual search results in the selection of an object. If the scene contains a visually salient object (e.g., a colored object in a black-and-white scene), attentional capture results in a selection of the object. Cued and affect-biased attention could also be what drives the selection of an object. Unless an object is selected, however, the object will appear as a diffusely attended blob with features that are insufficient for confident identification. Attention is thus a precondition for object recognition. Yet, as we have already seen, the predictive account fails to provide a satisfactory account of at least two forms of attention that may assist in the selection of an object, viz. voluntary object attention and affect-biased attention (Ransom et al., 2017, 2020).

The predictive account also lacks the resources to account for the zoom-out attention that occurs during scene gist recognition. Recall that over the course of an eye fixation, which takes around 100 ms, attention is initially highly focused at the center of the visual field and then covertly diffuses from the center to the periphery (Glaholt et al., 2012; Larson et al., 2014). This zoom-out attention also operates during each eye fixations in saccades and visual searches of complex scenes at later stages in the process of object recognition.

The predictive account construes attention as turning up the gain on an expected high precision signal, but although expectations modulate attention (Sørensen et al., 2015; Vangkilde, Coull, & Bundesen, 2012), scene gist extraction seems to occur during a single eye fixation and depends on zoom-out attention (Larson et al., 2014). Nevertheless, in zoom-out attention, covert attention is diffusely distributed around the visual center for about 50–75 ms of eye fixation, and it then diffusely expands outward

during the subsequent 25–50 ms of fixation. During initial zoom-out attention, the visual system expects a low-precision LSF signal extracted from the scene, not a high-prediction signal. As the system does not expect a high precision signal, it should not turn up the gain on the signal, and so, PP is unable to accommodate zoom-out attention; PP thus lacks the resources to account for object recognition on multiple levels.

5.6 The Dark Room Problem

The main problem with PP, as we have seen, lies in its excessive emphasis on top-down predictions. However, you can have too much of a good thing. So, perhaps an easy fix is to tone down the emphasis on top-down predictions. One option is a proposal that replaces prediction error minimization with the ratio of top-down processes to bottom-up processes as the overarching unifying principle (TD:BU ratio) (Herz et al., 2020). Top-down predictions guide the bottom-up signals by enhancing expected signals and inhibiting unexpected signals, whereas bottom-up processes are free of top-down guidance (or disturbances). The ultimate perceptual output is shaped by the TD:BU ratio.

As argued by Herz et al. (2020), different states of minds lead to different TD:BU ratios, including different moods, attentional scopes, and thinking styles. A broader thinking style, for example, entails a lower TD:BU ratio. Reduced guidance and inhibition by top-down processes enables increased associative activation and non-linear thought processes. Narrower thinking, by contrast, is linked to a higher TD:BU ratio. The increased top-down processing helps prevent distraction by competing thoughts, inhibits free associative activation, and thus results in a narrower and more ruminative style of thinking (e.g., Smith & Alloy, 2009).[6]

The state of mind framework may help address the so-called dark room problem for PP (Klein, 2018; Mumford, 1992; Sun & Firestone, 2020). The gist of the dark room objection is that if PP is right that the principle of prediction error minimization lies at the heart of all of our mental processes, then we should be biologically driven to stay inside a dark room. The dark room that drives the objection is not just pitch-dark, but also quiet, non-smelly, unfelt, and so on.[7] So, inside the dark room, no sensory inputs from the environment reach us. With no sensory inputs entering our brain, there cannot be a mismatch, or prediction error, between sensory inputs and our predictions about our environment. As long as we stay inside the dark room, no evidence could ever cause us to update our predictions. No matter how tame or wild we predict that the dark room really is, our predictions go unchallenged. So, staying in the dark room seems to be a much more effective way of minimizing prediction errors than entering the outside chaotic world (however, see also Van de Cruys, Friston, & Clark,

2020). But, as a matter of fact, we do not stay in a dark room, in fact most seem motivated to gather in "noisy" cities. So, taking the prediction error minimization principle to be fundamental to our brain and mind, as PP does, seems misguided.

This is not the place to consider PP's replies to the dark room problem.[8] It should, however, be clear that if we take the TD:BU ratio to explain the mind, then the dark room problem no longer rears its shady head. Rather, if the mind operates in different modes, involving different TD:BU ratios, then staying in a dark room should be appealing to us only when we are in a hyper-narrow state of mind entailing a sky-high TD:BU ratio – the TD:BU ratio characteristic of people in depressive states characterized by complete apathy.

The states of mind proposal can be seen as an augmentation of PP, but it should be emphasized that accepting the states of mind proposal entails denying some of the central claims made by PP, for instance, that prediction error signals are the only signals processed bottom-up and that prediction error minimization is the overarching principle explaining all our mental processes.

Notes

1 Predictive processing is far from the only Bayesian approach to the brain and cognition. For a review of Bayesian approaches, see, e.g.,Talbott (2016), Spratling (2017).
2 While perceptual states are the product of unconscious, subpersonal inferences, mental states themselves, including judgment and desire, are personal-level states (Clark, 2020; Wiese & Metzinger, 2017).
3 For the original demonstration, see KSU, Vision Cognition Laboratory, https://www.k-state.edu/psych/vcl/images/beach%20loop.gif, retrieved Oct 31, 2018.
4 TVA assumes that sensory evidence matched with templates in long-term memory provides the initial basis for stimulus encoding that is modulated by additional top-down mechanisms of pertinence and bias. However, the specific mechanism in this template matching procedure is not entirely clear, and we propose that PP could in fact be that exact mechanism. Thus, PP may be a key mechanism in perception, but not an exclusive unified mechanism in perceptual processing.
5 It may be argued that attentional capture is the only form of exogenous attention. However, here we leave room for other forms of exogenous attention, for example, stimulus-driven diffuse attention and what Azenet Lopez (2020, ch. 4) calls "spillover attention." According to Lopez, spillover attention is attentional allocation to a vicarious or secondary target, such as the bearer of a feature in the case of feature attention (for the most recent version of her view, see Lopez, 2022). The most intuitive cases of spillover attention are instances of endogenous attention. But given that attentional capture entails attentional selection, presumably spillover attention could be exogenous as well.
6 For a review of the relationship between rumination and depression, see, e.g., Thomsen (2006).

7 On Mumford's (1992) variation on the dark room problem, you are to envisage a place, "like the oriental Nirvana … when nothing surprises you and new stimuli cause the merest ripple in your consciousness" (1992, p. 247, fn. 5). Here, there are sensory inputs, but they don't move you the least bit. This version of the dark room is more akin to a depressive state characterized by complete apathy.

8 Sun and Firestone (2000) argue that various intuitive responses to the dark room problem ultimately fail. For example, it's highly predictable that we will get hungry in the dark room, but as Klein (2018) notes "predicting hunger is not the same as being motivated by it." Sun & Firestone acknowledge that Friston's (2013) reply succeeds in solving the problem but only by introducing a new one. Friston argues that the dark room problem rests on the mistaken assumption that the dark room is not surprising. As he puts it, "the state of a room being dark is surprising, because we do not expect to occupy dark rooms." The problem with this reply, Sun & Firestone argue, is that it makes PP trivially true, as no behavior can count as evidence against the view: "Why do we dance? Because we predict we won't stay still. Why do we donate to charity? Because we predict we will do good deeds. Why do we seek others? Because the brain has a prior which says 'brains don't like to be alone'" (pp. 347–348). See also Kwisthout et al. (2017), who present an interesting variation on the dark room problem based on the idea that coarse-grained predictions are more likely to minimize prediction error than fine-grained predictions.

References

Aminoff, E. M., Kveraga, K., & Bar, M. (2013). The role of the parahippocampal cortex in cognition. *Trends in Cognitive Sciences*, *17*, 379–390.

Auckland, M. E., Cave, K. R., & Donnelly, N. (2007). Non-target objects can influence perceptual processes during object recognition *Psychonomic Bulletin Review*, 14, 332–337.

Bar, M. (2003). A cortical mechanism for triggering top–down facilitation in visual object recognition. *Journal of Cognitive Neuroscience*, *15*, 600–609.

Bar, M. (2004). Visual objects in context. *Natural Reviews Neuroscience*, *5*, 617–629.

Bar, M. (2009). The proactive brain: Memory for predictions. *Philosophical Transactions of the Royal Society B*, *364*, 1235–1243.

Bar, M., Kassam, K. S., Ghuman, A. S., Boshyan, J., Schmidt, A. M., Dale, A. M., Hamalainen, M. S., Marinkovic, K., Schacter, D. L., Rosen, B. R., & Halgren, E. (2006). Top-down facilitation of visual recognition. *Proceedings of the National. Academy of Sciences U.S.A*, *103*, 449–454.

Beck, J. (1966). Effect of orientation and of shape similarity on perceptual grouping. *Perception & Psychophysics*, *1*, 300–302.

Berry, M. J., & Schwartz, G. (2011). The retina as embodying predictions about the visual world. In Bar, M. (ed.), *Predictions in the brain: Using our past to generate a future* (pp 295–308). Oxford: Oxford University Press.

Biederman, I., Mezzanotte, R. J., & Rabinowitz, J. C. (1982). Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology*, *14*(2), 143–177.

Broadbent, D. (1958). *Perception and communication*. London: Pergamon Press.
Brogaard, B., & Sørensen, T. A. (2023). Perceptual variation in object perception: A defence of perceptual pluralism. In A. Mroczko-Wąsowicz, & R. Grush (eds.), *Sensory Individuals: Unimodal and Multimodal Perspectives* (pp 113–129). Oxford: Oxford University Press.
Bundesen, C., & Habekost, T. (2008). *Broadbent, D. (1958). Perception and communication*. London: Pergamon Press.
Castelhano, M. S., & Krzyś, K. (2020). Rethinking space: A review of perception, attention, and memory in scene processing. *Annual Review of Vision Science*, *6*(1), 563–586.
Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(3), 181–204.
Clark, A. (2015). Embodied prediction, https://uberty.org/wp-content/uploads/2017/06/Embodied-Prediction.pdf
Clark, A. (2016). *Surfing uncertainty: Prediction, action, and the embodied mind*. New York: Oxford University Press.
Clark, A. (2017). Predictions, precision, and agentive attention. *Consciousness and Cognition*, *56*, 115–119.
Clark, A. (2020). Beyond desire? Agency, choice, and the predictive mind. *Australasian Journal of Philosophy*, *98*(1), 1–15.
Dall, J. O., Wang, Y., Cai, X., Chan, R. C., & Sørensen, T. A. (2021). Visual short-term memory and attention: An investigation of familiarity and stroke count in Chinese characters. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 47(2), 282–294. https://doi.org/10.1037/xlm0000950
Dall, J. O., Watanabe, K., & Sørensen, T. A. (2016, February). Category specific knowledge modulates capacity limitations of visual short-term memory. In 2016 *8th international conference on knowledge and smart technology* (KST). IEEE, 275–280.
Davenport, J. L., & Potter, M. C. (2004). Scene consistency in object and background perception. *Psychological Science*, *15*(8), 559–564.
Deutsch, J. A., & Deutsch, D. (1963). Attention: Some theoretical considerations. *Psychological Review*, *70*(1), 80–90.
Feldman, H., & Friston, K. (2010). Attention, uncertainty, and free-energy. *Frontiers in Human Neuroscience*, *4*, 215.
Fernandes, S., & Castelhano, M. S. (2019). The foreground bias: Initial scene representations across the depth plane. *PsyArXiv*. https://doi.org/10.31234/OSF.IO/S32WZ
Fiser, J., & Aslin, R. N. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological Science*, *12*, 499–504.
Fiser, J., & Aslin, R. N. (2005). Encoding multi-element scenes: Statistical learning of visual feature hierarchies. *Journal of Experimental Psychology General*, *134*, 521–537.
Friedman, A. (1979). Framing pictures: The role of knowledge in automatized encoding and memory of gist. *Journal of Experimental Psychology General*, *108*, 316–355.
Friston, K. J. (2003). Learning and inference in the brain. *Neural Network*, *16*(9), 1325–1352.

Friston, K. J. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *360*, 815–836.

Friston, K. J. (2009). The free-energy principle: A rough guide to the brain? *Trends Cognitive Science*, *13*(7), 293–301.

Friston, K. J. (2010). The free-energy principle: A unified brain theory? *Natural Reviews Neuroscience*, *11*(2), 127–138.

Friston, K. J. (2013). Active inference and free energy. *Behavioral and Brain Science*, *36*, 212–213.

Glaholt, M. G., Rayner, K., & Reingold, E. M. (2012). The mask-onset delay paradigm and the availability of central and peripheral visual information during scene viewing. *Journal of Vision*, *12*(1), 9.

Gordon, R. D. (2004). Attentional allocation during the perception of scenes. *Journal of Experimental Psychology Human Perception and Performance*, *30*, 760–777.

Green, C., & Hummel, J. E. (2006). Familiar interacting object pairs are perceptually grouped. *Journal of Experimental Psychology Human Perception and Performance*, *32*, 1107–1119.

Herz, N., Baror, S., & Bar, M. (2020). Overarching states of mind. *Trends in Cognitive Sciences*, *24*, 184–199.

Hock, H. S., Gordon, G. P., & Whitehurst, R. (1974) Contextual relations: The influence of familiarity, physical plausibility, and belongingness. *Perception and Psychophysics*, 16, 4–8.

Hoffman, J. (1996). Visual object recognition. In W. Prinz, & B. Bridgeman (Eds.), *Handbook of perception and action* (Vol. *1*, pp. 297–344). New York: Academic Press.

Hohwy, J. (2012). Attention and conscious perception in the hypothesis testing brain. *Frontiers in Psychology*, 3(96), https://doi.org/10.3389/fpsyg.2012.00096.

Hohwy, J. (2013). *The predictive mind*. Oxford: Oxford University Press.

Hohwy, J. (2020). New directions in predictive processing. *Mind & Language*, *35*(2), 209–223.

Hollingworth, A., & Henderson, J. M. (1998). Does consistent scene context facilitate object detection. *Journal of Experimental Psychology General*, *127*, 398–415.

Josephs, E. L., & Konkle, T. (2019). Perceptual dissociations among views of objects, scenes, and reachable spaces. *Journal of Experimental Psychology Human Perception and Performance*, *45*(6), 715–28.

Klein, C. (2018). What do predictive coders want? *Synthese*, *195*, 2541–2557.

Kondo, H. M., van Loon, A. M., Kawahara, J.-I., & Moore, M. C. J. (2017). Auditory and visual scene analysis: An overview. *Philosophical Transactions of the Royal Society B Biological Science*, *372*(1714), 20160099.

Kveraga, K., Boshyan, J., & Bar, M. (2007). Magnocellular projections as the trigger of top-down facilitation in recognition. *Journal of Neuroscience*, *27*, 13232–13240.

Kwisthout, J., Bekkering, H., & Rooij, I. (2017). To be precise, the details don't matter: On predictive processing, precision, and level of detail of predictions. *Brain and Cognition*, *112*, 84–91.

Larson, A. M., Freeman, T. E., Ringer, R. V., & Loschky, L. C. (2014). The spatiotemporal dynamics of scene gist recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(2), 471–487.

Lopez, A. (2020). *Information gating and the structure of consciousness,* Doctoral dissertation at the University of Miami.

Lopez, A. (2022). Vicarious attention, degrees of enhancement and the contents of consciousness. *Philosophy and the Mind Sciences*, *3*(1). https://doi.org/10.33735/phimisci.2022.9194

Man, L., Krzys, K., & Castelhano, M. (2019). The foreground bias: Differing impacts across depth on visual search in scenes. *PsyArXiv*. https://doi.org/10.31234/OSF.IO/W6J4A

Meyers, L. S., & Rhoades, R. W. (1978). Visual search of common scenes. *Quarterly Journal of Experimental Psychology*, *30*, 297–310.

Mumford, D. (1992). On the computational architecture of the neocortex. *Biological Cybernetics*, *66*(3), 241–251.

Neisser, U., & Becklen, R. (1975). Selective looking: Attending to visually specified events. *Cognitive Psychology*, *7*(4), 480–494.

Oliva, A., & Torralba, A. (2007). The role of context in object recognition. *Trends in Cognitive Sciences*, *11*(12), 520–527.

Palmer, S. E. (1975). The effects of contextual scenes on the identification of objects. *Memory and Cognition*, *3*, 519–526.

Posner, M. I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology*, *32*, 3.

Ransom, M., Fazelpour, S., & Mole, C. (2017). Attention in the predictive mind. *Consciousness and Cognition*, *47*, 99–112.

Ransom, M., Fazelpour, S., Markovic, J., Kryklywy, J., Thompson, E. T., & Todd, R. M.. (2020). Affect-biased attention and predictive processing, *Cognition*, 203, https://doi.org/10.1016/j.cognition.2020.104370.

Schyns, P. G., & Oliva, A. (1994). From blobs to boundary edges: Evidence for time- and spatial-dependent scene recognition. *Psychological Science*, *5*, 195–200.

Spratling, M. W. (2017). A review of predictive coding algorithms. *Brain and Cognition*, *112*, 92–97.

Smith, J. M., & Alloy, L. B. (2009). A roadmap to rumination: A review of the definition, assessment, and conceptualization of this multifaceted construct. *Clinical Psychology Review*, *29*(2), 116–128.

Sun, Z., & Firestone, C. (2020). The dark room problem. *Trends in Cognitive Sciences*, *24*(5), 346–348.

Sørensen, T. A., Vangkilde, S., & Bundesen, C. (2015). Components of attention modulated by temporal expectation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(1), 178–192.

Talbott, W. (2016). Bayesian epistemology, In Edward N. Zalta (ed.) (2016) *The stanford encyclopedia of philosophy*. https://plato.stanford.edu/archives/win2016/entries/epistemology-bayesian/

Thomsen, D. K. (2006). The association between rumination and negative affect: A review. *Cognition and Emotion*, *20*(8), 1216–1235.

Torralba, A., Oliva, A., Castelhano, M., & Henderson, J. (2006). Contextual guidance of attention in natural scenes: The role of global features on object search. *Psychological Review*, *113*, 766–786.

Trapp, S., & Bar, M. (2015). Prediction, context and competition in visual recognition. *Annals of the New York Academy of Sciences*, *1339*, 190–198.

Treisman, A. (1982). Perceptual grouping and attention in visual search for features and for objects. *Journal of Experimental Psychology: Human Perception and Performance*, *8*(2), 194–214.

Van de Cruys, S., Friston, K., & Clark, A. (2020). Controlled optimism: Reply to sun and firestone on the dark room problem. *Trends in Cognitive Sciences*, *24*(9), 1–2.

Vance, J. (2021). Precision and perceptual clarity. *Australasian Journal of Philosophy*, *99*(2), 379–395.

Vangkilde, S., Coull, J. T., & Bundesen, C. (2012). Great expectations: Temporal expectation modulates perceptual processing speed. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(5), 1183–1191.

Võ, M. L. H., & Henderson, J. M. (2009). Does gravity matter? Effects of semantic and syntactic inconsistencies on the allocation of attention during scene perception. *Journal of Vision*, *9*(3), 24–24.

Wiese, W., & Metzinger, T. (2017). Vanilla PP for philosophers: A primer on predictive processing. In *philosophy and predictive processing*. Frankfurt am Main: MIND Group.