Forthcoming in *Synthese*

Perceptual justification in the Bayesian brain: A foundherentist account

Paweł Gładziejewski

Nicolaus Copernicus University in Toruń, Department of Cognitive Science

ul. Fosa Staromiejska 1a, 87-100, Toruń, Poland

Abstract: In this paper, I use the predictive processing (PP) theory of perception to tackle the question of how perceptual states can be rationally involved in cognition by justifying other mental states. I put forward two claims regarding the epistemological implications of PP. First, perceptual states can confer justification on other mental states because the perceptual states are themselves rationally acquired. Second, despite being inferentially justified rather than epistemically basic, perceptual states can still be epistemically responsive to the mind-independent world. My main goal is to elucidate the epistemology of perception already implicit in PP. But I also hope to show how it is possible to peacefully combine central tenets of foundationalist and coherentist accounts of the rational powers of perception while avoiding the well-recognized pitfalls of either.

Key-words: Bayesian brain; coherentism; epistemic justification; foundationalism; perceptual justification; predictive processing

1. Introduction

Philosophers sometimes posit that we need to carefully distinguish the causes of our beliefs from the reasons for our beliefs (Davidson 1986; McDowell 1996; Sellars 1956). Acting as a reason means playing a normative role: some mental states serve as reasons for other mental states by epistemically justifying or warranting those states. However, nothing counts as a reason for a mental state simply by being causally involved in producing that state. And the empirical sciences of cognition, the story goes, only have a purely causal story to tell. For example, in the case of perception, cognitive (neuro)science uncovers the causal transactions that lead from physical energies impinging on the sensory apparatus to perceptual states representing the external environment, which in turn causally shape beliefs. It seems that a causal story like this by itself cannot solve the problem of how perceptual states could play a rational role.

Intriguingly, a particular development in scientific theorizing about causal mechanisms of perception seems to blur the reason/cause divide. According to the predictive processing (PP) account that I have in mind, perceiving involves inverting a generative model of the environment to estimate the most likely causes of sensory stimulation. This inversion is achieved by minimizing the prediction error or the discrepancy between estimate-based predictions and the actual sensory stimulation.

Proponents of PP usually posit it as a mechanistic or process-level theory that aims to reveal the causal mechanisms responsible for perception (Badcock, Friston, Ramstead et al. 2019; Hohwy 2018). However, the ways in which proponents of PP sometimes describe these mechanisms are rich in epistemically normative notions and ideas. Perhaps most relevant here is the idea that perception is unconscious inference in which the brain performs approximate Bayesian updating to make sense of the environment. That is, the causal transactions that output perceptual states conform to a normative rule of statistical inference. In fact, some PP-based explanations of cognitive phenomena constitute rational reconstructions of sorts: they purport to show that the brain solves perceptual problems in a way that conforms to the way it *should* solve those problems *if* it were a probabilistically rational system (see Hohwy, Roepstorff, Friston 2008).

The purpose of this paper is to argue that PP indeed comes with a built-in normative account of the rational powers of perception that is both original and philosophically valuable. It is original because it is neither foundationalist nor coherentist, but rather combines central tenets of both positions to yield a particular variant of hybrid or 'foundherentist' view (for earlier expositions of this sort of intermediate position, see: Gupta 2006; Haack 1993; I follow Haack in using the term 'foundherentism'). It is philosophically valuable because it addresses two major problems that have plagued traditional accounts of perceptual justification.

The first problem relates to the question of how perceptual states acquire their ability to confer justification on other mental states. Here, the background idea is that epistemic justification is transferred from one mental state to another *via* the inferential links between respective states or their contents. On a familiar foundationalist picture, this inferential chain terminates in perceptual states, which act as 'unjustified justifiers,' that is, have the ability to confer justification on other states without requiring to be themselves inferentially justified. But a dilemma can be raised here (BonJour 1985; Sellars 1956). If perceptual states are considered representational or contentful in a way that allows them to *inferentially* transmit their justification to beliefs, then they are precisely the sort of states that require to be *inferentially* justified by other, antecedent representational states. This conclusion, however, would be at odds with the very idea of foundationalism. If we instead treat perceptual states as contentless raw feels or mere sensory registrations, then even if they can causally shape beliefs, it is hard to see how they could feature in inferential, justification-conferring relations with representational states.

The second problem about perceptual justification arises if we abandon foundationalism and instead settle on a view that perceptual states draw their power to justify from the fact that they themselves are inferentially justified. In particular, on such a coherentist picture, perceptual states can be regarded as justified in virtue of participating in a broader coherent system of inferentially linked states. For example, percepts could be 'cognitively spontaneous' beliefs (BonJour 1985). Although exogenously *caused* rather than arrived at through inference, such beliefs are inferentially *justified* by other beliefs belonging to a wider coherent set (e.g., meta-beliefs about what the content of the relevant belief is, about the relevant belief being cognitively spontaneous, about the conditions in which the relevant belief is formed, and about how reliable the perceptual apparatus is when producing cognitively spontaneous beliefs with such contents in such conditions). This is where the problem I have in mind – the 'input' problem – arises. How can mere coherence among mental states make those states epistemically responsive to how things really are in the mind-independent world? Traditionally, it was perception that was thought to act as an epistemic interface of this sort. However, on the coherentist view, the justification of perceptual states, just like the justification of beliefs, is a matter of purely internal relations among mental states. To put this in different terms: since there could be many alternative, equally coherent but mutually incompatible sets of beliefs *which include cognitively spontaneous beliefs* (BonJour 1997), it is hard to see how perception could be truth-conductive just in virtue of featuring in a coherent system.

I will argue that the PP's nuanced view of how perceptual states are formed accounts for perceptual justification in a way that addresses both those problems. It answers the question of where the rational power of perception originates along the coherentist lines. That is, percepts can justify other mental states because they acquire a positive epistemic status (justification) in virtue of having a rational etiology of their own. Simultaneously, the input problem is dealt with by pointing to how sensory signals constrain perceptual inference. Although sensory signals are mere registrations devoid of intentional content, they embody a statistical structure that – when properly hooked to a wider Bayesian machinery – drives perceptual inferences to produce accurate representations of the environment.

The discussion is structured as follows. In section 2, I briefly discuss the PP account of perception, focusing on the notion of perception as approximate Bayesian inference. In Section 3, I discuss how perceptual estimates of the environment owe their positive epistemic status to their rational etiology. In Section 4, I supplement this picture with a (quasi-)foundationalist element, focusing on the evidential role of statistical patterns in the sensory signal. In Section 5, I use the notion of epistemic convergence to outline how perception could generate justification that is unconditional or absolute. I briefly conclude in Section 6.

2. Perception, Bayes, and predictive processing

PP is founded on an idea that perceiving amounts to unconsciously inferring hidden worldly causes of sensory stimulation (for extensive discussions, see Bogacz 2017; Clark 2013b, 2016; Friston 2005, 2010; Hohwy 2013, 2020; Rao, Ballard 1999; for major historical precursors, see Gregory 1980; Helmholtz 1855; Peirce 1900, CP-5.181; the particular version of PP that I employ in the present article originates from Friston 2005). The causes are considered 'hidden' because the causal structure affecting the organism's sensory apparatus cannot be directly 'read off' from the sensory stimulation. Sensory states are underdetermined by their external causes: multiple different causes could produce the same sensory effect, and the same cause can produce many distinct sensory effects depending on circumstances. In fact, in most natural circumstances, the flow of sensory input is produced by multiple interacting causes, and there are many alternative ways of 'unmixing' the input to produce a perception of a scene.

According to PP, to deal with sensory ambiguity, the brain or the perceptual system attempts to infer the causes of sensory stimulation in a way that approximately conforms to Bayes rule:

$$p(h|d) = \frac{p(d|h)p(h)}{p(d)}$$

The idea, then, is that a particular hypothesis (about the causes of sensory stimulation) is selected from some hypothesis space H in a way that maximizes posterior probability $p(h|d)$. The posterior probability measures how likely a hypothesis is to be true, given the sensory data. Computing the posterior requires combining the prior belief about the probability of the hypothesis being true, $p(h)$, and the likelihood $p(d|h)$ which measures how likely the data is, given the hypothesis.

However, calculating the probability of the evidence itself – $p(d)$ in the right-hand side denominator above – is intractable for a hypothesis space of non-trivial size. Realistically, then, a biological system can at most approximate exact Bayesian inference.

In PP, the central theoretical posit for explaining how this approximation proceeds is the generative model. The proposal is that the brain's feedback or top-down synaptic connections implement a statistical model of the environment. The model is 'generative' because it aims to capture how the sensory input depends on the causal structure that generates it. Formally, the generative model constitutes a Bayesian network that encodes joint probability distribution $p(h, d)$ where h stands for hypotheses about the distal causes of the sensory input and d stands for the sensory input. Nodes comprising this network encode the hidden causes, and the edges capture the statistical dependencies between those causes and between the causes and their sensory effects.

With the generative model of the environment in hand, approximating the posterior becomes achievable using variational inference instead of exact inference. An initial guess is made about the true posterior distribution, denoted $q(h)$. This guess is then iteratively updated to bring it ever closer to the true posterior, i.e., the posterior probability that exact Bayesian inference would yield, under the generative model. Formally, this is equivalent to minimizing the Kullback-Leibler divergence (KL-divergence) between the estimated and the true posterior, which is a measure of the difference between those two distributions.

Here, however, another complication emerges. The KL-divergence cannot be directly estimated (after all, the brain or the perceptual system cannot 'know' in advance what the true posterior is to compare it to the approximate posterior). Still, the KL-divergence can be *indirectly* minimized through minimizing another quantity, known as variational free energy (VFE, denoted below as *F*):

$$F = \sum\nolimits_{h \in H} q(h) \, \log \frac{q(h)}{p(h, d)}$$

The lower the VFE, the lower the KL-divergence between the true and approximate posterior. Crucially, the right-hand side of the above equation only includes the 'guess,' denoted by $q(h)$, and the internally-encoded generative model, denoted by $p(h, d)$. This means that approximating the posterior can be indirectly achieved by computing over quantities accessible for the brain or the organism 'from the inside' (this will become relevant later).

The notion of prediction error minimization aims to capture how neural information processing mechanisms perform variational inference. Notice that the generative model is equal to the product of a prior distribution over the hypotheses and the likelihood distribution expressing the probability of obtaining the data, given the hypotheses:

$$p(h, d) = p(d|h) \, p(h)$$

As such, the generative model can be used to simulate the flow of sensory input through drawing 'fictive' samples: one takes the most likely hypothesis from the prior distribution and then, using the likelihood distribution, generates mock input that is most likely under this hypothesis (Greshman 2019).

In perception, these fictive samples – implemented as top-down 'prediction' signals – are compared to actual sensory samples obtained by interacting with the environment. Prediction error measures the difference between the sensory input and the internal predictions (we can regard prediction error as equivalent to variational free energy). The signal conveying the prediction error is propagated bottom-up, which in turn drives the adjustment of perceptual hypotheses or estimates.[1] Over short timescales, this process entails inverting the generative model to yield an approximate posterior, given current sensory data. This is perceptual inference. Over longer timescales, the generative model's parameters can be optimized (through gradient descent on the prediction error) so that the model becomes increasingly effective at reducing average prediction error, calculated over many instances of perceptual inference. This is perceptual learning.

Heuristically, the reasoning behind all this is the following. Imagine building a statistical model of your environment by engaging in exact Bayesian inference. You start with some priors (perhaps initially set at random), then sample data, calculate the posteriors to accommodate the samples as they come in, and, through conditionalization, iteratively use the posteriors to update your priors. Over time, this process should allow you to develop a model that is increasingly better at minimizing the prediction error: new data should be more and more predictable, given

[1] Throughout this paper, I employ the terms 'perceptual state', 'perceptual representation', 'perceptual hypothesis' and 'perceptual estimate' interchangeably. I also take it that perceptual *experiences* are, simply, conscious perceptual states. However, I do not take the perceptual state's being conscious as either necessary or sufficient for it to be epistemically appraisable or to play a justificatory role (see Berger, Nanay, Quilty-Dunn 2018).

your updated priors. The crux of PP lies in the idea that a biological system performs approximate Bayesian inference, so to speak, by flipping the exact Bayesian inference: ‘It is reasonable to turn around the observation that inference leads to less prediction error, and say that, if a system is somehow able to continuously minimise prediction error, then the system will *approximate Bayesian inference*’ (Hohwy 2020, p. 3).

The remainder of this paper will focus on the epistemological consequences of the notion that approximate Bayesian inference underlies perception. Before I move on, however, let me briefly discuss three other central tenets of PP to make the exposition of the theory more comprehensive.

First, PP postulates that the generative model is hierarchically structured, reflecting the nested causal structure producing the sensory data. Lower levels predict the flow of rapidly changing regularities in sensory input (e.g., edges in the visual field), while higher levels predict increasingly invariant patterns which emerge over longer timescales (e.g., objects in space and time). An exemplary three-level generative model may be expressed as $p(h_1, h_2, h_3, d)$, with h_1–h_3 corresponding to distinct inferential levels, and can be thus decomposed into the product: $p(d|h_1)p(h_1|h_2)p(h_2|h_3)\ p(h_3)$.

Second, PP postulates the existence of a mechanism that regulates the degree to which perceptual inference relies on the antecedent assumptions versus the need to fit incoming data. In particular, the brain or perceptual system is thought to actively estimate the precision of the input signals (precision is an inverse of variance, effectively tracking the credibility of the input). Estimations of precision regulate the weight given to the prediction error signals in perceptual inference, with perceptual inference becoming more reliant on priors when input is estimated as noisy. In PP, precision estimation is postulated as an explanation of attention.

Third, the notion of prediction error minimization can be extended to encompass motor control (see Friston 2010; Smith, Friston, Whyte, preprint). The resulting theory, dubbed

‘Active Inference,’ is premised on a notion that along with priors about the causes of sensory stimulation, the generative model encodes priors about ‘policies’ or ways in which an organism expects itself to act. This way, action becomes an attempt to bring the world in line with predictions, thus minimizing prediction error minimization. One interesting corollary of the mathematics used to express Active Inference is that organisms acting to minimize prediction error should sometimes engage in ‘epistemic’ active inference, actively exploring their environments to reduce uncertainty about the causal structure producing the sensory signal.[2]

3. Predictive processing and the rationality of perception

In this section, I lay out a thoroughly inferentialist reading of perceptual justification under PP. In 3.1, I introduce the view that perceptual inferences confer epistemic justification on resulting perceptual states. In 3.2, I argue that perceptual justification is conditional, i.e., relativized to the rational standing of the generative model, and that PP’s coherentist story cannot be turned foundationalist by appeal to the reliable acquisition of perceptual priors.

[2] It needs to be noted that with the introduction of Active Inference, the philosophical interpretation of the resulting theory arguably changes. This shift puts into question whether the theory can be applied to epistemological issues along the lines I advertise in the present paper. In particular, forceful arguments have been put forward for the claim that Active Inference is best read along enactivist rather than representationalist and inferentialist lines (c.f. Bruineberg, Kiverstein, Rietveld 2016; Ramstead, Kirchoff, Friston 2019). Moreover, instrumentalist reading of the generative models in Active Inference has been proposed, construing them as a tools for scientifically modeling the organism and its behavior, rather than as a brain-implemented representational structures (van Es 2020). I will have to set these issue aside here. It may be worthwhile to add, however, that some authors working at the forefront of Active-Inference-based modeling have argued that there is still some room for (inferentially-updated) internal representations within this framework (see Constant, Clark, Friston 2020; Ramstead, Friston, Hipolito 2020).

3.1. Predictive processing involves inferential justification of perceptual states

The preceding discussion established that perception, according to PP, is inferential in the sense that the causal transitions involved in turning raw sensory data into perceptual states representing the environment conform to a normative, rational rule of inference. As stated in the previous section, the causal processes that output perceptual states do not comply with *exact* Bayesian inference. For example, perception on the PP view need not involve a step in which the marginal probability of sensory evidence, $p(d)$, is computed. Still, per what the Bayes rule prescribes, to arrive at a perceptual hypothesis, the brain or perceptual system relies on predictions drawn from the prior and the prior likelihood distributions encoded in the generative model. Although the resulting perceptual state, produced by minimizing prediction error, can initially 'wobble around,' deviating from the true posterior (see Bogacz 2017; Hohwy 2017), it is bound to eventually converge on a value that is close (to some approximation) to the true posterior. Even if the overlap between Bayesian inference and the approximation realized by the causal mechanisms of perception is not perfect, it seems non-trivial enough to render perception genuinely Bayesian.[3]

[3] Note that there on a liberal usage of the term, for an algorithm to 'approximate' Bayesian inference, it is sufficient that this algorithm reliably produces output states that match the ones that exact Bayesian inference would output (given certain priors). On this approach, Bayesian inference could in principle be approximated by using a look-up table (see Maloney, Mamassian 2009). Arguably, more than a simple table-lookup is required for a process to count as *inferential*, so the output-matching approach allows for non-inferential processes to count as approximating Bayesian inference. It is important, then, to stress that the notion of 'approximating' Bayes at use in this paper is *not* this liberal. In particular, on the present view, the *process* by which the outputs (i.e. approximated posteriors) are generated mimics certain crucial aspects of exact Bayesian inference. Most importantly, the present view assumes that that to compute posteriors, the brain uses as input the prior and likelihood distributions encoded in the generative model. Error-minimization is what drives the input-output

A natural construal of perception from this point of view is in terms of inference to the best explanation (Hohwy 2013). As the sensory apparatus registers streams of data, there will be multiple alternative ways of explaining or interpreting the data in terms of their distal causes. Maximizing the posterior probability of hypotheses is a strategy for finding a unique interpretation (explanatory hypothesis), among many possible, which is most likely true or accurate, given the generative model. But posterior probability can also be interpreted as a measure of the inferential fit or coherence between a candidate perceptual hypothesis and the priors[4] encoded in the generative model. Thus, constituting the best hypothesis or explanation means counting as such in light of the strength of the inferential connections to the antecedent model.[5]

As a toy example, take a simple, non-technical PP-based attempt to explain binocular rivalry (Hohwy, Roepstorff, Friston 2008). Two different pictures (say, a face and a house) are shown to each eye of the subject. Instead of subjectively perceiving a fusion of a face and a house (a face-house hypothesis), subjects report alternating between experiencing a face and a house.

transitions, enabling them to reliably match the transitions that exact Bayesian inference would produce. Furthermore, by virtue of operating on a Bayesian network, the processes underlying perception exhibit systematicity and productivity (for extensive discussions, see Clark 2013a; Gładziejewski, forthcoming-a; Kiefer 2017, 2019), features often taken as hallmarks of inference. In sum, it is reasonable to claim that in PP, Bayesian inference is approximated with a genuinely inferential process.

[4] In the remainder of this paper, I use the term 'prior' to denote both the prior distribution and the prior likelihood distribution.

[5] Kiefer (2017) proposes a different way of capturing the role of coherence in perception on the PP view. He focuses on how perception, in PP as well as computationally related schemes, involves a more global representational shift. This shift involves not only the formation of a perceptual hypothesis, but also a global change in a whole model comprised of inferentially linked hypotheses, such that the coherence of this model is maintained against sensory perturbations. I take Kiefer’s proposal to be compatible with the present approach.

PP's explanation of this effect points to the inferential links between the three alternative hypotheses and the prior assumptions (combined with other considerations that aim to explain why the visual system never eventually decides on a single stable interpretation of the input). Although the face-house hypothesis fits the sensory evidence best (its likelihood is high), the visual system is described as assigning significantly lower prior unconditional probability to encountering face-houses as compared to either faces or houses. Binocular rivalry, on this proposal, stems from the brain's attempt to keep its perceptual hypotheses coherent with antecedent assumptions about what sorts of objects are most likely to cause sensory states.

I propose that the notion that perception is inferential in this way – that it relies on antecedent mental states to produce a perceptual state, in a way that accords to a normative rule of inference – is epistemologically relevant (see also Clark 2018; Ghijsen 2018; Munton 2018; Vance 2015). My overall point can be captured in four interrelated claims.

First, in virtue of their rational etiology, perceptual states count as bearers of epistemic value (for a seminal defense of the view that the mental processes underlying the formation of perceptual states render those states epistemically appraisable, see Siegel 2017). Perceptual states are evaluable as justified (or unjustified) in light of their inferential connections to prior assumptions on which they are based. Of course, those prior assumptions, encoded in the generative model, may not correspond to personal-level beliefs, and the perceptual inferences are not consciously accessible for the subject nor under the voluntary control of the subject. Despite those facts, I take perceptual inferences in PP to be truth-preserving transitions between representational states (see also Kiefer 2017). These inferences can endow perceptual states with epistemic justification because they are similar enough to other unconscious, nonvoluntary inferences that we take for granted as underlying justification transfer (Ghijsen 2018; Siegel 2017; see also Carter, Rupert 2020).[6]

Second, perceptual states' epistemic standing *originates* from their inferential relations to antecedent states instead of being merely *modulated* by them. On the latter, modulatory

[6] Some readers may still raise concern that I am too quick to count subpersonal priors as epistemically evaluable and capable of featuring in rational processes. Perhaps genuine epistemic evaluability resides exclusively at the personal level. Addressing this worry fully requires a paper of its own (see Carter, Rupert 2020 for an excellent discussion). But let me sketch out what I take to be a plausible line of argument. The question we should ask is: What features do personal-level states (and processes operating on them) have, such that having such features renders those states/processes epistemically evaluable *and* distinguishes them from subpersonal states/processes. Some of the plausible candidates involve conscious accessibility or the fact that subjects enjoy voluntary control over their personal-level states. However, as stated in the main text, it seems that epistemically evaluable personal-level states/processes sometimes – perhaps often – run unconsciously and beyond the voluntary control of the subject (see Jenkin 2020; Siegel 2017). Arguably, then, the fact that subpersonal priors and inferences over them are unconscious and not voluntarily controlled does not automatically exclude them from the domain of epistemic evaluability. Another plausible view may posit that personal-level states are epistemically evaluable in virtue of being revisable in light of evidence or reasons. But as will transpire in Section 5, on the PP approach, subpersonal priors are also rationally adjustable in light of contrary (sensory) evidence. Yet another option would be to claim that what makes personal-level states special with respect to having epistemic import is that they are involved in action guidance. However, when combined with Active Inference, PP views action as a sort of prediction error minimization, which is guided (more or less directly) by subpersonal policy priors (for philosophically-oriented treatments of action and decision making in PP/AI, see Smith, Ramstead, Kiefer, preprint; Tate 2019). I hope that these remarks show that plausible candidates for features that endow mental states with epistemic status either turn out not to be necessary for having epistemic status (conscious accessibility, voluntary control) or are in fact shared between personal-level states and subpersonal priors (reason responsiveness, action guidance). Hence, treating subpersonal priors as bearers of epistemic import is not as hopeless as it may seem at first – in fact, I think that the burden of proof is on the proponents of the epistemic-evaluability-on-personal-level-only view. I thank an anonymous reviewer for pressing me on this point.

approach, perceptual states are said to enjoy some degree of justification regardless of their inferential connections to other mental states, and only then can inferences either downgrade or upgrade this original standing (see Siegel 2017). On the PP view, however, inferences are inherently involved in forming perceptions. If the constitutive function of perception is to represent the mind-independent world rather than merely register physical energies impinging on the sensory apparatus (Burge 2010), then, in PP, this job is done by inferring the causes of sensory states. Relatedly, if perception solves the underdetermination problem mentioned earlier, then, on the PP view, the problem is solved through inference. Because of how pervasive the role of inference in perception is, according to PP, it is reasonable to see it as a source of epistemic justification rather than a modifier of non-inferential justification (see also Ghijsen 2018; Jenkin 2020; Vance 2015).

Third, because perceptual inferences comply with a rational rule of inference, the epistemic status of resulting perceptual states is *positive*. Perceptual states are justified in light of the antecedent assumptions (the generative model) on which they are based.[7] In the next subsection, I will add a crucial caveat to this claim.

Fourth, given that perceptual states count as bearers of a positive justificational status, they can, in principle, transfer this status to other mental states, like beliefs. This effectively answers

[7] This point invites a question about whether PP is compatible – and if so, then how – with the idea that perceptual inference sometimes *downgrades* the epistemic standing of perceptual states (Siegel 2017). An interesting route to addressing this issue lies in PP-based models of psychopathologies, which explain perceptual disturbances (like hallucinations in schizophrenia) in terms of disturbances of Bayesian updating (see Adams, Stephan, Brown et al. 2013; Fletcher, Frith 2009). But can neurotypical Bayesian brains also undergo perceptual downgrade? This is a subject that requires a separate treatment, so I have to set it aside here (for an interesting take, see Clark 2018).

the question of how perceptual states acquire their ability to confer justification on other mental states.[8]

3.2. In predictive processing, perceptual justification is conditional

The inferentialist/coherentist view just depicted puts, or so it seems, a substantial limitation on the rational powers of perception. We, or at least those of us who subscribe to PP, need to forgo the traditional empiricist hope that perception could directly acquaint the perceiver with a set of pristine, non-inferentially justified propositions. Instead, perceptual justification is conditional (Gupta 2006). Formally, this idea has already been implicitly conveyed in the equation for variational free energy (VFE) introduced in section 2. Inferring a posterior involves searching the space of hypotheses in the generative model for one that (best) minimizes the VFE. This effectively means comparing one representation, that is, the candidate perceptual estimate, to another representation, that is, the model of the causal structure of the environment (see also Constant, Clark, Friston 2020; Kiefer, Hohwy 2017). In other words, inferring the best causal explanation for the current sensory input means inferring the best explanation *relative to the generative model brought to bear on interpreting this input.*

[8] The crucial assumption here is, of course, that perceptual states are capable of engaging in inferential (justification-transmitting) relations with those other mental states. Here, I will not discuss at length how this can be achieved. But suffice it to say that in the hierarchical scheme that PP postulates, estimates at higher levels of the generative model minimize the prediction error with respect to lower-level hypotheses situated closer to the sensory periphery. Thus, the justification transfer from lower (perceptual) to higher (cognitive) levels is approximately Bayesian (see Brössel 2017 for an account of how justification transfer from perception to belief could work in a broadly Bayesian cognitive system).

Philosophically, the upshot is that the rational standing of perceptual states is conditional on the rational standing of the generative model. Perceptual states are justified conditional on the justification of the prior assumptions on which they are based.

For illustration, consider two perceivers, A and B, who harbor generative models which differ significantly with respect to priors that enable the extraction of shape information from information regarding illumination. In particular, A's model ascribes the highest prior probability to there being a single source of illumination located above the perceiver, and B's model ascribes the highest prior probability to there being a single source of illumination placed below the perceiver. A and B are shown an image of a round object illuminated towards its upper side and gradually shaded towards its bottom part. Given some further assumptions that I leave out here for simplicity, upon being presented such image, A perceives as of a convex shape, and B perceives as of a concave shape (if we permit chickens to count as perceivers in this scenario, then Hersherberger 1970 describes an attempt to induce it experimentally). Both percepts result from a rational process and are thus conditionally justified. Their justification is conditional on the justification of the prior assumptions regarding, among other things, the most likely location of the light source. Note, however, that based on their perceptual states alone, the two perceivers are incapable of finding a neutral common ground to decide which perception (or the perceptual belief based on it) is justified in some absolute, unconditional sense.

I think that these considerations rule out the possibility of reconciling the PP's account of perceptual justification with pure or non-hybrid foundationalism that relies on unjustified justifiers. Consider an intriguing attempt, due to Ghijsen (2018), to turn PP's view of perception foundationalist after all. Although Ghijsen agrees that perceptual states owe their justification to their inferential connections with priors, he claims that it is priors themselves that terminate the chain of inferential justification. On his proposal, the priors obtain their ability to affect perceptual states' epistemic status not by being inferred from other states but by virtue of the reliable causal chains (leading to the world itself) through which they are acquired.

I do not intend to deny that priors are reliably acquired, but I doubt that saying that they are acquired *non-inferentially* gives justice to PP. One of PP's interesting tenets is its use of 'empirical Bayes' to propose that a large chunk of the prior knowledge is bootstrapped from raw sensory data during development (see Clark 2013a, 2013b; Hohwy 2020). Given an initial generative model (which may even start with parameters set at random), posteriors arrived at through perceptual inference can incrementally shape priors for future iterations of inference. The rate of revision, initially rapid, can slow down as the priors are increasingly more grounded in learning history. The point is that acquiring and adjusting priors, just like perceptual inference, is a matter of minimizing the prediction error (see Bogacz 2017; Friston 2005). The difference lies in the timescale: perceptual inference involves finding the approximate posterior, given the generative model, while learning or adjusting the priors involves gradually optimizing the parameters of the generative model over multiple instances of perceptual inference. I will revisit the idea that the generative model is (at least in part) inferred from sensory data in section 5. For now, suffice it to say that priors are inferentially acquired and adjusted.[9]

This unrelenting inferentialism still applies (contra Ghijsen 2018, p. 16) if we decide to bracket out the history of learning and restrict the source of justification of perceptual states to priors that are operative in a single episode of perceptual inference. In realistic scenarios, perceiving involves unmixing complex sensory input to reveal multiple interacting hidden causes that jointly comprise a dynamic scene (say, a rooster chasing a cat). The flow of the sensory signal depends, in part, on how those causes interact with each other. In the generative

[9] PP leaves room for the possibility that at least some of the priors used in perception may be innate. But note that PP view allows that such innate priors can be *maintained* or *adjusted* through iterations of perceptual inference, so that at least their maintenance/adjustment is inferential (I will come back to this in Section 5).

model, these sorts of interactions are encoded as conditional dependencies between variables encoding the hidden states. Thus, perception of a dynamic scene at time t is inferred from a representation of an immediately preceding world state at t_{-1}, adjusting, of course, for the prediction error coming from the lower level(s). As a result, the choice of which priors are brought to bear on the interpretation of the current sensory input at least partially depends on their inferential connections *to other priors* (arguably, in dreams and imagery, these types of dependencies are the only drivers of state transitions within the generative model; see Hobson, Hong, Friston 2014; Williams 2020). Perceptual justification, in PP, is inferential all the way.

4. Sensory states and the input problem

In this section, I introduce a (quasi-)foundationalist element to the account to show how it can solve the input problem mentioned in the introduction. In 4.1, I posit that the statistical structure of the sensory signal, when appropriately connected to the inferential machinery of the generative model, allows perceptual states to be produced in a truth-conductive manner. In 4.2, I clarify the epistemological commitments of my proposal by discussing a PP-inspired variant of the new evil demon scenario.

4.1. Sensory receptivity solves the input problem

Let me now turn to the second issue regarding perceptual justification mentioned in the introduction – the input problem. How could inferentially derived percepts remain responsive, in some epistemologically relevant sense, to a mind-independent world? The idea that perceptual states are selected so as to remain coherent with the preexisting model does not, by itself, establish how they can also be truth-conductive. In Kantian terms, the story so far focused on the ‘spontaneous,’ constructive aspect of perception. Here, I will discuss a ‘receptive’ aspect of perception that puts some external constraint on how perceptual states are produced, enabling perception to reveal reality.

In PP, this receptive side of perception is found, I want to claim, at the very bottom of the inferential hierarchy, that is, at the level of streams of raw data produced by the world at the organism’s sensory boundary. I propose that sensory states are not representational and, as such, cannot (hence, do not) engage in inferential relations with representational states. Nonetheless, they still play a quasi-foundational *evidential* role in perception in a way that effectively solves the input problem. Now, I will unpack this idea.

By saying that sensory states are not representations, I mean that they are not the sort of states to which we could justifiably attribute content or accuracy conditions. They are mere registrations of physical energies affecting the organism rather than representations of the distal causes of the stimulation (Burge 2010). Take the visual modality, where the sensory states consist in activations of photosensitive cells in the retina. These states do not have the prototypical features of representations. For example, they seem incapable of misrepresenting anything. Even though retinal registrations can be noisy (say, on a foggy day) and thus mislead perceptual inference, it would be presumably mistaken to claim that the *retina itself* misrepresents. The representational error lies in perceptual estimates of the environment rather than in the sensory input itself. Furthermore, sensory states do not serve the roles of representations in the cognitive system. For example, photoreceptor activations do not allow off-line or stimulus-free processing that is often associated with representations. They react to the world, but do not stan-in for (parts of) it, like representations do. Of course, one might note that sensory states still meet the conditions of serving as receptors, in that their biological proper function is to reliably react to states of the environment. But it has been forcefully argued that serving as a receptor in a cognitive system is not sufficient for serving as a representation

(Ramsey 2007). Although sensory states are (as I clarify below) rich in information precisely in virtue of acting as receptors, this information is best construed in terms of raw streams of bits – to be interpreted in terms of their distal causes by the generative model – rather than as contentful states or states with accuracy conditions.

Importantly, sensory states exhibit receptivity, understood in a broadly Kantian sense as a passive capacity to be affected by things. Because they are not representational, they are pure or theory-neutral with respect to the generative model. Sensory states depend on what the worldly causes *are*, not on what they are *represented* or *inferred* to be. Mathematically, the dependence of sensory states on their worldly causes can be captured in the following equation (Friston 2005; Wiese 2017):

$$s = g(c) + \omega$$

Here, s denotes sensory states, rendering them as a function of their worldly causes, expressed as $g(c)$, plus random noise, denoted ω. The functional dependence can also be described as a 'generative process,' where the 'true' causal structure of the world produces sensory states (Smith, Friston, Whyte, preprint).[10]

[10] As pointed out to me by an anonymous reviewer, the claim regarding the receptivity of the senses requires certain qualifications. First, there is a sense in which the sensory signal does depend on the generative model. Simply, how a person samples her environment – where she directs her eyeballs or how she explores the immediate surroundings with touch receptors – is often driven by the model-derived estimates of the causes of sensory states (see e.g. Friston, Adams, Perrinet, Brakspear (2012) for an Active Inference model of how people visually sample an image with saccadic movements under an assumption that the object depicted is a face; see also Lupyan 2017). However, this sort of dependence of the sensory signal on the generative model does not endanger the claim that the latter is pure or passive in the relevant sense. After all, once the perceiver has decided to actively sample the environment in a particular way, it is no longer up to the her what signal will be received by her sensorium – it

From this perspective, we can construe the brain or the perceptual system as attempting to predict the flow of sensory states, s, by reconstructing, to some biologically affordable level of approximation, $g(c)$ in an internal generative model (accounting, through estimations of precision, for context-dependent degrees of noise). This is achieved through repeated trial-and-error in which the generative model is gradually optimized with respect to its ability to predict s, as measured by the long-term, average prediction error. Note that it is ultimately s itself that serves as a 'tribunal' against which the generative model is tested. In this sense, the sensory states serve as an evidential basis for the generative model. This procedure is truth-conductive – with some restrictions to be discussed in the following subsection – because the model's ability to predict s depends on the degree to which it recapitulates the causal structure (the generative process) that produces s.[11]

Another way to spell out this idea is by saying that sensory states embody (but not: represent) statistical patterns or correlations[12] that systematically depend on the causal structure

depends on the external causes themselves. Second, it may be said that certain 'predictions' about the environment are already present at the level of sensory systems, simply in virtue of the fact that those systems embody certain feedforward biases. For example, the on/off receptive fields of ganglion cells in the retina can be said to 'predict' that local patches of natural images are uniform in light intensity, such that the firing a ganglion cells encodes a prediction error with respect to this prediction (Srinivasan, Laughlin, Dubs 1982). However, I think that these sorts of feedforward biases are best seen as non-representationally attuned to environment statistics, rather than as representations (Gładziejewski, forthcoming-a). Hence, I do not think it is fair to conclude that the senses are theory-laden in an epistemically relevant sense or that they store 'unjustified justifiers'.

[11] The somewhat vague term 'recapitulation' can be clarified by appeal to the notion of structural similarity between the relational organization of the generative model and the causal structure generating the sensory signal (Gładziejewski 2015; Kiefer, Hohwy, 2017).

[12] The existence sensory patterns is mathematically grounded in their compressibility, that is, in the possibility of expressing those patterns using fewer bits of information (Dennett, 1991). In PP, the generative model can be seen as a sparse (compressed) encoding of the sensory patterns.

that produces them. PP postulates that the brain uses unsupervised learning algorithm(s) to extract information about the causal structure producing the input from those very statistical patterns. The loss function in perceptual learning (and inference) is specified by the average prediction error, which is ultimately determined by the difference between internal predictions and the patterns in the input itself. This way, again, it is relative to the raw sensory patterns that the generative model is evaluated and corrected.

To illustrate this with an example, consider object perception under PP. Let us start with an assumption that perceiving objects involves a binding process, whereby an unorganized set of features (e.g., shapes, colors, textures, movement) is integrated to yield a representation of a unified object that has those features (Treisman 1996). In PP, as well as in Bayesian approaches more generally, this binding process is modeled as causal inference (c.f. Hohwy 2013; Parise, Spence, Ernst, 2012; Shams, Beierholm, 2010; Wiese 2018). What we experience as objects are entities inferred as common causes that produce and sustain sensory patterns.

Take, then, a complex visual sensory pattern comprised of: (1) a sequence of collinear edges that close in to form a shape that is retained over time, (2) a uniform color patch, (3) the fact that the initial position and the motion of the color patch overlaps with the initial position and the motion the sequence of edges.[13] Given that a stable pattern such as this is unlikely to emerge randomly, it may be inferred as being generated and sustained by a single external object in the environment – a common cause underlying the pattern. From the common-cause explanation, further predictions can be derived about how the pattern should evolve: (1) manipulating the

[13] In reality, moving 'edges' or 'color patches' correspond not to sensory states as such (here, retinal activations) but to perceptual estimates formed at low level(s) of the generative model. Edge-talk and color-talk I employ here is a shorthand for talking about actual sensory patterns (for example, 'edges' may correspond to sequences of activations of on/off retinal ganglion cells, while 'colors' may correspond to patterns of activations of red-, green-, and blue-sensitive cones).

object (the purported common cause) should result in a correlated change of elements of the pattern; (2) the elements of the pattern should turn out mutually statistically independent, conditioned on the common cause (e.g., given the position of the object, the position of shape gives no additional information about the position of the color patch). These predictions can be put to the test through active inference, that is, by acting on the external causal structure to induce a stream of sensory data and computing the prediction error. Crucially, the sensory signal's statistical structure plays an *evidential* role here by either conforming to the common-cause perceptual estimate or by disconfirming it.

Admittedly, treating non-representational things (like sensory states on the present view) as evidence may strike some readers as implausible.[14] However, I think that on a closer look, this position is more natural than it initially appears. Behind my proposal is the idea that 'evidence' is a functional notion. Consider the evidential value of registrations made by scientific apparatuses, like the fMRI machinery reacting to electromagnetic waves caused by the changing spins of protons in hydrogen atoms that comprise water molecules present in the brain, or a telescope registering streams of radio waves generated by astronomical sources. Considered separately, these registrations are just events in worldly causal chains. However, they can become *evidence for* someone if they participate in a larger cognitive or epistemic economy. For a human equipped with a theory or an interpretative scheme, the existence of these registrations can *serve as evidence* with respect to hypotheses about the workings of the brain or the existence of distant supernovae (these hypotheses can be also thought of as candidate abductive explanations of registrations). The registrations constitute evidence in virtue of their use or functional role. Serving this role in a larger system does not require them to be representational: it is the explanatory hypotheses that represent the world, not the registrations themselves. Crucially, I think that a non-representational registration can play this

[14] I thank two anonymous reviewers for urging me to address this issue.

sort of evidential role in a larger economy that does *not* involve a full-blown human being as the interpreter. In PP, the patterns received by the senses allow the generative model to infer their external causes. No homunculus is required to interpret the registration, but the functional role of the registration is similar enough to the scientific case to still count as evidential.

To summarize, two features underly the epistemic role of the sensory states in PP. First, sensory states are non-representational, passive registrations of physical energies impinging on the sensory apparatus. In virtue of being such registrations, they embody statistical patterns which are generated by the external causal structure, thus encoding (non-semantic) information about the latter. Second, although the sensory states do not furnish the mind with epistemically basic representations, their epistemic role in perception is still recognizably foundational in spirit: they serve as a theory-neutral tribunal for generative models (and perceptual estimates drawn from them). Of course, their ability to play this role is necessarily dependent on them being appropriately connected to the larger Bayesian machinery. Without the generative model, sensory states are causal registrations with no epistemological significance. With the generative model in place, they become a source of sensory evidence, making the model epistemically responsive to the world. This is how PP solves the input problem.

4.2. Predictive processing, the nature of epistemic justification, and the new evil demon

To further clarify my proposal, let me set it against a backdrop of a wider discussion regarding the very nature of epistemic justification. One might argue that by invoking the sensory signal to address the input problem, I thereby introduce an externalist or reliabilist thread (see Goldman 2008; Lyons 2009) to a story that has thus far relied on a broadly internalist notion of epistemic justification.[15] It is easy to interpret my proposal as simply pointing to sensory states as intermediaries that establish a reliable causal chain connecting internal models and perceptual estimates to the states of the environment. However, the view on offer here is more subtle and remains in line with the internalist view of epistemic justification even after the sensory states enter the picture. The idea is that the sensory input's role lies not simply in truth-conductive causal mediation but in the evidential or support relation holding between the *statistical patterns that arise at the sensorium* and the generative models used to predict those patterns.

To see this, consider the following variant of the 'new evil demon' scenario (Cohen 1984). Imagine you have an epistemic twin. Throughout her life, she undergoes the same series of mental states and processes as you. She is indistinguishable from you in terms of epistemically relevant processes, like reasoning, memory-based belief formation, all the way to perception. If perception works in accordance to PP, this means that your twin receives a stream of sensory input identical to the one you receive, performs Bayesian perceptual inferences based on this input, forming the same perceptual hypotheses as you, and over time learns a generative model of her (purported) causal milieu that is exactly like the model that your brain harbors. The difference between you and your epistemic twin consists in the twin being systematically misled by an evil demon who intentionally conceals the truth from her. There is thus no reliable causal chain that connects, *via* the senses, her perceptual states to their external causes.

Because the streams of sensory data the demon feeds to your twin embody the same statistical patterns, and because she updates her perceptual states and, over time, her generative model in a Bayes-rational manner just like you, the twin's perceptual states are rationally

[15] The account so far relied on an internalist view of epistemic justification at least in the sense that it postulated that features internal to the cognitive life of the perceiver, such as the generative model, determine the epistemic standing of perceptual states (see Conee, Feldman 2001).

acquired, just like yours. Given all available sensory evidence, your twin's perceptions are as epistemically justified as yours, and so their rational involvement in her cognitive life matches the rational involvement of your perceptions. Focusing on the sensory states: whatever evidential value the sensory patterns have in your case, they have the same evidential value in the case of your epistemic twin. Thus, sensory states can bear evidential value even if the perceptions and internal models based on them are systematically off-track with respect to reality. Truth-conduciveness or reliability is not a necessary condition on the sensory states' having evidential value.

Importantly, the point here is not to altogether detach perceptual justification from truth. A much more reasonable conclusion to draw from the PP-inspired new evil demon case is to say that it puts a limit on the degree to which perceptual justification can be a guide to truth.

Consider a class of 'normal' scenarios, where a perceiver inhabits a world construed as a complex causal structure, such that the patterns that arise at her sensorium are directly sampled from this structure. By definition, normal scenarios involve no demonic mischief. In normal scenarios, the perceiver should, through long-term prediction error minimization, form a generative model that recapitulates (to some relevant degree of approximation) the external causal structure, whatever it may be. As a result, the perceiver should produce largely accurate perceptual estimates. As long as one finds oneself in a normal scenario, perception is truth-conductive on the PP view. However, normal scenarios do not exhaust the range of possible causal structures underlying sensory input. In particular, there may be 'devious' causation of the input, where the sensory patterns are intentionally produced so as to deceive the perceiver regarding the actual causes of her input. My claim, then, is that perception allows perceivers to rationally select, in a truth-conductive way, a model that best explains the patterns in their sensory input under an assumption that a normal scenario holds. However, perception is illsuited to allow perceivers to establish whether the latter assumption is true, that is, whether the causal structure producing the input is devious or normal.

So perceptual justification *is* a guide to truth, albeit within limits. These limits presumably overlap with the range of truths that can be accessed – in an information-processing sense, which may not involve *conscious* access – from the 'animal's point of view' (Eliasmith 2005), that is, given only the internal model and the sensory states. This conclusion should not be seen as surprising, as it simply restates a claim that borders on a truism: one is unable to refute skepticism by appealing to the deliverances of the senses.

5. Unconditional perceptual justification through epistemic convergence

In section 3.2, I stated that in PP, perceptual states are justified conditionally, as their justification depends on the justificational status of the generative model on which they are based (see also Gupta 2006). This implies that if the model itself lacks appropriate rational standing or justification, then the justification of resulting perceptual states is significantly undermined. The rational powers of perception turn out to be seriously limited. Is there a way for perceptual justification to go beyond this limit? Can the view defended in this paper allow perceivers to at least sometimes undergo perceptual states whose justification is, in some sense, absolute or unconditional? If what we perceive is molded by our antecedent models, how could two subjects with significantly different models ever rationally settle their disputes by appealing to a common perceptual ground?

The crucial jigsaw for solving this issue lies, I want to argue, in perceptual learning. Although in PP, perception relies on priors encoded in the generative model, it is by no means destined to dogmatically stick to a given set of priors in a self-confirming inferential loop. Instead, priors themselves are open to revision in light of incoming sensory evidence. A substantial body of computational modeling work on hierarchical Bayesian models has established that overpriors – abstract priors stored at high levels of the generative whose job is to constrain the space of hypotheses available at the lower levels – can be induced, in an unsupervised way, from raw sensory data (c.f. Tenenbaum, Kemp, Griffiths, Goodman, 2011; for extensive philosophical discussions of Bayesian accounts of learning, see also Clark 2013a; Colombo 2018). Within the conceptual framework of PP, learning priors consist of tuning the generative model parameters through gradient descent on long-term, average prediction error (see Friston 2005; Bogacz 2017). Hence, the idea of rationality of perception applies to learning priors just as well as it applies to inferring posteriors, given priors.

Importantly, for the overall claim about the revisability of priors to hold, no commitment is required to a strong view that (all) priors are bootstrapped from blank states. Even under the assumption that at least some priors are innate or unlearned in some sense, we should still expect those priors to be open to adjustment in light of contrary sensory evidence, and we should expect them to be maintained only under the condition that they survive empirical testing. So even if some priors do not originate from a rational process, they can at least be rationally adjusted or maintained in light of incoming sensory evidence, such that the very fact that a subject sustains those priors is explained by a rational process that is answerable to sensory evidence.

There is substantial empirical support for the notion that perceptual priors are malleable in light of sensory evidence. To mention a couple of examples: (1) the light-from-above prior that underlies the extraction of shape from shading (see the example from section 3.2) can be modified through exposing subjects to visual stimuli paired with haptic feedback (Adams, Graf, Ernst 2004); (2) the prior expectation that perceived objects are stationary or move at slow speeds (Weiss, Simoncelli, Adelson 2002) can be modified by exposing subjects, repeatedly over a couple training days, to a stimulus containing fast-moving lines (Sotiropoulos, Seitz, Seriès 2011); (3) the prior that relates the perceived weight of an object to its size can be adjusted through repeated training consisting of lifting blocks whose weight is inversely correlated with their volume (Flanagan, Bittner, Johansson 2008).

How do all these considerations allow us to address the question of unconditional perceptual justification? To see this, we may need to suspend intuitions inherited from traditional foundationalist empiricism (if we have such). Instead of treating unconditionally justified percepts as starting points in a chain of justification, perhaps we should treat them as arising *at the limit* of an extended rational learning process. The crux of this approach has been famously captured by Peirce:

> Different minds may set out with the most antagonistic views, but the progress of investigation carries them by a force outside of themselves to one and the same conclusion. This activity of thought by which we are carried, not where we wish, but to a fore-ordained goal, is like the operation of destiny. No modification of the point of view taken, no selection of other facts for study, no natural bent of mind even, can enable a man to escape the predestinate opinion. This great law is embodied in the conception of truth and reality. (Peirce 1878/2011, p. 63)

The idea, then, is that a prolonged rational inquiry, where the subject gradually modifies her beliefs in light of incoming evidence, should eventually converge on a certain set of beliefs. Once learned, these beliefs can no longer be revised by future evidence. And when exposed to a common pool of evidence (e.g., series of observations), distinct rational learners should eventually converge on those beliefs regardless of the differences in their starting assumptions. On this broadly Peircean picture, then, unconditionally justified beliefs are beliefs that constitute the endpoints of epistemic convergence (for an ingenious treatment of epistemic convergence to which the present proposal owes a lot, see Gupta 2006). Importantly, the very idea of epistemic convergence is consistent with Bayesian statistics and Bayesian epistemology (for useful discussions, see Hawthorne 2018; Hutteger 2015). Very roughly, Bayesian learners that update their priors in light (common) evidence are expected to eventually converge on a common set of priors, such that any initial differences between them vanish in the long run.

We may now see how the notion of epistemic convergence can find a local application in theorizing about perception. Regardless of whether inquiry at large epistemically converges to some ultimate endpoint, it is not unreasonable to claim that the *perceptual systems* – as long as they work in accordance to PP – undergo at least partial convergence over developmental timescales. That is, at least some perceptual priors may be such that: (1) over individual development, once those priors are learned, they remain stable in light of future sensory evidence, under the assumption that the relevant environmental statistics remain stable (alternatively: the priors in question are innate but adjustable during development, such that they are either stable in light of sensory evidence, or gradually adjusted during development until they reach a point where they remain stable in light of new sensory evidence); (2) distinct subjects, even if they initially differ, eventually converge on those priors as long as they sample sensory states from a common environment.[16]

Let us now revisit perceivers A and B from section 3.2, who differ with respect to the light-from-above prior. Upon registering an image of a round object illuminated towards its upper side and gradually shaded towards its bottom part, A forms a perception as of a convex shape,

[16] By ‘common environment,’ I mean here that distinct subjects inhabit learning environments that are identical or closely similar with respect to statistics that drive the learning or adjustment of the prior in question (in the case of the light-from-above prior, these would be environments which are the same or closely similar with respect to the average number of light sources and their average position(s) relative to the perceiver).

and B forms a perception as of a concave shape. As already stated, both percepts are conditionally justified, that is, justified relative to the generative models that A and B harbor, respectively. However, with the notion of epistemic convergence in hand, we may see how one of them may be *unconditionally* justified. Assuming that A and B inhabit a common world like ours, we may suppose that A's prior expectation of a singular source of light placed above is epistemically convergent (unless experimentally tinkered with on purpose, as in: Adams, Graf, Ernst 2004). It is A, and not B, that is unconditionally justified (or: just right) in her perception of the object as convex. Furthermore, under the assumption that A's light-from-above prior is epistemically convergent, B should be able to adjust her prior when continually exposed to sensory evidence until she eventually agrees with A.[17] Consequently, perception is never epistemically trapped in an antecedent model but is epistemically responsive, over time, to the way things are.

Of course, my overall point here may generalize to other priors. These may include, for example, overpriors that structure the way humans, from about the age of eighteen to twenty-four months onwards (Piaget 1954), parse their perceptual world into ordinary objects like cats and chairs (Gładziejewski, forthcoming-b).

To summarize this, I propose that perceptual states can be unconditionally justified to the degree to which they are inferred from perceptual priors on which perceptual learning epistemically convergences. 'Unconditional' justification does not mean here that the

[17] The idea of B being able to update her prior over time is not mere speculation, as empirical evidence suggests that, in humans, the light-from-above prior undergoes calibration in early development, until it converges to a certain value, remaining stable during adulthood (Stone 2011; Thomas, Nardini, Mareschal 2010).

perceptual state is non-inferentially justified. Rather, the point is that it is inferred from prior assumptions that constitute endpoints of rational learning.[18]

6. Conclusion

Whether perception is a matter of prediction error minimization is far from settled, and doubts have been recently raised about the empirical merits of PP, as well as about whether the

[18] Two additional clarifications may be in order here (both invited by an anonymous reviewer, to whom I am grateful). First, it may seem too strong to say that a perceiver is unconditionally justified in her perception only under the condition that she bases her percept on a convergent prior(s). I think this issue could be fixed by allowing a graded notion of unconditional justification, where a subject is unconditionally justified in her perception *to the degree t*o which the priors she uses to infer her estimates match the values of convergent priors (that is, how 'close' her internal model is to the model on which rational learning converges). So, a perceiver could be more or less unconditionally justified in her perception depending on how close the values of her priors are to the values of convergent priors, and only reach full unconditional justification when relying on convergent priors. Of course, additional work would need to be done to fully elucidate this graded notion of unconditional justification. Second, the present proposal does not rule out the possibility that there are priors on which perceptual systems of distinct perceivers *fail* to ever converge. An anonymous reviewer points out that the persistent perceptual disagreement about "The Dress" image (which some people perceive as black and blue, and some as white and gold) may be due to distinct people relying on different and non-converging perceptual priors. However, although the present proposal certainly leaves open the possibility that priors sometimes fail to converge (and thus fail to generate unconditional justification), I take this to be a feature of the view on offer here, and not necessarily a bug. Furthermore, I suspect that genuine failure of convergence is not common enough to raise serious danger of skepticism about perceptual justification. In The Dress example, the lack of convergence may be caused by the fact that ambient lightning statistics differ for perceivers of different circadian types (Wallisch 2017). That is, somewhat surprisingly, the lack of convergence in this case may be explained by the fact that the 'common environment' condition is not met among all perceivers – some draw their priors from daylight conditions, and others adjust their priors to artificial ambient illumination.

theory can be derived from first principles, as some of its advocates maintain (Litwin, Miłkowski 2020; Williams, preprint). However, if perception is prediction error minimization (or is underpinned by some other, sufficiently similar type of Bayesian processing) [19], then, I argued, this fact would shed light on its rational involvement in cognition. On the proposed interpretation, perception's rational power lies neither in furnishing the mind with pristine representations whose justification is immediate or given, nor is it solely a matter of internal coherence among representational states. Instead, in PP, the architecture of perceptual justification is foundherentist. In line with coherentism, perceptual states have a (Bayes-)rational etiology and are constructed so as to cohere with a larger inferential structure of the generative model. This way, they obtain their ability to confer justification on other mental states. But, in keeping with the spirit of foundationalism, the inferential machinery of perception is constrained from outside by the patterns that arise at the sensory boundary, making perceptual states epistemically responsive to worldly states of affairs. Against the notion that an insurmountable gulf separates reasons from causes, here is a thoroughly causal story that allows perception to act as a provider of reasons.

Acknowledgments

I thank the two anonymous reviewers for their insightful comments on the paper. A previous version of the paper was discussed at the Philosophy of Cognitive Science seminar held at the Institute of Philosophy and Sociology at the Polish Academy of Sciences, and I am indebted to

[19] Although the focus in this paper is on PP, perhaps my proposal paper may also capture, at least partially, the epistemology of perception inherent to 'Bayesian brain' views more generally (c.f. Pouget, Beck, Ji Ma, Latham 2013; Rescorla 2021), as well as to control-theoretic accounts grounded in notions of forward models and Kalman filtering (see Grush 2004; Rao 1999).

the participants for their criticisms and suggestions. My work was supported by the National Science Center in Poland (grant no. 2019/33/B/HS1/00677).

Compliance with Ethical Standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Adams, R. W., Stephan, K. E., Brown, H. R., Frith, C. D., Friston, K. J. (2013). The computational anatomy of psychosis. *Frontiers in Psychiatry*, 4, 47.

Adams, W. J., Graf, E. W., Ernst, M. O. (2004). Experience can change the 'light-from-above' prior. *Nature Neuroscience*, 7(10), 1057–1058.

Badcock, P. B., Friston, K. J., Ramstead, M. J. D., Ploeger, A., Hohwy, J. (2019). The hierarchically mechanistic mind: an evolutionary systems theory of the human brain, cognition, and behavior. *Cognitive, Affective & Behavioral Neuroscience*, 19, 1319–1351.

Berger, J., Nanay, B., Quilty-Dunn, J. (2018). Unconscious perceptual justification. *Inquiry*, 61(5-6), 569–589.

BonJour, L. (1985). *The Structure of Empirical Knowledge*. Harvard: Harvard University Press.

Bogacz, R. (2017). A tutorial on the free-energy framework for modelling perception and learning. *Journal of Mathematical Psychology*, 76, Part B, 198–211.

Brössel, P. (2017). Rational relations between perception and belief: The case of color. *Review of Philosophy and Psychology*, 8, 721–741.

Bruineberg, J., Kiverstein, J., Rietveld, E. (2016). The anticipating brain is not a scientist: the free-energy principle from an ecological-enactive perspective. *Synthese*, 195, 2417–2444.

Burge, T. (2010). *Origins of Objectivity*. Oxford: Oxford University Press.

Carter, J. D., Rupert, R. (2020). Epistemic value in the subpersonal vale. *Synthese*. DOI: https://doi.org/10.1007/s11229-020-02631-1

Clark, A. (2013a). Expecting the world: Perception, prediction and the origins of human knowledge. *Journal of Philosophy, CX*, 469-496.

Clark, A. (2013b). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3), 181–204.

Clark, A. (2016). *Surfing Uncertainty. Prediction, Action, and the Embodied Mind.* Oxford: Oxford University Press.

Clark, A. (2018). Priors and prejudices: Comments on Susana Siegel's 'The Rationality of Perception'. *Res Philosophica*, 95(4), 741–750.

Cohen, S. (1984). Justification and truth. *Philosophical Studies*, 46: 279–295.

Colombo, M. (2018). Bayesian cognitive science, predictive brains, and the nativism debate. *Synthese*, 195, 4817–4838.

Conee, E., Feldman, R. (2001). Internalism defended. *American Philosophical Quarterly*, 38, 1-18.

Constant, A., Clark, A., Friston, K. J. (2020). Representation wars: Enacting an armistice through Active Inference. *Frontiers in Psychology*, 11, 598733.

Davidson, D. (1986). A coherence theory of truth and knowledge. In: E. Lepore (ed.). *Truth and Interpretation: Perspectives on the Philosophy of Donald Davidson* (pp. 307–319). Oxford: Blackwell.

Dennett, D. (1991). Real patterns. *Journal of Philosophy*, 88(1), 27–51.

Eliasmith, C. (2005). A new perspective on representational problems. *Journal of Cognitive Science*, 6, 97–123.

Flanagan, J., Bittner, J., Johansson, R. (2008). Experience can change distinct size-weight priors engaged in lifting objects and judging their weights. *Current Biology*, 22, 1742–1747.

Fletcher, P. C., & Frith, C. D. (2009). Perceiving is believing: A Bayesian approach to explaining the positive symptoms of schizophrenia. *Nature Reviews Neuroscience*, 10, 48‑58.

Friston, K. J. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B*, 360 (1456), 815–836.

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

Friston, K. J., Adams, R. A., Perrinet, A., Brakspear, M. (2012). Perceptions as hypotheses: saccades as experiments. *Frontiers in Psychology*, 3, 151.

Ghijsen, H. (2018). Predictive processing and foundationalism about perception. *Synthese*, doi: https://doi.org/10.1007/s11229-018-1715-x

Gładziejewski, P. (2015). Predictive coding and representationalism. *Synthese*, 193, 559–582.

Gładziejewski, P. (forthcoming-a). Predictive processing, implicit and explicit. In: R. Thompson (ed.). *The Routledge Handbook of Philosophy and Implicit Cognition*. Routledge.

Gładziejewski, P. (forthcoming-b). Un-debunking ordinary objects with the help of predictive processing. *British Journal for the Philosophy of Science*.

Goldman, A. (2008). Immediate justification and process reliabilism. In: Q. Smith (ed.). *Epistemology: New Essays* (pp. 63–82). Oxford University Press.

Gregory, R. L. (1980). Perceptions as hypotheses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 290(1038), 181–97.

Greshman, S. J. (2019). The generative adversarial brain. *Frontiers in Artificial Intelligence*, 2, Article 18.

Grush, R. (2004). The emulation theory of representation: motor control, imagery, and perception. *Behavioral and Brain Sciences*, 27, 377–96.

Gupta, A. (2006). *Empiricism and Experience*. Oxford: Oxford University Press.

Haack, S. (1993). *Evidence and Inquiry: Towards Reconstruction in Epistemology*. Oxford: Blackwell Publishers.

Hawthorne, J. (2018). Inductive logic. In: E. Zalta (ed.). *The Stanford Encyclopedia of Philosophy*. Retrieved 03-2-2021 from: https://plato.stanford.edu/entries/logic-inductive/

Helmholtz, H. (1855). Über das Sehen des Menschen (pp. 85–117). In *Vorträge und Reden von Hermann Helmholtz*. 5th ed. Vol.1. Braunschweig: F. Vieweg.

Hershberger, W. (1970). Attached-shadow orientation perceived as depth by chickens reared in an environment illuminated from below. *Journal of Comparative and Physiological Psychology*, 73(3), 407–411.

Hobson, J. A., Hong, C. C-H., Friston, K. J. (2014). Virtual reality and consciousness inference in dreaming. *Frontiers in Psychology*, 5, 1133.

Hohwy, J. (2013). *The Predictive Mind*. Oxford: Oxford University Press.

Hohwy, J. (2017). Priors in perception: Top-down modulation, Bayesian perceptual learning rate, and prediction error minimization. *Consciousness and Cognition*, 47, 75–85.

Hohwy, J. (2018). The predictive processing hypothesis. In A. Newen, L. Bruin, S. Gallagher (eds). *The Oxford Handbook of 4E Cognition* (129–146). Oxford: Oxford University Press.

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

Hohwy, J., Roepstorff, A., & Friston, K. J. (2008). Predictive coding explains binocular rivalry: an epistemological review. *Cognition*, 108: 687–701.

Huttegger, S.M. (2015). Bayesian convergence to the truth and the metaphysics of possible worlds. *Philosophy of Science*, 82(4), 587–601.

Jenkin, Z. (2020). The epistemic role of core cognition. *The Philosophical Review*, 129(2), 251–298.

Kiefer, A. (2017). Literal perceptual inference. In T. Metzinger, W. Wiese (eds). *Philosophy and Predictive Processing*. MIND Group. Available online at: https://predictive-mind.net/papers/literal-perceptual-inference.

Kiefer A. (2019). *A Defense of Pure Connectionism*. Unpublished dissertation. The Graduate Center, City University of New York. DOI: 10.13140/RG.2.2.18476.51842.

Kiefer, A., Hohwy, J. (2017). Content and misrepresentation in hierarchical generative models. *Synthese*, 195, 2387–2415.

Lupyan, G. (2017). Changing what you see by changing what you know: The role of attention. *Frontiers in Psychology*, 8, Article 553.

Lyons, J. (2009). *Perception and Basic Beliefs: Zombies, Modules and the Problem of the External World*. Oxford: Oxford University Press.

Maloney, L. T. Mamassian, P. (2009). Bayesian Decision Theory as a model of human visual perception: Testing Bayesian transfer. Visual Neuroscience, 26, 147–155.

Munton, J. (2018). The eye's mind: Perceptual process and epistemic norms. *Philosophical Perspectives*, 31(1), 317–347.

Parise, C. V., Spence, C., Ernst, M. O. (2012). When correlation implies causation in multisensory integration. *Current Biology*, *22*(1), 46–49.

Peirce, C. S. (1878/2011). How to make our ideas clear. In: R. B. Talisse, S. F. Aikin (eds.). *The Pragmatism Reader: From Peirce Through the Present* (pp. 50–65). Princeton (NJ): Princeton University Press.

Peirce, C. S. (1934). *Collected Papers of Charles Sanders Peirce, Volume V*. C. Hartshorne, P. Weiss (eds). Cambridge (MA): Harvard University Press.

Piaget, J. (1954). *The Construction of Reality in the Child*. New York: Basic Books.

Pouget, A., Beck, J. M., Ji Ma, W., Latham, P. E. (2013). Probabilistic brains: knowns and unknowns. *Nature Neuroscience*, 16, 1170–1178.

Rao, R. P. N. (1999). An optimal estimation approach to visual perception and learning. *Vision Research*, 39(11), 1963–1989.

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

Ramsey, W. (2007). *Representation Reconsidered*. Cambridge: Cambridge University Press.

Ramstead, M. J., Frston, K. J., Hipolito I. (2020). Is the free-energy principle a formal theory of semantics? From variational density dynamics to neural and phenotypic representations. *Entropy*, 22(8), 889.

Rescorla, M. (2021). Bayesian modeling of the mind: From norms to neurons. *WIREs Cognitive Science*, 12(1), e1540.

Sellars, W. (1956). Empiricism and the philosophy of mind. In: H. Feigl, N. Scrived (eds.). *Minnesota Studies in the Philosophy of Science* (pp. 253–329). Minneapolis, MN: University of Minnesota Press.

Shams, L., Beierholm, U. (2010). Causal inference in perception. *Trends in Cognitive Sciences*, 14(9), 425–432.

Siegel, S. (2017). *The Rationality of Perception*. Oxford: Oxford University Press.

Smith, R., Friston, K. J., Whyte, Ch. J. (preprint). A step-by-step tutorial on Active Inference and its application to empirical data. PsyArXiv: https://psyarxiv.com/b4jm6/

Smith, R., Ramstead, M., Kiefer, A. (preprint). Active inference models do not contradict folk psychology. PsyArXiv: https://psyarxiv.com/kr5xf/

Sotiropoulos, G., Seitz, A. R., Seriès, P.(2011). Changing expectations about speed alters perceived motion direction. *Current Biology*, 21, R883–R884.

Srinivasan, M. V., Laughlin, S. B., Dubs, A. (1982). Predictive coding: A fresh view of inhibition in the retina. *Proceedings of the Royal Society of London, B*, 216(1205), 427–459.

Stone, J. V. (2011). Footprints sticking out of the sand, Part 2: Children's Bayesian priors for shape and lighting direction. *Perception*, 40(2), 175–190.

Tate, A. J. M. (2019). A predictive processing theory of motivation. *Synthese*, https://doi.org/10.1007/s11229-019-02354-y

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., Goodman, N. D. (2011). How to grow a mind: Statistics, structure and abstraction. *Science*, 331, 1279–1285.

Thomas, R., Nardini, M., Mareschal, D. (2010). Interactions between "light-from-above" and convexity priors in visual development. *Journal of Vision*, 10(8), Article 6.

Treisman, A., (1996). The binding problem. *Current Opinion in Neurobiology*, 6(2), 171–178.

Vance, J. (2015). Cognitive penetration and the tribunal of experience. *Review of Philosophy and Psychology*, 6, 641–663.

Van Es, T. (2020). Living models or life modelled? On the use of models in the free energy principle. *Adaptive Behavior*, doi: https://doi.org/10.1177/1059712320918678

Wallisch, P. (2017). Illumination assumptions account for individual differences in the perceptual interpretation of a profoundly ambiguous stimulus in the color domain: "The dress". *Journal of Vision*, 17(4), 1–14.

Weiss, Y., Simoncelli, E. P., Adelson, E. H. (2002). Motion illusions as optimal percepts. *Nature Neuroscience*, 5, 598–604.

Wiese, W. (2017). What are the contents of representations in predictive processing? *Phenomenology and the Cognitive Sciences*, 16, 715–736.

Wiese, W. (2018). *Experienced Wholeness: Integrating Insights from Gestalt Theory, Cognitive Neuroscience, and Predictive Processing*. Cambridge, MA: The MIT Press.

Williams, D. (2020). Imaginative constraints and generative models. *Australasian Journal of Philosophy*, 99(1), doi: https://doi.org/10.1080/00048402.2020.1719523

Williams, D. (preprint). Is the brain an organ for prediction error minimization? http://philsci-archive.pitt.edu/id/eprint/18047