AVANT, Vol. XII, No. 1
ISSN: 2082-6710 avant.edu.pl/en
DOI: 10.26913/avant.2021.01.04

Understanding Predictive Processing. A Review

Michał Piekarski
Institute of Philosophy
Cardinal Stefan Wyszyński University in Warsaw
m.piekarski@uksw.edu.pl

Received 4 August 2020; accepted 31 August 2021; published 5 September 2021.

Abstract

The purpose of this paper is to provide a systematic review of the Predictive Processing framework (hereinafter PP) and to identify its basic theoretical difficulties. For this reason, it is, primarily, polemic-critical and, secondarily, historical. I discuss the main concepts, positions and research issues present within this framework (§1-2). Next, I present the Bayesian-brain thesis (§3) and the difficulty associated with it (§4). In §5, I compare the conservative and radical approach to PP and discuss the internalist nature of the generative model in the context of Markov blankets. The possibility of linking PP with the free energy principle (hereinafter FEP) and the homeostatic nature of predictive mechanisms is discussed in §6. This is followed by the presentation of PP's difficulties with solving the dark room problem and the exploration-exploitation trade-off (§7). I emphasize the need to integrate PP with other models and research frameworks within cognitive science. Thus, this review not only discusses PP, but also provides an assessment of the condition of this research framework in the light of the hopes placed on it by many researchers. The *Conclusions* section discuss further research challenges and the epistemological significance of PP.

Keywords: predictive processing; Bayesian brain; Bayesian inference; Bayesian models; prediction; prediction error; generative model; hierarchical inference; top-down processing; free energy principle; active inference; Markov blanket; perceptual inference; precision; perception; mechanisms; philosophy of mind; philosophy of cognitive science; epistemology

1. Introduction

There will be no exaggeration in claiming that the "fashion" for PP continues in the philosophy of mind and cognitive sciences. Papers and studies on this research framework and related issues are constantly being published.[1] There are at least a few reasons for PP's popularity. First, it is a research framework that promises to provide a naturalistic and biologically reliable explanation of complex interactions between perception, action, and environment. Second, if we are to believe the prominent representatives of this framework, PP can offer a unification of theories in cognitive science (cf. Clark, 2013; Hohwy, 2015; Seth, 2015). Third, PP offers to explain perception and actions using Bayesian modeling (which many researchers consider to be normative), which in practice means that the cognitive system is to implement (to some extent) the Bayesian belief network (cf. Harkness & Keshava, 2017; Hohwy, 2013). Fourth, many researchers (cf. Hohwy, 2014; 2015; Friston & Kiebel, 2009; Seth, 2015) believe that PP comes from a more basic, biological theory of life based on the so-called free energy principle (FEP). According to this principle, living organisms are systems that maintain their existence by minimizing the free energy of their internal states (cf. Friston, 2009; 2010; 2012; 2017; Friston, Stephan, 2007). Thanks to this, PP may constitute an important element of the general theory explaining life, mind and cognition (cf. Adams, Brown, Friston, 2014; Allen & Friston, 2018; Friston, 2013b; Friston, Fortier & Friedman, 2018).

According to PP, some organisms[2] entail, are embodied in or sometimes are a multi-level, hierarchical generative model of their environment, which stores or processes information in a cascading manner: it sends top-generated predictions whose aim is to minimize the so-called prediction error connected with bottom-up signals coming from sensory input. This error relates to the disproportion between expectations based on the internal parameters of the model and the variable information reaching the model through the senses. According to this view, an organism behaves effectively in its environment if it can efficiently minimize potential prediction errors. In the opinion of some PP supporters, this approach is supposed to explain many, if not all, phenomena related to cognition and what the philosophical tradition defines as the mind.[3]

I claim that we can now look at PP as a movement or research tradition.[4] Such a view, however, requires a certain amount of criticism and distance. Successive researchers refer more or less

[1] Google Scholar notes published 1,450,000 records containing the term PP as one of the keywords in the period from 2012 to 2020 (as of 24.08.2021).

[2] It is not entirely clear whether this is true for every living organism or only e.g., humans and primates. See Hohwy, 2020a, p. 220 on the problems with the use of PP for living organisms. Note that some authors claim that even self-organizing systems like plants must embody generative models (cf. Calvo & Friston, 2017).

[3] Clark (2013) labels the PP framework as *Grand Unified Theory* (GUT), making clear allusions to the physical *Grand Unified Theory*. In this sense, PP can offer an explanation of phenomena such as perception, attention, memory, motor control, actions, phenomenal experience, illusions, autism disorders and even consciousness (cf. Supplementary Table 1 in Hohwy, 2020a, for a thorough overview of topics in the field of PP with literature).

[4] Larry Laudan (1977) understood the research tradition as a set of general assumptions about entities and processes in a certain field of research and the correct methods that must be used to consider problems and construct

critically to PP analyses. The purpose of this review is to comprehensively discuss the basic concepts, positions and directions of research related to PP. It discusses the most important philosophical issues associated with PP and indicates their potential difficulties, without favoring any of the approaches or positions. Indeed, it can be treated as a charitable and critical approach to the PP framework, focused not so much on its rejection or weakening as on its improvement and development. The paper is directed primarily at philosophers of cognitive science, philosophers of mind, philosophers of perception as well as cognitive psychologists and those researchers who are interested in PP's philosophical aspirations.

So far, two philosophically oriented PP reviews have been published. The first is the work of Wanja Wiese and Thomas Metzinger entitled *Vanilla PP for Philosophers: A Primer on Predictive Processing* (2017). It provides an excellent take on the basic concepts of PP and a general introduction to what PP is. However, this work lacks critical dimension and does not discuss many important issues such as internalism in PP or the dark room problem. Another review was recently published by Jacob Hohwy (2020a). It discusses the main positions in PP and outlines the potential directions of development of this research framework. Hohwy describes PP in terms of his own and (in what I will later call a "conservative") reading of the framework, he proposes his own interpretation which is not without controversy (cf. Hohwy, 2013). His approach also lacks the critical aspect. This review does not want to repeat the work of these authors, but rather propose another, more critical, or rather complementary perspective.

2. General characteristics and basic concepts

At the outset, it should be noted that there is no single generally accepted approach to PP. First, there is also no single term that defines this framework.[5] In the literature, except PP, one can find such terms as "prediction error minimization" (Hohwy, 2013; 2020a), "predictive coding," which are most often used by neuroscientists (Rao & Ballard, 1999; cf. Spratling, 2017) and "active inference" (cf. Friston et al., 2017; Ramstead, Kirchoff & Friston, 2020). I decided not to use the first of these terms because it is closely related to the interpretation of PP proposed by Hohwy. It does not fit with the approach developed here, but it does owe a lot to Hohwy. The term "predictive coding" refers mainly to those coding strategies in which the predicted part of the input signal code is removed from the actual signal. In this way, it is only the difference between these signals that is passed on as output to the next stage of information processing.[6] Here the (Bayesian) brain can be cast as minimizing prediction errors. In such

theories. Research tradition, in contrast to theory, only defines the field of application belonging to the theory: it does not explain and does not predict phenomena. It seems that PP is currently at this stage of development and it is premature to refer to it as a scientific theory or research paradigm (on this topic see also Litwin & Miłkowski, 2020).

[5] In a very general sense, one can speak of PP as probabilistic inference modelling.

[6] Rao and Ballard (1999) proposed a visual information processing model in which information transmitted from the higher to lower layers of the cerebral cortex includes predictions about the activity of neurons present on its lower layers. The descending (top-down) connection carries information on the predictions about the neuronal

approaches, less attention is paid to the idea of a hierarchical and multilevel model generating predictions (Clark, 2016, p. 25-26; Kiefer & Hohwy, 2018, p. 9).[7] Importantly, predictive coding in and of itself has nothing to say about action. In PP, action is an important element of investigation and plays a non-trivial explanatory role. PP sometimes appeals to active inference which extends the principles of prediction to cover action, planning, decision-making and so on. Active inference is strictly related to the FEP (cf. Friston, 2009, see §6).

Many researchers associate the idea of PP with the work of the nineteenth century physicist, physician, and psychologist with neo-Kantian leanings Hermann von Helmholtz (1867).[8] He was the first thinker who explicitly stated that the brain is a hypothesis testing machine. Helmholtz says that the "mental activities" on which perception is based are more unconscious than conscious, because we only have direct access to events received by our nervous system, which means that we only feel the effects of external objects (Helmholtz, 1867, p. 430). Therefore, referring to Helmholtz, perception should be understood as unconscious inference (Hohwy, 2013, p. 18).[9] This approach, however, raises objections from many authors and leads to the formulation of various non-Helmholtzian views of PP (see § 5).

Some researchers claim (cf. Orlandi, 2018) that the currently developed PP approaches draw inspiration not only from neuroscience research, but also from image transmission research focused on television (cf. Harrison, 1952; Kretzmer, 1952; Oliwer, 1952).

Also important for the philosophical inspirations of PP are the works of such researchers as MacKay (1956), Neisser (1967), and Gregory (1980), who belong to the group of cognitive psychologists advocating "analysis-by-synthesis." According to this view, the brain does not build an internal model of the world by collecting (bottom-up) information received by the senses, but rather tries to create representations which it then compares with this information to choose those that are compatible.[10]

activity of the lower cortical layers. From the bottom up, information is generated about possible errors between top-down predictions and actual neuronal activity. Information about these errors is then used by the brain to create current predictions about the nature of the signal received through the sensory inputs. Based on the correction made, new predictions are created. Rao and Ballard's idea was used by other researchers and is still a fertile cognitive hypothesis within neuroscience today.

[7] For data compression approaches using predictive coding, see Shi & Sun, 1999.

[8] Jakob Hohwy claims that the sources of PP should be traced back to the thought of an Arab philosopher living in the late 10th and early 11th century, i.e., Ibn al-Haytham. In his work entitled *Optics*, Ibn al-Haytham explained that the process of seeing objects consists in light being reflected off their surface and then directed to the eyes. He defended the view that many visible properties of objects are recognized through judging and reasoning (inference) (Hohwy, 2013, p. 5).

[9] Dan Zahavi (2018) points to the relevant elements of the nineteenth and twentieth century neo-Kantian conception, which are present in the considerations of at least some supporters of PP.

[10] For example, Gregory claims that perception is based on top-down information processing. Information coming from the environment and reaching the brain through sensory inputs is ambiguous and indeterminate. To be able to interpret it, the cognitive system must use its previous experience and accumulated knowledge. With their help, the brain concludes what is perceived. Perception, according to Gregory, can therefore be compared to a process of making hypotheses in science. Both hypotheses in science and perceptive processes make it

It should be said that the authors who develop various approaches and positions regarding PP share, to a greater or lesser extent, the following beliefs (however, they differ in their understanding of the basic concepts and their scope):

1. The brain is or entails a multilevel, hierarchical model generating predictions (the so-called generative model);[11]

2. The hierarchical structure of the model can be mapped onto the hierarchical anatomy of the cerebral cortex;

3. The purpose of the model is to minimize prediction errors that result from the difference between the predictions made by the model and the information coming from the environment through the senses (i.e. exteroceptive inputs: visual, tactile and auditory) but also interoceptive inputs (e.g. heartbeat and states of arousal), proprioception and kinesthesis (cf. Garfinkel et al., 2015; Seth, 2013);

4. The model is hierarchical because it covers many levels of information processing;

5. Information in the model is processed in two directions: top-down (predictions about the information reaching the model) and bottom-up (information about prediction errors). This means that each level of the model that processes information provides (i.e., generates) predictions about what is happening at the level below, while receiving from that level information on the size of the prediction error (weighting precision).

2. 1. Prediction

The concept of prediction is crucial for the PP framework. It can be understood in four basic ways: (1) predicting the current state of affairs; (2) predicting future instances or statistical features of phenomena; (3) correlation, i.e., a statistical relationship between the values of various variables (e.g., the relationship between the length of school time and self-assessment; or the relationship between signals received by the eyeballs—signals from the left eye correlate signals from the right eye); and (4) inference and hypothesis testing. Some suggest (Anderson & Chemero, 2013, p. 204-205) that in PP, predictions are understood as correlations and inferences. However, predictions of sensory correlations do not assume the existence of a model and a network of connections between pieces of information (knowledge) (Clark, 2013, p. 236), whilst predictions understood as inferences assume such a model.[12] The latter

possible to recognize relevant situations and objects in various, variable contexts, based on residual and unclear information (Gregory, 1980, p. 182).

[11] The notion of a generative model is highly contentious. It is not just a question: is the whole brain a model, or part of it, or maybe (all or part) of the body?; but also the problem of whether the system has a model or is a model? Is the model an explicatory representation or a completely implicit model built into the structure of the system? E.g., some suggest that the generative model should be considered in terms of the whole organism (cf. Bruineberg, Rietveld, 2014; Bruineberg, Kiverstein & Rietveld, 2018; Friston, 2013a; cf. § 5. See also Conant & Ashby, 1970).

[12] Logically, predictions can be considered a kind of judgments about the future (e.g., in the form of the sentence "It will definitely rain") or hypothetical judgments (e.g., "If I wear a coat, I probably won't get wet").

are particularly important in PP, because they are such predictions that, based on an internal model, predict the causal structure of the world. With their help, the generative model recapitulates, i.e., synthesizes in some form (as a surrogate) the statistical structure of the world (Clark, 2013, p. 182; see also Gładziejewski, 2016).

Low-level predictions (e.g., about the shape of a perceived object) are more detailed, refer to smaller space-time vectors, and are short-term. High-level predictions (e.g., regarding the behavior of other agents or making decisions under risk conditions) are more general in nature, along with being more abstract and long-term. Thus, a prediction will be, for example, that a moving object is heading towards me with high probability; as well as anticipating that in difficult weather conditions, car drivers usually drive more carefully and conservatively. It should be added that the more abstract and general the prediction is, the more it is parametrized (i.e. expressed as a function of some parameters) by internal (e.g., Bayesian beliefs[13] (priors), heuristics, etc.) (cf. Friston, Wiese & Hobson, 2020; Millidge et al., 2020) and external constraints (structure of the environment, others agents, socio-material norms, etc.) (cf. Clark, 2018). New predictions are made on the basis of old ones and are supplemented with information on the prediction error. Predictions understood in this way are used to minimize prediction errors at every level of information processing carried out by the organism in relation to a specific perceptual, cognitive or non-cognitive task.

Prediction errors should be understood as disproportions that arise between hypotheses created on the basis of the model of the world possessed by a given cognitive system (e.g., brain or whole organism), and information from the world (specific data sets), which are provided by sensory inputs. In other words, the prediction error is the degree of mismatch between two signals (where greater mismatch means greater error). It can be concluded that any discrepancy between the prediction and the sensory signal results in the appearance of prediction error messages. In this sense, prediction error is the difference between sensory information and prediction (Hohwy, 2020a, p. 210). It should be added that machine learning distinguishes between reducible and irreducible prediction errors. A reducible prediction error is a mismatch between the hypothesis and the information from input. This relationship is not directly observed and can be estimated. An irreducible prediction error arises from the fact that the information from input doesn't completely determine the hypothesis. So it means that there are other variables outside and independent of data that still have some effect on the hypothesis. In the case of reducible prediction errors, the error can be minimized by generating more and more accurate information from input estimation. Thus an accurate estimation does not guarantee that the model won't be error-free because of irreducible errors. In this sense, irreducible prediction errors can be viewed as information that the model cannot extract from the data and that affects that data. Differentiating between reducible and irreducible prediction errors, resp. uncertainty requires an estimation of expected uncertainty based on previous prediction errors. The model must constantly evaluate uncertainty and devise strategies for reducing these irreducible errors, resp. uncertainty (cf. Gottlieb, 2012).

[13] Note that beliefs in this approach are non-propositional. See also footnote 17.

2. 2. Generative model

The concept of the generative model in cognitive science and neuroscience is not new (cf. Ng & Jordan, 2001; Pickering & Clark, 2014). However, none of its earlier applications had as many philosophical implications as its use in PP. Generally, generative models are statistically understood as mathematical models that capture the relationships between the values of a set of random variables. These dependencies are represented by model parameters, and the variables can be observed (i.e., some of their values can be provided directly by the data) or hidden. In PP, the generative model is understood as a statistical model whose variables are hidden, i.e., it is impossible to unambiguously assign the relevant data to the causal processes producing them. Statistical models are as accurate as the predictions they generate. These, in turn, are based on specific parameters that the model learns about based on the data it receives. The predictions generated by the model are described in terms of probability distributions. For this reason, generative models are referred to as probabilistic. The generative model is a statistical model of a joint distribution of probabilities where some random variables *X*, *Y*, ... attribute the probability that the variables *X*, *Y*, ... take specific values. A joint probability distribution can also be expressed as a common probability density function (for continuous variables) or a joint probability mass function (for discrete variables).[14] PP researchers claim that the generative model is a Bayesian probabilistic model, which means that it (in some way) implements the Bayesian rule to rationally (i.e. in accordance with the axioms of probability) combine existing and uncertain information with new evidence.[15] This means that speaking of the Bayesian nature of the generative model, one means that the predictions generated by it are based on priors (Bayesian beliefs) and as such are encoded in the internal parameters of the model, and then compared with incoming data and used to update parameters, thereby determining new probability distributions of specific variable values.

Generally speaking, the generative model can be understood as a coherent structure capable of generating a series of phenomena in a way that models the actual process by which these phenomena are generated. A good example are computer programs that generate realistic, two-dimensional or three-dimensional images of human faces, based only on a few input variables that relate to specific parameters, such as the location of facial muscles, eye embedding or skin texture (Kiefer & Hohwy, 2018, p. 3). Such programs implement a process that generates a given phenomenon as it were in relation to another generative process (e.g., by means of which base variables in the real world contribute to the appearance of real human faces). The extent to which the generative process can model another process or phenomenon depends on the similarity between the two processes. Other researchers (cf. Nair, Susskind & Hinton, 2008) add that the generative model as one that explains the investigated phenomena by determining the set of factors that could cause them. Both of these concepts of the generative

[14] We could say that generative model is a model of the conditional probability of the observable *X*, given a target *Y*. It means that generative model is statistical model of how observations are generated.

[15] Most Bayesians in this framework accept three assumptions: (1) probabilities are subjective degrees of belief; (2) Bayesian conditioning to update beliefs (with Bayes rule coming in handy to relate posterior to prior and likelihood); (3) inference follows rules of probability theory (see §3).

model, i.e., (1) a causally effective structure generating phenomena, and (2) a model that explains phenomena by referring to their causes are different. From a theoretical point of view, generative models only describe a certain probabilistic relationship, and whether we interpret it as causal or, for example, informative depends on the context in which they are used. To simplify, we can say that a model is such a structure that is capable of producing a given phenomenon and offers the best explanation of this phenomenon (cf. Seth, 2015; Kiefer, 2017).

Hierarchical models in PP explain the complexity of specific phenomena and are meant to show the causal relationships that connect them. This means that as the complexity of the model increases, the chances of generating the corresponding phenomenon using structurally different sets of causes increase too.

After these necessary conceptual and definitional remarks, we can now proceed to initial PP characteristics. According to the PP framework, the basic task of the brain, understood as a multilevel, hierarchical generative model, is to minimize prediction errors. Prediction errors should be understood as disproportions that appear between hypotheses created on the basis of the model of the world possessed by a given cognitive system, and the information from the world (specific data sets) that is provided by sensory inputs. Minimization of prediction errors is crucial for the organism, because—in accordance with PP—all perception is subordinated to the goal of ensuring that the actions of the organism in its environment are effective.[16] To minimize potential prediction errors, the generative model continuously creates statistical predictions about what is happening or can happen in the world. These predictions refer to the present and future shape of the information reaching the brain or organism through sensory modalities.

2. 3. Precision

Effective minimization of prediction errors presupposes estimation of the degree of their precision. By estimating or weighing precision it is possible to determine how precise a given error is, i.e., whether the information it transmits is reliable for the system or not. Precision is understood here as in statistics, i.e. as inverse variance. This means that the greater the average divergence from its mean, the lower the precision of a random variable (and conversely) (Friston, 2010). According to PP, precision-weighting is a process by which the model increases the gain on the prediction errors that are estimated to provide the most reliable sensory information, conditional on the higher-level prediction. For example, if the agent expects a given prediction error to be particularly reliable or highly probable, then the agent increases its weight, i.e., the extent to which it can affect model parameters. Precision is therefore a measure of the degree of uncertainty. The concept of precision refers to the estimation of uncertainty (Feldman & Friston 2010). Weighting of prediction errors according to their estimated precision is attention (Friston, 2009). High attention means that perception is based more on bottom-up incoming information from the sensory signal, because any small aberration is quickly

[16] Prediction error can be minimized in two ways: by changing model parameters or by interfering with the structure of the sensory signal (active inference) (Friston, 2010, p. 129; cf. § 6).

noticed. A low degree of attention means that, in a given situation, perception is based more on top-down predictions, e.g. only a vague outline of a given object is perceived, etc. Note that the terms used in PP such as precision, uncertainty, or surprise refer to the properties of probability distributions. Specifically, the inference process requires the representation of probability distributions with respect to the possible causes of sensory stimulation. These distributions allow for the estimation of many values. For example, a wide distribution encodes a high degree of uncertainty about a particular cause. The uncertainty is associated with the degree of surprise about the possible causes of incoming information. Therefore, precision can also be defined as certainty regarding a given belief (Adams et al. 2013, p. 1).[17]

Precision is also associated with accuracy, i.e., the degree of approximation between a measurement and the actual measured value. On the other hand, precision is the extent to which repeated measurements give the same results under unchanged conditions. Thus, precision of the prediction error is one thing, and its accuracy is another.[18] Therefore, precision estimation involves noise (for non-random quantities) and error (for random quantities).

2. 4. Hierarchy

After this discussion, we must return to the structure of the model. Predictions that minimize predictions errors are hierarchically organized and generated at any individual levels of the generative model. Thanks to this, predictions present at higher levels of the model refer to information about prediction errors that are present at lower levels. It means that predictions organize incoming bottom-up sensory information from the top-down to minimize surprise[19] as much as possible, surprise which is associated with ignorance of the causes of sensory stimulation. In other words: predictions are sent down the hierarchy and the information about the size of the prediction error is returned upwards. This information is precisely estimated in order to determine its reliability.

The above structure of the model relates to the observation according to which the brain is organized hierarchically. Further cortical areas can be considered as forming a hierarchy. The lower areas (levels) of this hierarchy are associated with sensual inputs, while the higher ones with multimodality and associative functions. The concept of hierarchy refers primarily to the so-called descending (forward) and ascending (backward) connections, which are based on the specificity of cortical layers responsible for internal connections in the brain (cf. Felleman & Van Essen, 1991; Rao & Ballard, 1999). In PP terms, cortical hierarchies, by implementing descending connections, generate predictions that relate to the information on the causal structure of certain states of affairs, which are contained in ascending connections and coded in the

[17] Belief in PP should be understood as the probability distribution concerning unknown states of affairs or properties of the world. In other words, belief is a systemic (prior) belief with a high degree of abstraction, i.e., relating to general knowledge of the world.

[18] This is convincingly justified by Kwisthout, Bekkering and van Rooij (2017) who argue that, for estimating precision, it is not necessary for error information to be very detailed.

[19] This organization has been best studied according to the visual system (cf. Rao & Ballard, 1999).

form of probability distributions. Uplink connections provide feedback on prediction errors to higher levels. The top-down and bottom-up connections are therefore implemented at every level of the hierarchy (Friston & Stephan, 2007, p. 443).

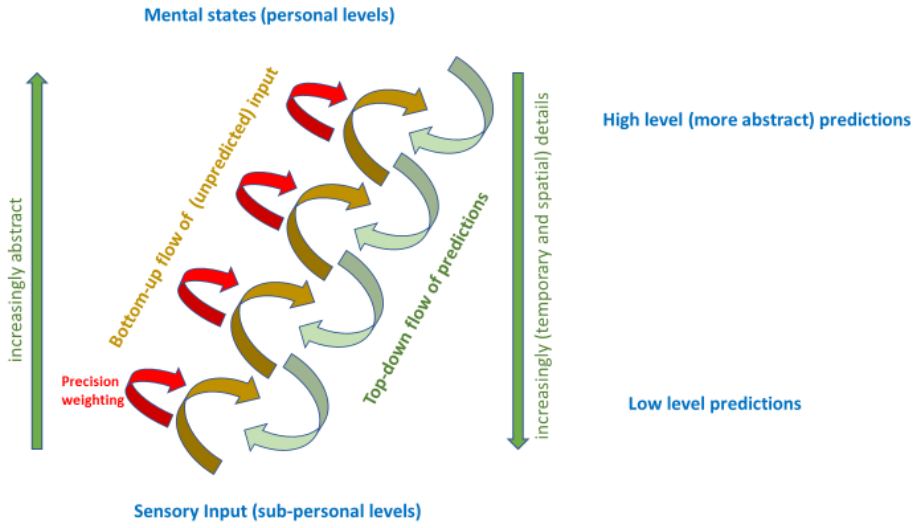

Fig. 1. Schematic representation of the levels of information processing by the generative model. Bottom up sensory information is processed in the context of the agent's knowledge and priors (i.e., internal parameters of the generative model) coming from the higher levels of the generative model. Unpredicted information from the sensory signal is carried up the hierarchy leading to the generation and adjustment of appropriate predictions. The error information is weighted with precision and its residuals are carried over to a higher level where they are minimized. The whole process is continuous and repeatable. (Adapted from Clark 2016, p. 30, 59).

The architecture of the generative model is related to space-time vectors. As its structure is bidirectional (in the sense of information processing) and hierarchical, the model enables processing of specific information only in relation to specific time scales and spatial dimensions relativized to them. It is limited to the difference between the higher and lower levels of the model (cf. Friston, 2003), namely predictions at the (arbitrary) level—let us mark it as *x*—minimize the prediction error at the *x-1* level, but they cannot do it at the *x-2* level. On the other hand, predictions at the level of *x-1* minimize the error at the level of *x-2*, etc. (cf. Friston & Kiebel, 2009; Hohwy, 2015a). The reason for this is the time differences between levels.[20] The higher we go in the hierarchy, the greater the time dimensions of subsequent model levels. At the same time, individual levels become increasingly abstract: from not very abstract, e.g., regarding edge detection, to more abstract, referring, for example, to processes of categorization. Therefore, it can be said that subsequent levels of the model are built on the basis of its

[20] This is one of the main objections against PP (cf. Williams, 2019a; 2020).

spatio-temporal structure, rather than specific environmental properties. In other words, they are not distinguished because of what they represent, i.e., because of the content, but because of the time scale in which they operate (Hohwy, 2013, p. 72).

3. Bayesian-brain thesis

To clarify the nature of perception, motor control, decision making, and so on, the key issue is the prediction that the cognitive system makes during each contact with external states of affairs. In this perspective, the issue of perception concerns the question of how real the hypotheses about the world are shaped and selected (Hohwy, 2013, p. 16). In other words: how, without knowing the causes for sensory stimulation, model can hypothesizes about their causes? Such a task would require extrapolating specific information from an uncertain data set. Therefore, the model processes information based on some form of statistical inference. These findings come from the observation that sensory information does not shape perception directly. Rather, it is actively selected and properly used. Predictions integrate relevant aspects of the perceived world. The issue of perception, therefore, concerns the possibility of using data that reaches the model via sensory inputs in such a way that the organism does not make prediction errors. These errors, however, come from ignorance of the causal sources of sensory stimulation. There is no unambiguous relationship between causes and effects as different causes can have the same effects. Ignorance of these causes can pose a practical threat to the organism.

Many researchers (cf. Fink & Zednik, 2017; Harkness & Keshava, 2017; Hohwy, 2013; 2014) adopt the Bayesian-brain thesis (cf. Knill & Pouget, 2004; Knill & Richards, 1996) according to which the generative model constructs and tests internal models of the external world by implementing cognitive processes that are an approximation[21] of Bayesian inference (Clark, 2013, p. 189).[22] Then, using hierarchically organized inferences, the brain creates appropriate multi-level generative models that generate hypotheses top-down to "interpret" bottom-up information from the sensory input. Each level of such a model minimizes prediction errors at a lower level—from neuronal processes to higher cognition. This means that the model does not directly compute true posterior distribution of hidden states, but iteratively updates (in tractable way) the approximate posterior via gradient descent to minimize prediction errors. Optimizing the model in this way allows one to find a distribution that approximates the exact posteriori (an approximate posterior distribution over states makes simplifying assumptions about the nature of the true posterior distribution) (cf. Sanborn, 2017). What does it mean? Let's consider an example: I am driving at night down an unlit, one-way road. I see two points

[21] Approximation of a given function using another, simpler, which is easier to study and apply.

[22] Some researchers defend a non-inferential interpretation of Bayesian inference, which can be understood in an enactive way as a type of action (cf. Ramstead, Kirchhoff & Friston, 2020). Depending on the interpretation, this approach may lead to the conclusion that Bayesian inference is not a computational process, but e.g., a process controlled by the dynamics of the system (see Kelso, 2012; cf. Bruineberg & Rietveld, 2014; Bruineberg, Kiverstein & Rietveld, 2018). These approaches are not free from problems. For example, Korbak (2021) shows, defending the position he describes as computational enactivism, that one can be a supporter of the computational theory of mind and semantic information, while accepting the premises of enactivism.

of light approaching. I predict that they are headlights of a car coming from the opposite direction. I also assume that this car is in the right lane. However, there is a risk that it is driving against the flow of traffic. I am not sure. So how can I decide what to do? If my representation of the movement of the car facing me turns out to be wrong, this error can ultimately cost me my life. The accuracy of my prediction about what is going to happen on the road not only depends on the information that reaches me through the senses, but also on my knowledge, experience, and beliefs.

According to PP, the prediction error will be minimized only when the model adopts the best possible hypothesis regarding the causes of the sensory signal source. Based on this hypothesis, predictions are then generated to condition the actions of the cognitive system or organism. The model is multi-level, thanks to which each level minimizes the prediction error at the lower level. In the example with the car, one level of the model (higher) will concern the possibility of recognizing light points as headlights, the other (lower) will refer, for example, to the detection of the edge of the perceived object, the next level will generate predictions regarding, e.g., two vehicles colliding, etc. At each level, the model estimates how precise a given prediction error is, so that the hypotheses adopted so far can be reviewed (Friston 2009, p. 299). So how is the process of creating and selecting hypotheses that have different probability values possible? The hypothesis that the lights approaching from the opposite direction are caused by a different car is more likely than the hypothesis that their source is a spaceship.

The view postulated by many researchers that the predictive brain is a Bayesian brain means that the generative model resembles to some extent the hierarchical Bayesian network[23] implemented in the brain. The general idea is this: the generative model creates and selects appropriate hypotheses, using Bayesian inference. Therefore, minimizing prediction errors is an empirical application of Bayes' rule.

On the basis of the Kolmogorov axioms,[24] the definition of prior P(A) and posterior P(A/B), the Bayes' theorem is proved:

$$\mathbf{P(A|B)} = \frac{\mathbf{P(B|A)P(A)}}{\mathbf{P(B)}}$$

[23] A Bayesian network (Jensen, 1996) (also referred to as belief networks, cf. Pearl 1988) is a pair (*D*, *P*), where *D* is a directed acyclic graph (DAG) and *P* is a probability distribution. Each node in network stores the distribution $P(Xi|\pi(i))$ where $X\pi(i)$ is the set of nodes corresponding to $\pi(i)$ predecessors (parents) of node *(i)*. In such a graph:

1. nodes represent random variables (property of a given item);
2. arcs represent specific (probabilistic) relationships between variables (the variable *x* has such and such influence on the variable *y*);
3. variables represented by specific nodes take on discrete properties (e.g., YES, NO);
4. each specific node has associated conditional probability tables, which determine the influence exerted on that node by its predecessors (parents) in the graph..

[24] These axioms impose the following conditions on the a priori probability: 1. for any element x: $0 \leq P(A) \leq 1$ (axiom of non-negativity); 2. P (T) = 1 (where T is a tautology) (normalization axiom); 3. $P(A \vee B) = P(A) + P(B)$ when $A \wedge B = \neg T$ (axiom of finite additivity).

Bayesian theorem (also called Bayesian rule) allows you to calculate the secondary probability (posterior) P(A/B) of e.g., a given hypothesis A when the inverse probability (likelihood) P(B/A) and the primary probabilities (priors) P(A) and P(B) are known. In other words, Bayes' theorem determines how evidence or data (e.g., empirical) affect the probability of hypothesis A.

Proponents of Bayesianism associate the method of updating the probability or the change of certain hypotheses under the influence of new data with the so-called principle of conditioning related to Bayesian confirmation theory. Note that the Bayesian rule is a mathematical equality but Bayesian conditioning is a philosophical position about updating beliefs in the light of new evidence. It does not follow logically or mathematically that this new belief in hypothesis A should be P(A|B).The very determination of the probability of a given hypothesis (synchronous aspect) does not raise any doubts (assuming many possible interpretations of the concept of probability), so the possibility of updating them raises a number of controversies (cf. Kawalec, 2003). This is important for the analyzes carried out here, because in accordance with the basic assumptions of PP, the generative model not only computes the probability of certain hypotheses and beliefs, but also has to update its parameters by changing or modifying the previously obtained knowledge. In order to be able to indicate how this is possible, it is necesary to say a few words, as in the PP framework, probability is understood as one of the possible interpretations of Bayesianism.

There are different interpretations of the concept of probability depending on what property the probability is to measure: (1) frequentist interpretation (R. von Mises); (2) logical (objective) (J. Keynes, R. Carnap; J. Hintikka); (3) epistemological (K. Ajdukiewicz, H. Kyburg) and (4) subjective (F. Ramsey; B. de Finetti; L. Savage and proponents of PP). In subjective interpretation probability is understood as a measure of a person's confidence in the truth of a given hypothesis, proposition or theory. The position itself is obviously not homogeneous and its various subtypes can be distinguished, e.g., behavioral, which associates beliefs with specific behaviors in decision-making situations.

Probability in PP is understood here as a measure of a person's conviction. For this reason, the Bayes' rule is referred to as the subjective probability rule.[25] The example of an unlit, one-way road illustrates this problem well. Let us repeat: driving a car, I see lights coming towards me from the opposite direction. I make the hypothesis that they are the lights of a car going against the traffic. I assume that if I see lights approaching me at night (this is the information from a sensory signal), they are the lights of a car going against the traffic (hypothesis explaining the source of the information). In the light of my previous knowledge and experience such a hypothesis is more likely than the hypothesis that the source of the approaching lights is a spacecraft or lighting installation.

[25] There are a number of examples that suggest that people do not apply Bayes' rules by intuitively assessing the likelihood of a given event in the light of new experiences. They make two mistakes: (1) they are conservative in revising the initial probabilities under the influence of new data; and (2) ignore the initial probabilities (cf. Tversky & Kahneman, 1979). However, it should be added that these objections with regard to PP are somewhat limited: PP assumes the unconscious nature of applying the Bayesian rule.

Each of the levels of the generative model that is involved in the perceptual situation just described applies Bayes' rule. Error information modulated by the attention mechanism (expected precision) goes to a higher level, where the *posterior* probability is calculated, the model parameter is changed, and the appropriate prediction is created to be tested. It can be said that the parameters of the internal probabilistic map of the environment are changing. Based on available data related to the discussed situation, the brain calculates the probability of prediction for specific levels of the model. They concern, among others:

1. hypotheses regarding shapes, colors or sounds (low levels of the model); and
2. hypotheses regarding the movement of cars, braking distance, evasiveness, etc.;

and

3. hypotheses related to traffic situations, moving vehicles, traffic lights etc.

The generated hypotheses are based on the information received from the senses.

The important thing in our example is that the probability of the hypothesis whose aim is to minimize the prediction error as much as possible, is based more on the subjective assessment of a given event than for example on some "objective" properties of reality (Rescola 2015, p. 4). This assessment, however, is not completely arbitrary because it is associated with the probabilistic nature of the information given in the sensory signal. However, it is subjective because we are not aware of causal sources that trigger sensory arousals. Therefore, one should accept their most likely (from the organism's point of view) explanation.[26]

It is widely believed that Bayesian models are located at Marr's computational level (Chapter et al., 2010; Friston, Fortier, Friedman, 2018, p. 23-24) because they provide an understanding of *why* a given system does what it does. From this perspective, it can reasonably be assumed that there is more explanatory power in those models that allow researchers to view other levels of analysis than the computational. In the approach developed here, the multilevel and hierarchically organized model generates subsequent predictions that minimize potential prediction errors present at lower levels of the hierarchy. At each level, the Bayes' rule is to be implemented for the organism to represent the world and act effectively in it (cf. Clark, 2013, Friston, 2005). In addition to descriptions at the computational level, various types of algorithms are being developed that approximately implement the Bayes' rule (cf. Sanborn, 2017; Spratling, 2017). Thus, they enable answering questions about the *how* of a given mechanism, i.e., to a certain extent, functionally explain a given system. As I mentioned earlier, the issue of a neuronal implementation of the proposed algorithms remains unclear, but many authors make efforts to indicate its possibility (cf. Friston, 2005; Gordon et al., 2018; Parr & Friston, 2018; Rao & Ballard, 1999).

[26] Using abductive inference (Seth, 2015).

4. Problems with Bayesian models

Bayesian confirmation theory[27] has faced a number of objections. The most important are related to the so-called paradox of the ravens,[28] the problem of logical omniscience,[29] the problem of old evidence[30] and the new problem of Goodman's induction. The last problem needs a closer look. This paradox generally concerns the fact that the applicability of the confirmation theory is determined by certain philosophical assumptions that must be adopted with it (explicitly or not). For if the confirmation theory is not based on certain philosophical assumptions (e.g., concerning the way of constructing a probabilistic space or its objects), it leads to absurd conclusions, e.g., confirming contradictory statements. Methodologically, Goodman's problem can be presented as follows:

[27] Bayesian theory of confirmation is quantitative, not a qualitative. It tells us whether a given piece of evidence confirms a hypothesis, and additionally tells us how much, or the degree to which, a piece of evidence confirms a hypothesis. This theory has to explain to possibility of updating new beliefs by the Bayesian generative model. This is made possible by the principle of conditioning, which allows the probability of certain hypotheses or beliefs to be updated. This principle is based on the systematic application of the Bayesian rule in such a way that the relative probabilities of the hypotheses from the earlier stages of the justification are treated as the a priori probabilities in the following stages. The confirmation theory based on the principle of conditioning serves Bayesian supporters (especially in the subjective version) to demonstrate scientific objectivity. Regardless of the preferences and beliefs of scientists in relation to a given hypothesis (i.e., the primary probability that scientists assign to it), taking into account the same data (e.g., empirical), after some time the differences in the preferences of scientists will be leveled (i.e., posteriors will be almost identical) (Kawalec, 2003, p. 116).

[28] This paradox is related to Hempel's observation that his confirmation theory (which forms the basis of Bayesian confirmation theory) predicts that the observation of the black raven (as expected) supports the hypothesis that all ravens are black. However, the observation of the non-black non-raven also supports this hypothesis. One can defend this theory by arguing that in the real world a confirmation based on the observation of black ravens is much more reliable than that based on the observation of non-black non-ravens. (cf. Fitelson, 2006).

[29] According to the standard axioms of the theory of probability, it seems that every logical truth has a probability equal to 1, that is, be sure. The use of the phrase "seems" is appropriate because Kolmogorov's axioms do not really entail this claim. However, if one takes the normal interpretation and concludes that the set of possibilities is some set of (logically or metaphysically) possible worlds and identifies the propositions with the set of worlds in which they are true, it means assigning this logical (or perhaps even metaphysical) omniscience to subject. (This problem is related to the problem of practical computability of Bayesian models). Bayesian supporters respond to this objection by claiming that it is not a descriptive theory, but a normative theory, therefore some idealizations are possible.

Other important constraints of Bayesianism are related to, inter alia, with the problem of setting priors or historical inadequacy (cf. Easwaran, 2011).

[30] This problem was formulated by Clark Gilmour (1980), who noted that some evidence that has been known for some time - that it is old evidence, and that there is a hypothesis or theory that has been analyzed for some time. At some point it is stated that this hypothesis implies this old evidence. Typically, this would mean that the hypothesis implies this evidence. However, according to Bayesian confirmation theory seems unable to explain how a previously known evidence could provide any new support for that hypothesis. For conditionalization to come into play, there must be a change in the probability of the evidence but if this evidence is old evidence, there is no change in its probability. Attempts to solve this problem are discussed, among others, by Howson (1991) and Eva and Hartman (2020).

For any quantitative confirmation theory of *T*, there are two (mutually translatable) interpretations of I_1 and I_2, such that for the sentence $\{\tau(\alpha) = x\} \in T$ w I_1 and I_2 the degrees of confirmation of this sentence are paradoxically divergent (in particular $\{\tau(I_1(\alpha)) = x\} \wedge \{\tau(I_2(\neg\alpha)) = x\}$, where τ is the confirmation function at *T* and $x \in [0, 1]$).

The Goodman paradox proves that it is impossible to construct a good confirmation theory without adopting minimal philosophical assumptions. These assumptions concern, inter alia, the following issues:

1. the method of constructing a probabilistic space;
2. assigning probabilities to original hypotheses;
3. assigning probabilities to conditions of hypotheses;
4. rules for updating the probability of hypotheses etc. (cf. Kawalec, 2003).

In the light of these remarks, it is obvious that the Bayesian PP framework adopts a number of philosophical assumptions (including those concerning the unconscious nature of the Bayesian rule implementation, the nature of predictions, or information from the environment). In PP, updating beliefs is directly related to assigning a certain weight to the prediction error. The weighting is based on the accuracies of the probability distributions (or in the case of the continuous probability density function), where precision is the reciprocal of the variance so that a highly precise distribution has very little variance. Weight expresses the speed of learning, which increases as precision is assigned to a belief. We can speak of a precision optimization mechanism. I.e., the model can generate diminishing precision predictions to create a learning rate variable that is sensitive to contextual information. The generative model, therefore, must be hierarchical, so that the weight of the prediction errors at any level can be modulated by learned patterns in several spatiotemporal ranges, i.e., at other levels. For example: the weight of the predicton error related to the perception of cars moving on the street will decrease when I drive a car with glasses (assuming I have poor eyesight) and when, for example, I drive a car on a road I know (Hohwy, 2020a, p. 210-211).

PP supporters can defend themselves against the above objections by claiming that (1) the generative model implements the Bayesian network in an approximate and not exact manner (cf. Sanborn, 2017; Hohwy, 2020a); (2) many of the arguments criticizing Bayesianism relate to the use of probability theory at the personal level, and much of the literature and research in the field of PP focuses on sub-personal processes; (3) the concept of probability in PP is also frequency-related, i.e., one can speak of a certain "objective" probabilistic structure measured in the *Natural Scenes Statistics* (Orlandi, 2016). This solution helps to explain why certain possibilities are excluded in advance in PP. This is because the world is in some way previously structured; (3) it can be argued that in some types of cognitive processes the brain does not carry out fully developed and formally correct complex calculations, but uses simplified heuristics which—as empirical research confirms—may in certain circumstances be more effective than the use of complex calculations (cf. Gigenzer, Brighton, 2009; Parpart, Jones & Love, 2018). This possibility is shown by some supporters of the active inference framework (cf. Constant et al., 2019; Veissière et al., 2020). It should be emphasized, however, that such attempts are absent from the literature on PP (cf. Colombo, Elkin & Hartmann, 2018).

The strongest arguments against the use of these models in cognitive science and neuroscience were presented by Bowers and Davis in the famous paper *Bayesian just-so stories in psychology and neuroscience* (2012). They argue that it is not modern Bayesian models that explain a number of issues related to the performance of various tasks by our brains, but traditional non-Bayesian approaches.[31] Also, Bayesian models and their assumptions tend to be fitted post-hoc around data (cf. Litwin & Miłkowski, 2020). The reason is that (1) Bayesian models used in psychology have little support in empirical evidence. This translates directly into the fact that these models predict phenomena to a much lesser extent than classic models; (2) the use of Bayesian models in neuroscience is even less empirically justified. There are many mathematical analyses that show how specific neuronal populations could calculate in a Bayesian manner, but there is very little evidence that would confirm these analyses empirically; (3) it is not justified to believe that psychological theories should be constrained only or primarily by rational analysis of what the brain should do. According to Bowers and Davis, the Bayesian approach cannot be accepted, because there are many other equally important constraints. These include evolutionary, biological or computational constraints. Ignoring these constraints leads to the conviction underlying Bayesian models that brain function is based on optimizing behavior correlated with specific tasks (see also Elqayam & Evans, 2011). In other words, we can talk about over-intellectualism regarding brain work. I think that, in part, the above remarks can be combined with the earlier observations. Namely, Bayesian models are developed primarily at the computational level of analysis demonstrated by Marr (cf. Colombo & Seriès, 2012).[32] It is different, as some argue, in the case of Bayesian modeling used in PP. If so, isn't Bowers and Davis' criticism focused on PP-based approaches? It is difficult to answer these questions unequivocally.

Other researchers develop the last of the arguments put forward by the authors of *Bayesian just-so stories in psychology and neuroscience*. Namely, if one looks at cognitive processes in the context of the biological evolution of the human species, the use of probability—or simply the Bayes' rule—is something unusual for man, appearing relatively late in the process of evolution (cf. Cosmides & Tooby, 1996; Williams, 2019b).[33]

The arguments given are strong and their acceptance may lead to a conclusion that denies the explanatory power of some, and perhaps even all Bayesian models. When analyzing them, it

[31] In response to this objection, Griffiths et al. (2012) states that Bowers and Davis ignore the fact that Bayesian approaches in science "[lead] us to new ideas about the mind and brain". This means that they cannot be replaced with other non-Bayesian research frameworks, because both Bayesian and non-Bayesian frameworks can provide valuable and irreducible information about the phenomena studied.

[32] Clark Glymour (2001) attempted to defend the implementation of Bayesian algorithms by arguing that Bayesian networks are a species of neural networks, which in turn are a species of graphical causal models. His claim is based on observations and research on people with brain damage. Glymour assumes that knowledge of patients not properly performing certain tasks provides information about the structure of the mind. For example, a failure in an acyclic neural network causes the emergence of new probabilistic relations of independence, which means that the Bayesian neural network hypothesis allows testable predictions about the damaged brain. Important for Glymour's analyses is the possibility of graphical modeling of processes carried out in the brain (this idea was developed by Glymour's student David Danks [2014]).

[33] See footnote 25.

should be noted that most Bayesian approaches create models of the ideal observer. As Knill and Pouget (2004, p. 712) claim, in Bayesian models, "human observers act as optimal Bayesian observers."[34] It seems that, in a predictive approach, due to algorithmic levels and implementation, this ideal observer ceases to be ideal. This is because, as some researchers claim, models in PP are oriented on actual processes, so they relate to specific physical components and causal relationships that are responsible for the implementation of such and such human behaviors (cf. Harkness & Keshava, 2017; Hohwy, 2015; Spratling, 2017).

Another complaint that should be mentioned is related to the so-called NP-hard problem. It concerns the practical computability of Bayesian models. This means that the Bayesian requirements (associated with explicit Bayesian inference) that the agent should meet are impossible to implement.[35]

The question of whether PP enables modeling cognitive phenomena at non-computational levels of description distinguished by Marr, i.e., algorithm and implementation, raises many doubts. As for the problem of mechanistic implementation, this issue raises a number of ambiguities. It looks different when used at the PP algorithm level. Here the situation is much more promising (cf. Spratling, 2017; Friston, Fortier & Friedman, 2018, p. 23-24).[36]

Jones and Love argue that, despite the general assumptions in Bayesian models, there is actually no application of Bayes' rule, but rather of "mathematically elegant formalization of abstract induction" (Jones & Love 2011, p. 178). In practice, this means that (1) the model is already equipped with prior knowledge, which, for example, allows it to recognize patterns (e.g., regularity in a string of words in which the dimensions of the letters matter), so the fact that the model recognizes this and not another pattern is not completely surprising or theoretically informative; the model, which is to explain the relevant cognitive function of a given system, assumes a way of its implementation; and that (2) strong assumptions are made here

[34] In this context, it is worth recalling the old objection that Bayesianism assumes that the cognizing subject is ideal as its beliefs are to be consistent with the axioms of probability theory. Proponents of Bayesianism claim that the theory is normative, not empirical, and therefore some idealizations are allowed. When discussing this objection in the context of PP, it should be remembered that beliefs are not understood as in classic Bayesian models. Rather, they denote hierarchically defined probability distributions in a multilevel generative model.

[35] In the context of PP, see Kwisthout & van Rooij 2019. Possible strategies to solve this problem are discussed by van Rooij et al. 2018. These researchers, however, claim that none of these strategies ultimately solves the NP-Hard. Note that the authors of this paper do mention a strategy that provably resolves the issue of NP-hardness of Bayesian inference. The other strategy for solving this problem is proposed by Thornton (2016), who claims that a generative model approximating Bayesian inference can implement a universal Turing machine (see also Sanborn, 2017; Friston et al., 2006).

[36] Spratling lists five possible theories (this list is certainly not closed) that use PP at an algorithmic level. These are linear predictive coding, predictive coding in the retina, predictive coding in the primary cerebral cortex based on the Rao and Ballard model, predictive coding in the primary cerebral cortex based on the BC model (Biased Competition) and predictive coding based on the FEP model. The algorithms discussed also differ in how the prediction coefficients relate to the relevant biological neural circuits. Some believe that these coefficients correspond to the synaptic weights of the lateral connections in the retina (Srinivasan, Laughlin & Dubs, 1982); others associate them with cortical pyramidal burn-out indices of feedback (Friston, 2009; Kiebel & Friston 2011) or with the responses of the corresponding conjugated pyramidal cells (Spratling, 2012).

regarding the environment in which the system operates, i.e., a formal model of the environment is created for which the system is to be adapted. In the approach, both of these model properties are associated with a strong assumption regarding the rationality of inferences present in traditional Bayesian models (cf. Oaksford & Chater, 2007). Classical literature on heuristic inferences (cf. Tversky & Kahneman, 1974; 1979) no longer allows for an uncritical transition to this assumption. The implementation of Bayes' theorem is therefore only the general hypothesis adopted in this approach.

These comments do not seem to be well-founded in relation to Bayesian PP, namely in PP there are no inductive inferences in the sense defined e.g., by Rational Analysis. There, the behavior of a given organism is understood as the optimal solution for a specific probabilistic inference that relates to the task arising from contact with the environment, based on a formal model of the environment.[37] In PP, hypotheses formulated on the basis of these inferences explain information best if they are treated as evidence in favor of the statements that the hypotheses make (Hohwy, 2014; cf. Hempel, 1965, p. 372–374). This means that in PP, the cognitive system "extracts" its priors from the raw data through the process of incremental model optimization by minimizing prediction errors. The model is thus constructed through continuous learning and error minimization, so that no specific beliefs need to be assumed in advance. This process can be called inductive or abductive, but the key thing is that it is unsupervised: the inputs are not a priori classified, and the beliefs at the starting point can be randomized and then progressively match the input statistics. For this reason, the generative model can be understood as self-evidencing (Hohwy, 2014). In PP we are not dealing here with a formal model of the world like in Rational Analysis, but with hierarchically organized probability distributions regarding relevant information transmitted through sensory inputs.

The above characteristic is purely formal because it does not show why and in what way the model should use Bayes' inference, and why it should be probabilistic rather than some other kind of inference (Colombo, Elin & Hartmann, 2018). Clark claims (2013b, p. 189) that in PP cognitive processes are an approximation of Bayesian inference. This means that an intractable problem is converted into a tractable optimization problem (Friston, 2011; cf. Buckley et al., 2017; Dayan et al., 1995).This explanation does not say much about how the brain would exactly follow the strategy prescribed by Bayesian rule.[38] It seems that the hyper-realistic interpretation, which understands the concept of approximation literally, i.e., in the sense that the brain directly uses Bayes' theorem, should definitely be ruled out.

[37] In this approach, the following steps can be distinguished (Oaksford & Chater, 2007, p. 71-72):

1. The goals of the cognitive system are identified;
2. A formal model of the environment to which the system is adapted is developed;
3. Minimum assumptions about the calculations are made;
4. The optimal behavior function, determined in steps 1-3 is obtained (this requires formal analysis using rational norms such as probability theory, logic or decision theory);
5. Empirical evidence is tested to see if the predictions about the behavior function are confirmed; 6. Steps 1-5 are iteratively repeated to refine the theory.

[38] Gładziejewski suggests one should speak here about the implementation of the Bayesian network, whose structure resembles a causal-probabilistic network of the environmental structure (Gładziejewski, 2016, p. 571).

Alex Kiefer (2017) argues that perceptual inference in PP should be treated literally, i.e., as a process of reaching the truth and maximizing the coherence of the model. He refers to Harman's view (1973) according to which inference is a search for coherence or a change of views, which is connected with the rational requirement to maintain the truth in relation to appropriate representations or beliefs.[39] The argumentation adopted by Kiefer is convincing because this author shows how Bayes' inference, understood in this way, can be used in computational models proposed e.g., by Sejnowski and Hinton (Hinton & Sejnowski, 1983) that inspired modern computational models. The fitting of new information to the model is treated as a special case of consistent information, or in other words: maintaining the coherence of internal model parameters. Therefore, increasing the coherence of the model allows for greater consistency in representing the relevant probability distributions. In PP, precision (Clark, 2013; Hohwy, 2013, Ch. 3) is used to specify the already existing knowledge (priors) and new information reaching the model in relation to subsequent levels of the hierarchy. The conclusions are a precisely weighted combination of prior knowledge and the likelihood of specific information appearing. Precision, therefore, is a way of building a compromise between knowledge and incoming information (Kiefer, 2017, p. 19). This way, the model, using new data, increases its internal coherence.

Colombo, Elkin and Hartmann (2018) put forward a number of objections that speak against the realistic approach shared by some PP supporters with regard to Bayesian inferences. First, the Bayesian approach in PP and more broadly in cognitive science does not implement special epistemic virtues that other alternative approaches would not share. The point is that the approach is neither simpler (with greater unification possibilities) nor, for example, more rational. Second, Bayesian algorithms have so far found poor support in empirical evidence. According to Colombo, Elkin and Hartmann, an anti-realistic approach should be adopted, which in practice means agnosticism according to Bayesian models.[40] So is the choice of Bayesian approach in PP arbitrary? There are reasons to think so.[41]

[39] The following objection can be made to Kiefer's solution: the search for coherence is an essentially undecidable problem, and the consistency is NP-Hard (though not NP-Complete). The problem of the subject's omniscience and the question of the infinite computability of the generative model arises here. This objection can be weakened by claiming that the generative model uses simplified inferences (heuristics) in its calculations. However, so far there is no such proposal in the literature on PP. One can also take the complete class theorem (Brown, 1981; Wald, 1947) according to which for any given pair of loss functions and decisions, there are some priors that render the decisions Bayes optimal. This means that every conceivable set of behaviors, resp. actions is explainable with respect to the (at least one) set of priors (Friston, Adams & Montague, 2012, p. 6).

[40] These authors also discuss other alternative methods of uncertainty modeling (for example the quantum probability) (Colombo, Elkin & Hartmann, 2018, p. 12-18).

[41] Colombo, Elkin and Hartman (2018) argue that Bayesianism is still the most popular approach to studying uncertainty and modeling cognitive processes. The choice of this approach is therefore a conservative move, one which favored the development of specialization within empirical sciences (Stanford, 2019). The search for new solutions is characteristic of modern institutionalized science, but research liberalism alone is not a sufficient argument for rejecting the conservative Bayesian approach.

5. Conservative and radical approaches to predictive processing

Even a sketchy reading of some PP studies leads to the conclusion that there is no single and comprehensive interpretation of this framework. We see that there are a number of different critical views that are often mutually exclusive. When analyzing this research approach, although the role it can play in cognitive science and philosophy is clear, one can notice a number of discrepancies when it comes to the understanding of basic concepts, terminology, and the explanatory or methodological status of every statement. There are at least two approaches: conservative and radical (cf. Clark, 2015a, 2015b; Dołęga, 2017; Gładziejewski, 2017b; Orlandi, Lee, 2018).[42]

5. 1. Conservative predictive processing

The conservative approach to PP (cf. Gładziejewski, 2016, 2017b; Hohwy, 2013; 2018; 2020a; Kiefer, Hohwy, 2018; Wiese, 2017) emphasizes that the mind is relatively isolated from the environment, which means that its cognitive contact with the outside world has its source in the neuronal activity of the brain and only in it. The internal model of the world is coded at the neuronal level (Clark, 2015a, p. 14). The generative model produces predictions that function as representations of what is happening in the external environment. This means that the relationship between the mind and the world is mediated by internal representations. In this framework, the world model is constructed on the basis of internal representations of the world which are isomorphic with its causal structure (Gładziejewski, 2016, Kiefer, Hohwy, 2018). Some describe these representations as structural because they operate on the basis of the structural similarity between the representation itself and what it represents (cf. Gładziejewski, 2016; Gładziejewski & Miłkowski, 2017; Kiefer & Hohwy, 2018; O'Brien & Opie, 2004; Shea, 2014). In this context, one can recall the words of Hohwy who claims that "the causal net of the environment and the causal net represented in the internal model will mirror each other" (Hohwy, 2018, p. 4), because the cognitive system is "an internal mirror of nature" (Hohwy, 2013, p. 220).

Gładziejewski (2017b) points to three basic commitments that are characteristic of conservative PP (Gładziejewski, 2016, 2017b; cf. Hohwy, 2013; 2018; Kiefer, Hohwy 2018; Wiese, 2017). These are (1) the commitment to representationalism; (2) the commitment to using the concept of inference as subserving perception and action; (3) the commitment to internalism as the position that cognitive mechanisms are based solely on the work of the central nervous system. This means that the content of mental representations is determined only by the internal states of the organism (cf. Lau & Deutsch, 2002).

The first commitment is particularly important for the analyses presented in this article. Gładziejewski claims that this commitment can be interpreted in two ways. The so-called weak (pragmatic) interpretation assumes that the content of mental states is attributed to the internal model of the world for purely pragmatic reasons. In this approach, internal representations do

[42] Zahavi, however, emphasizes that his criticism primarily concerns the conservative PP and—to a much lesser extent—its non-representationalist alternative (Zahavi, 2018, p. 55). For this reason, I will discuss the conservative PP and will focus on demonstrating that this position has no anti-realistic consequences.

not have real content, but they are only postulated by some researchers who want to explain the cognitive functions of the mind (cf. Egan, 2014; Downey, 2017). Strong (realistic) conservative PP postulates that mental states contain real and causally relevant representative content. For this reason, it should be stated that the generative model has real content and represents the world in a non-trivial way. This means that (1) it generates environment-oriented predictions that guide actions; (2) the function of guiding action depends on the resemblance between functional relationships among encoded variables and the causal structure of the environment; and (3) the effectiveness of the model in action depends causally on the degree of structural similarity between the environment and the model.

When assuming a strong interpretation, it should be stated that the commitments to the inferential nature of perception and to inferentialism are closely related and result from the commitment to representationalism (Gładziejewski, 2017b, p. 106, 111; Kiefer, 2017).[43]

5. 2. Radical predictive processing

The approaches that can be included in the radical PP are much more diverse than the proposals under the conservative approach. It can be said that what is common to them is criticism or rejection of one or more commitments characterizing conservative PP. Radical PP is exemplified by the works (among others) of Andy Clark (2015a; 2016 etc.), Nico Orlandi (2016; 2018; Orlandi & Lee, 2018), Jelle Bruineberg, Julian Kiverstein and Erik Rietveld (Bruineberg 2017; Bruineberg, Kiverstein & Rietveld, 2018; Bruineberg & Rietveld, 2014) and Michael Kirchhoff (2018; Kirchhoff & Robertson, 2018). These authors emphasize that, firstly, action and perception stand in close relation to each other. Secondly, and more importantly in this context, they also claim that some levels of the hierarchical generative model are directly representational, whilst others are only indirectly so,[44] being related to the world in an enactive way, which means that representations "aim (is) to engage the world, rather than to depict it in some action-neutral fashion" (Clark, 2015b, p. 4).

Radical PP is inspired, on the one hand, by ecological psychologists developing ideas proposed by James J. Gibson (1966; 1979), and on the other hand by enactivists referring to the works of Francisco Varela and Humberto Maturana (Varela, Maturana & Uribe, 1974; Varela, Thomson & Rosch, 1992). Ecological approaches are characterized aptly by Pfeifer and Bongard by means of the so-called *Principle of Ecological Balance*. It states that, "first… that given a certain task environment there has to be a match between the complexities of the agent's sensory, motor, and neural systems… second... that there is a certain balance or task-distribution between morphology, materials, control, and environment" (2007, p. 123).

[43] The point is that the brain abductively "infers" about the causes of sensory stimulation (input sensors) in such a way that it presents hypotheses that best explain information coming from the environment (Kiefer 2017; see also Gregory 1980).

[44] Some researchers belonging to radical PP reject the existence of representation at all (cf. Kirchhoff & Robertson, 2018; Orlandi, 2016; 2018; van Es, 2020a).

Despite many controversies pointed out by various critics, the enactive tradition has many supporters also among researchers dealing with PP. Because of its programmatic anti-representationalism, it is often associated with and combined with ecological psychology and philosophy (cf. e.g., Bruineberg, Kiverstein & Rietveld, 2018; Bruineberg & Rietveld, 2014). However, one should keep in mind the differences between these two approaches.

Refeference		Conservative PP	Radical PP
Background	Helmholtz, Psychology „analysis-by-synthesis"	√	X
	Phenomenology; Ecological Psychology; Enactivism	X	√
Commitment to	Representationalism	√	X
	Inferentialism	√	X
	Internalism	√	X
Character of perception	Direct	X	√
	Indirect	√	X
Model of explanation	Mechanistic	√	X
	Others	√	√

Table 1. Comparison of conservative and radical PP due to their inspiration, theoretical commitments and the adopted model of explanation.

Some supporters of the radical approach (see, among others, Bruineberg, Kiverstein & Rietveld, 2018; Orlandi, 2016; 2018; van Es, 2020a) criticize the conservative approach because of its internalism and representationalism. They claim that internalism means a position that—in contrast to extended and embodied approaches to cognition—emphasizes that the limit within which one should think and study cognition is the limit of the nervous system. According to the strong interpretation,[45] conservative PP assumes some of the premises of methodological individualism, i.e., a position whereby the explanation of cognitive processes and phenomena is carried out by analyzing and explaining the cognitive mechanisms implemented in the brain. However, it can be concluded that the dispute is to some extent apparent. The brain (and the generative model in particular) comes to recapitulate the causal/statistical structure of the environment. This means that the environment plays an indirect but absolutely essential role in the psychological explanations offered by what is here called conservative PP.

5. 3. Internalism and Markov blankets

Many researchers claim that the boundaries of living systems are best marked by their Markov blankets (Hohwy, 2017; Friston, Wiese, Hobson, 2021; see also Hesp et al., 2019; Kirchhoff et al., 2018). In Bayesian networks, Markov blankets are described using the concepts of parents and children. Explaining a child's behavior requires addressing the behavior of his parent and other children. Therefore, it is not necessary to know the states of the dots that precede the child's parents, i.e., grandparents, great-grandparents, etc. In practice, this means that if you want to predict the condition of a given node, all you need to know are the states of the knots that make up his Markov blanket. The following illustration shows it well:

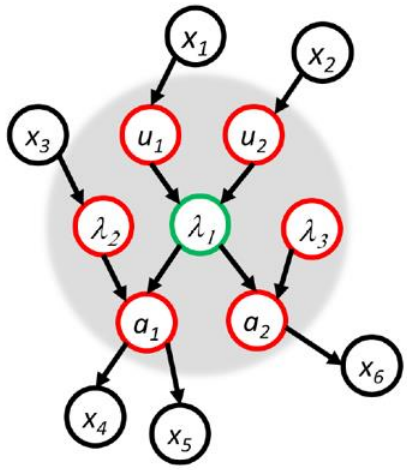

Fig. 2. DAG of Markov blanket (Hohwy, 2017, p. 3).

[45] Strong (realistic) conservative PP postulates that mental states contain real and causally efficacious representative content.

The Markov blanket (b) around internal states λ_n - where all other (external) variables are marked as x_n - is defined as a set of variables that make λ conditionally independent from external states. Mathematically, it is written as follows:

$$\boldsymbol{\lambda \perp x|k \Leftrightarrow P(\lambda, x|k) = P(\lambda|k)\, P(x|k)}$$

This equation illustrates the structure of the dependence in the factorization of the joint distribution conditioned on the states of the blanket and into two independent distributions; by definition, two variables are conditionally independent if, and only if, their joint probability, conditioned by some third variable, is equal to the product of their individual probability, conditioned by that third variable. There is often talk of random variables separated in this way by Markov blankets—and conditional dependencies related to them (in terms of parents and their children) (Hipólito et al., 2021, p. 90).

In this approach, the internal states of a given organism can be defined as Markov blankets that separate the system from its environment. These states are therefore independent (in a statistical sense) of environmental states. This means that internal states are designed to minimize prediction errors to maintain structural and functional integrity that leads to homeostasis and self-organization. The condition for minimizing prediction errors, resp. free energy is therefore a strict separation of the interior of the system from its surroundings, i.e., the ability of this system to distinguish itself from the environment (Friston, 2013b). Markov blankets enable statistical demarcation of system boundaries. A Markov blanket induces a statistical partitioning between internal (systemic) and external (environmental) states, the latter of which can be associated with neuronal, bodily, or worldly states depending on the relevant partitioning of the system in question. Blankets comprise a bipartition into active and sensory states.[46] They mediate exchanges between systemic and environmental (neuronal, bodily and so on) states (Ramstead et al., 2021, p. 4).[47] Hohwy claims that there is a fundamental difference between the internal known causes as they are inferred by the model and the external hidden causes on the other side of the Markov blanket. According to this, internalism means that all perceptual and cognitive processing happens within the internal model, or, equivalently, within the Markov blanket (Hohwy, 2017, p. 7). It should therefore be said, that in the conservative PP cognitive contact depends on where the Markov blanket is placed, its location being independent of any facts about the brain, patterns of neuronal activation and so on. For example, Hohwy's claim that the blanket is best placed around the brain is a conditional arrangement related to,

[46] This means that the understanding of Markov's blankets in PP and the active inference framework is far beyond Pearl's approach, in which there is no separation into active and passive states.

[47] It should be noticed that there is a divergence between ontic interpretations of Markov blankets in terms of Pearl "instrumental" blankets and Friston "realist" blankets (Bruineberg et al., 2020). Pearl interpretation does not go beyond pure formalism and Markov blankets themselves turn out to be a mathematical construct used for making inferences about the e.g. generative model. In Fristonian approach (used by many PP advocates) the properties of blankets are projected onto target systems (e.g. living organisms). Such an approach is not "philosophically innocent", as convincingly demonstrated by Bruineberg et al., 2020.

among others, the explanatory strategy, the research interests, and the phenomenon under study.[48] For this reason, it should be said that the conservative PP's internalism is not essentialist, but pragmatic.[49]

The rejection of essentialism regarding the boundaries of cognitive systems can lead to weakening or even negating the sharp distinction between internalism and externalism (see Ramstead et al., 2021), thereby weakening the distinction into conservative and radical PP. The above analyses lead to the statement that one can be an internalist in PP while not postulating neuro-representationalism; it all depends on where the Markov blanket will be placed. In other words, internalism does not have to imply and does not actually imply neuro-representationalism.[50] For this reason, one can agree with Gładziejewski, who emphasizes that a consistent reading of the conservative approach (even with a strong interpretation of the commitment to representations) includes it in the 4E approach to cognition.[51] Jakob Hohwy (2018) adds that this framework is able to better explain how embodied agents interact dynamically with the environment, and that the boundary between the mind and the world is on the one hand self-evidencing (Hohwy, 2014), which means that the causes of sensory stimuli are indirectly known by inference about the information coming from the sensory inputs, while on the other hand it is causal. There is dynamic feedback between the mind and the world, and this is made possible by perception and actions in the world.

6. Predictive processing and the free energy principle

It can be said that PP is a combination of two ideas (Gładziejewski, 2019; Klein, 2018): the first rests on the conviction that this framework is a set of various statements about what the brain is and how it works (cf. Clark, 2013; 2016; Hohwy, 2013, 2020a; Wiese, Metzinger, 2017); the second is that the basis of the brain or mind, or rather the whole human body, lies in the formal conception of theoretical biology called the free energy principle (cf. Friston,

[48] From the perspective of mechanistic explanation, determining the boundaries of a cognitive system is always dependent on a given research task and thus on how we want to explain such and such cognitive phenomenon, thereby indicating the mechanism responsible for it. Determining the limits of the cognitive system would then be carried out anew, without any prejudices as to its width or centralization around the brain, but using a mechanistic criterion to distinguish components of the system from contextual conditions (Wachowski, 2018, cf. Hutchins, 2014; Kaplan 2012).

[49] By essentialism according to Ramstead, I understand the belief according to which "there is a uniquely defining boundary or unit of analysis from which best to understand and investigate cognition" (Ramstead et al., 2019, p. 2).

[50] Zahavi accuses conservative PP neuro-representationalism (2018).

[51] He formulates a number of arguments for this claim. First, representations are not static images of reality, but internal, guiding actions or structural representations that allow the recognition of representational errors. They are modal in nature and their content is constrained by the way the body is embodied and embedded in the environment. Secondly, the key concept of inference for PP is liberal, which means that the representations have truthfulness conditions, and the way they are updated is active rather than reactive. Thirdly, the conceptual resources related to PP allow for an interesting connection between this approach and other existing 4E approaches (Gładziejewski, 2017b).

2009; 2010; 2012; Friston & Kiebel, 2009; Friston, Kilner & Harrison, 2006; Friston & Stephan, 2007; Hohwy, 2015a; Kiverstein, Sims; 2021; Smith, Friston & White, 2021). There is no consensus among researchers as to how these two ideas—let us call them architectural and homeostatic—are related to each other and whether the architectural approach must necessarily be considered in the context of the homeostatic approach. It seems natural to look for some fundamental biological principle that will explain the predictive nature of the human brain. Since human cognition can be explained by the Bayesian PP (cf. Clark, 2013; Howhy, 2013; 2015a), then, as some researchers claim, it should be possible to include the unification framework determined by this approach in one formal rule that is implemented or realized on subsequent levels of the generative model. This framework is supposed to be provided by FEP which expresses the homeostatic nature of living organisms. Therefore, if the architectural approach has a definite explanatory power, then we should look for its full justification in the biological, and even better, metaphysical nature of the objects explained. Following this lead, it should be stated that the explanatory power of the architectural approach comes from the explanatory and unifying power of FEP. Analyzing the above remarks, two conclusions come to mind: (1) the relationship between PP and FEP depends on the explanatory possibilities that we assign to the architectural approach; and (2) determining the dependence of the architectural approach on the homeostatic one implies a number of solutions regarding what FEP explains and how.

According to FEP, any self-organizing system that is at a non-equilibrium steady-state with its environment must minimize its (variational) free energy (Friston & Stephan, 2007; Friston, 2013b). In other words: any "thing" that achieves a non-equilibrium steady-state can be construed as performing a Bayesian inference with posterior beliefs that are parameterised by the thing's internal states. This means that this principle applies to all biological systems and is associated with their innate ability to counter the natural tendency to disorder (cf. Constant, 2021; Colombo, Palacios, 2021). If a system exists, its trajectory over phase space by definition satisfies a functional variational free energy. In line with the FEP, every living organism is an ergodic process, i.e., one for which statistics averaged over time are the same as its statistics averaged over the process. Internal states of the body can be defined in terms of Markov blankets (cf. Friston, Wiese & Hobson, 2020; Kirchhoff et al., 2018), which, in order to maintain their integrity and autonomy, must minimize free energy (in the information-theoretic sense) – an upper bound on surprise that is equivalent to the negative log probability of an outcome given a generative model (Friston, 2010). Heuristically speaking free energy is the difference between expectations about the world (how it is modelled) and what the world is like. Minimizing free energy is crucial for the organism because this mechanism reduces the degree of uncertainty resp. surprisal.[52] Proponents of applying FEP in PP suggest that the minimization of free energy should be directly associated with the Bayesian minimization of prediction errors.[53] FEP reduces surprise by tracking and minimizing discrepancies between data

[52] Surprisal is the negative logarithm of the probability of an event. In other words, it shows how unlikely a given event is in relation to the model of the world.

[53] Minimizing free energy entails minimizing prediction errors under the Laplace assumption (Friston, 2008; cf. Wiese & Metzinger, 2017, p. 12).

sent through the senses and top-coded predictions (this involves functional asymmetry of predictions and errors, which is based on the interaction of two types of pyramidal cells (cf. Clark, 2013; Park & Friston, 2013)), modifying their precision based on neuromodulatory mechanisms. In other words: free energy will be minimized when either the model parameters, i.e., beliefs (priors) and predictions about the sources and nature of sensory information (perceptual inference) are changed, or when there is a change in the environment, i.e., the causal structure of certain world states changes (by selectively sampling unsurprising sensations [cf. Badcock, Friston & Ramstead, 2019, p. 8; Friston, Daunizeau & Kiebel, 2009]). Minimizing free energy increases the model evidence, thereby reducing (long-term average) surprise, resp. prediction error.

Friston claims (2010, p. 129) that prediction errors can be minimized in two ways: (1) by passive (perceptual) inference, i.e., by revising the generative model (changing its internal parameters) and the hypotheses formulated about the statistical sensory signal; (2) through active inference, i.e., through such action in the world that will allow us to maintain the appropriate hypothesis formulated by the model in a way that makes "our predictions come true" (Clark, 2016, p. 121).[54] By making an active inference, the active agent interferes with the causal structure of relevant states of affairs. Here, active inference is understood, in a sense, as an action that minimizes uncertainty.[55] These types of action are adaptive because their goal is (in accordance with FEP) to maintain homeostasis. It can therefore be said that active inference is the use of FEP for action (Buckley et al., 2017).[56] It consists in the active agent wanting to stay alive by maintaining its homeostasis.

By combining information theory with the thesis on the adaptive nature of biological life in terms of Bayesian networks and the theory of optimal decision, FEP makes it possible to explain the nature of the decision-making processes, motor control and information costs (Friston, 2012, p. 2102), as well as many aspects of the anatomy and physiology of certain organisms (Friston, 2009, p. 295). Schematically, the implementation of FEP at individual levels in the life of organisms can be represented as follows:

[54] It should be added that perceptual inference defines limitations (by modifying relevant model parameters) for active inference (Hohwy, 2020b, p. 8).

[55] In newer approaches, active inference is associated with Markov blankets and variational neuroethology (cf. Ramstead et al., 2017). This approach can be described as "broad active inference" or as "active inference framework". The active inference framework is intended to provide a normative explanation of why agents must necessarily infer and minimize surprise by using their inner, latent states to maintain physiological integrity (Friston, 2012; Morville et al., 2018, p. 16). It should be added that the formalism of active inference is to provide an integrative framework for many normative approaches (Friston et al., 2017, p. 4) and is directly related to dynamic causal modeling (Friston, Harrison & Penny, 2003).

[56] This means that if FEP is the normative rule for all living organisms, then active inference is a normative framework explaining the optimal actions and behaviors of these organisms (cf. Schwartenbeck et al., 2013).

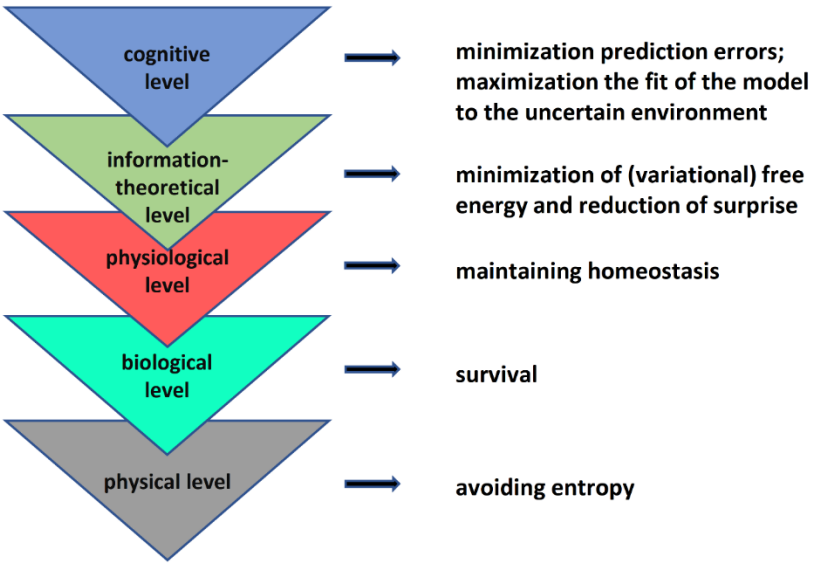

Fig. 3. The scope of FEP according to living organism.

The PP literature provides arguments for supporters of binding this research framework with FEP (cf. Friston & Kiebel, 2009; Hohwy, 2013; 2015; Ramstread, Kirchhoff & Friston, 2020; Bruineberg & Rietveld, 2014; Bruineberg, Kiverstein & Rietveld, 2018), its opponents (cf. Colombo & Wright, 2017; Orlandi 2016; van Es, 2020b) as well as authors showing ambivalent or agnostic attitude (cf. Clark, 2016; Gładziejewski, 2016; 2019; Sims, 2016; 2017). In this review, I will signal the problem of the homeostatic nature of predictive mechanisms, which seems to be crucial for the issues discussed here.

According to FEP, every living organism is an ergodic process, i.e., one for which statistics averaged over time are the same as its statistics averaged over the process (cf. Friston, 2013b). From the point of view of this type of biological system, it is beneficial to minimize free energy in order to maintain homeostasis understood as energy balance (in the information-theoretic sense, not thermodynamic) with the environment.[57] In the Fristonian approach, the mechanism of minimizing prediction errors is homeostatic and its function is to generate predictions (Friston & Kiebel, 2009). This means that it can be explained using FEP.[58] Godfrey-Smith, however, draws attention to a significant limitation associated with talking about homeostatic

[57] The concept of homeostasis comes from the work of the French physiologist Claud Bernard, who stated that the invariability of the internal environment is a condition of a free and self-contained life (1865).

[58] Colombo and Wright point out that it is unclear what kind of explanation FEP offers. According to these researchers, FEP is not compatible with either organicism or mechanisms. In their view, it should be noted that the inference leading to FEP takes the form of a transcendental argument, the conclusion of which is that FEP is a condition for the existence of adaptive systems, including information processing systems and non-biolo-

mechanisms. Namely: the actual homeostatic mechanisms consist in maintaining the stability of those organic properties that are not trivially associated with survival. That is, when a complex activity of the organism contributes to its survival, it is truly homeostatic only if there is some indirect organic property, bringing real contribution to survival, whose homeostasis is maintained through this complex activity. Thus, for example, the intelligent use of fire by people to maintain body heat is not a real case of homeostasis implemented by some cognitive mechanism. It is therefore a complex organic ability that allows man to maintain a constant body temperature in changing environmental conditions, thus enabling survival in these conditions. However, there is no reason to believe that everything that happens through cognition is trivially homeostatic. Therefore, if effective perception and coordination of activities enables the body, e.g., to avoid a predator, it is not a homeostatic mechanism even though it helps the organism maintain the basic integrity of the organic system. However, this would not be possible without the stability of some indirect organic properties, such as temperature and blood sugar levels. Cognitive mechanisms are adaptive, but an explanation of why they are this way requires reference to more basic organic homeostatic mechanisms (Godfrey-Smith, 1996, p. 79).

Godfrey-Smith's remark is crucial for analyzing homeostatic mechanisms as viewed according to Friston. It proves that one should be very careful when deciding that something serves to maintain homeostasis or is a homeostatic mechanism. [59] It is difficult to consider the mechanism of minimizing prediction errors as an actual homeostatic mechanism. This mechanism can ensure the survival and operation of an organism in the environment as long as it refers to organic homeostatic mechanisms. In this sense, minimization of prediction errors is a way to implement homeostasis. Seth and Friston claim: "cognitive processes are grounded in fundamental evolutionary imperatives to maintain physiological homeostasis" (Seth & Friston, 2016, p. 2). Therefore, it is justified to doubt the possibility of explaining predictive mechanisms as homeostatic mechanisms of a given type.[60]

gical systems such as social networks and artifacts (Friston, 2009, p. 293): "Systems that do not minimize free energy cannot exist" (Friston, 2013b, p. 2). Without going into the details of the analysis by Colombo and Wright, it must be said that the transcendental argument justifying FEP as the first principle in life sciences cannot be fully applied in any of the dominant views on the explanation of life. Cf. Andrews, 2021, where there is a response to the criticism of the FEP carried out, among others, by Colombo and Wright. On the normative nature of FEP as a regulatory idea and conceptual framework for life sciences, see Hohwy, 2020b.

[59] It is possible be that a cognitive mechanism does not to have biologically significant consequences, which means that it cannot be directly reduced to a more basic homeostatic mechanism. The sexual behavior of the praying mantis is a good example of this. The male mantis can see the female mantis better, but still falls prey to it during mating. It is therefore difficult to argue that the male mantis' perceptual mechanisms are indirectly homeostatic. The male's successful minimization of potential prediction error does not at all lead to maintaining the stability of its organism (cf. Lelito, Brown, 2006.)

[60] However, one can defend Friston's approach by claiming that only some mechanisms are homeostatic, while others are allostatic. Allostasis, or "stability through change" is the process of achieving homeostasis through physiological and behavioral changes. It is necessary to maintain the internal vitality of the organism in changing conditions (Sterling, Eyer, 1988). It can provide compensation in response to various problems of the organism (e.g., compensates for problems with the heart, kidneys or liver). Unfortunately, allostasis is unstable and com-

7. Limits of explanation in predictive processing

PP currently has a lot of supporters, but it also attracts important criticism. Some accuse PP of having weak empirical foundations in the sense that its biological and computational underpinnings are not clear. For this reason, PP actually does not offer homogeneous and systematic explanations, but redescriptions of already known phenomena (cf. Litwin & Miłkowski, 2020). Others (cf. Williams, 2019a; 2020) emphasize that the strong assumption of PP regarding the hierarchy of the generative model does not enable explaining thoughts in terms of both their generality and compositionality (Fodor, 1975). Another difficulty of PP is the general weakness of hierarchical models of perception when it comes to explaining the so-called perceptual organization (cf. van Leeuwen, 2015a; 2015b). I will not discuss these objections here further, but I will focus on the dark room problem, which, I claim, concerns the basic difficulty of PP associated with the explanatory monism of this framework. I will argue (cf. §7. 2) that strong explanatory PP is PP integrated with other models and approaches in cognitive science (cf. Colombo & Wright, 2021).

7. 1. The dark room problem

The dark room problem refers to the basic objection against PP (and sometimes FEP), which concerns the assumption that continuous minimization of prediction errors is optimal and necessary for agents acting in the world. The dark room problem can be expressed as follows: the dark room is a state in which the agent could find itself if it minimized the sum of all potential prediction errors so that no states of the world could surprise it.[61] The total absence of stimuli ensures optimal and maximum efficiency in minimizing any uncertainty. It seems that in accordance with the requirement of constant and long-term minimization of prediction errors, such a state should be desirable for the agent. But is that really the case?[62]

---

pensation can quickly end. Some FEP proponents argue that since homeostasis does not explain the rich diversity of regulatory processes, allostasis should be referred to. The latter enables proactive preparation of the organism for potential regulatory changes, thus contributing to minimizing free energy, resp. prediction errors (cf. Corcoran & Hohwy, 2019; Corcoran, Pezzullo & Hohwy, 2020; Kiverstein, Sims, 2021).

The proposed solution certainly allows for a better and more complete application of FEP to explain a number of phenomena. However, the same objection immediately arises that can be formulated with respect to homeostatic mechanisms: one should be very careful about stating that a mechanism is allostatic. Certainly, some of the introceptive mechanisms can be considered allostatic, but this does not mean that the generative model as such serves allostasis, i.e., it can be satisfactorily explained only by FEP.

[61] In another approach, an agent wishing to minimize the sum of all potential prediction errors should look for "a dark, unchanging chamber, and stay there" (Friston, Thornton, Clark, 2012).

Mumford: "In some sense, this is the state that the cortex is striving to achieve: perfect prediction of the world, like the oriental Nirvana (…) when nothing surprises you and new stimuli cause the merest ripple in your consciousness" (1992, p. 247, footnote 5).

[62] Clark had already expressed this doubt in the 2013 paper: "How can a neural imperative to minimize prediction error by enslaving perception, action, and attention accommodate the obvious fact that animals don't simply seek a nice dark room and stay in it? Surely staying still inside a darkened room would afford easy and nigh-perfect prediction of our own unfolding neural states?" (Clark, 2013, p. 191).

The dark room problem describes the clear contrast between the rich repertoire of real living creatures' behavior and the requirement to adjust all behaviors, actions, and decisions of the agent armed with the generative model to the requirement of minimizing prediction errors, resp. surprise and uncertainty (i.e., maximizing the model evidence). It is not so that living, embodied and embedded creatures are always looking for a niche that would meet the requirements of the dark room. Rather, the environment (not only physical, but also cultural and social) and other agents motivate us to take various actions, including those that do not minimize uncertainty, but on the contrary increase it. Is it the case then that PP as a theory that explains, if not all, at least some of the processes and phenomena of perception and cognition cannot take account of the fact that there are such actions and behaviors that do not always lead to minimizing prediction errors? Or maybe every action we take ultimately allows us to minimize surprise in the long term?[63]

In these considerations, I am analyzing some of the responses made by PP and FEP proponents regarding the dark room problem. Next, I will examine some criticism of these proposals and consider whether the attempts to solve the dilemma that has been implied by the dark room problem to date are satisfactory.

In one of the first papers on the dark room problem, Friston, Thornton and Clark (2012) state that the dark room is not a completely fictitious idea, because there are troglophiles, i.e., creatures (Dark-Room agents) that are evolutionarily adapted to life and navigation in dark places such as caves, abandoned bunkers, or underground waters. Therefore, from the perspective of the dilemma discussed here, one can ask why such animals exist? The answer is simple. Every living organism strives for such a minimization of free energy, or uncertainty, which consists in changing the sensory signal so that it can carry out actions that correspond to its predictions and the model of the world. In this sense, each such organism is a real (i.e., evolutionary) solution to the problem of minimizing surprise in a changing world. The evolutionary and developmental history of troglophiles is different from that of man. That is why people do not live in "dark rooms" as troglophiles. This means that the dark room provides a low level of surprise only if the agent has been optimized by evolution to stay and act in it (Ramstead, Kirchhoff & Friston, 2020, p. 230).

However, one may ask what happens when an agent such as a human enters into the dark room? Is it not the case that being in such a room reduces the level of uncertainty and surprise?[64] Clark believes it is not, because animals like us live and function in a world that is

[63]The dark room problem is directly linked to the exploration-exploitation trade-off. It is a notion derived from machine learning research, but it has wider application. Generally speaking, this trade-off concerns situations in which one chooses between what is known, what can be foreseen (it can meet our expectations) (exploitation) and what is not certain, but there is a strong supposition that it offers some novelty—in the form of information, experience or skills (exploration) (cf. Jasrasaria & Pyzer-Knapp, 2018).

[64] It may also be that being in such a room could be very predictable yet highly surprising as it does not match our expectations of how events unfold: agents usually expect and predict novel information. (I thank the reviewer for this remark).

Sims (2017) suggests that the questions related to the dark room can be reduced to the following: if actions are only a minimization of surprise, why are we not trying to minimize all possible stimuli?

constantly changing and requires something from us (cf. Clark, 2018). Therefore, we expect something to happen constantly: "Change, motion, exploration, and search are themselves valuable for creatures living in worlds where resources are unevenly spread and new threats and opportunities continuously arise" (Clark, 2013, p. 193). It is hard to disagree with Clark, but his answer raises a question that concerns the mechanism regulating the choice between exploitation (staying in the dark room) and exploration (leaving the dark room). Just because we know we are going to leave does not mean we have a good explanation for it.

Friston provides another answer. In his opinion, the first thing we do after entering a dark room is to turn on the light. Lighting up is an action that minimizes uncertainty (i.e., expected free energy). Namely: it is not that, entering a dark room, we expect it to be dark. Rather, we expect it to be bright. Therefore, minimizing uncertainty is associated with the expectation that a dark room can be illuminated, not that it will remain dark (Friston, Fortier & Friedman, 2018, p. 26). Friston's answer, like the one formulated by Clark, reveals a trade-off that interests us rather than explains it.

Pezzulo Rigoli and Friston (2015, p. 32) suggests that the FEP perspective on living organisms (homeostatic approach) explains why agents have no problem with the "dark room". In their opinion, homeostatic regulation implies constant updating of empirical priors of action, whose dynamics are dictated by an uninterrupted stream of interoceptive messages flowing between the brain and the body. This stream constantly provides information about new values and goals. Therefore, every organism equipped with the body must face reality, not live in a darkened room. This view explains the reasons why agents avoid dark rooms much better, but it does not explain completely how something present in the environment becomes a value or a goal for them. In other words, what mechanism makes the environment a carrier of value for the agent? Why does the agent see environment as having some meaning?[65]

An interesting critique of PP and FEP was carried out by Klein in the paper *What do predictive coders want?* In his approach, the whole dark room problem can be reduced to a question about motivation: "how motivating states like desires can exist in a predictive coding framework?" (Klein, 2018, p. 3; see also Sun & Firestone, 2020).[66] It is important that motivation cannot be reduced to predictions. Klein believes that the dark room problem demands an explanation of why and how we operate by minimizing uncertainty. Hohwy's thesis that we don't have to minimize every prediction error, but the average prediction error given over longer periods of time (being in a dark room makes that impossible) does not seem to explain much (Hohwy, 2013, p. 85, 175).[67] Hohwy suggests: let us imagine that we mistakenly perceive dogs as sheep. In the long term, the inability to distinguish between dogs and sheep will increase

[65]It should be added that the dark room problem raises concerns not only among the critics of PP, but also among its supporters, especially those who associate it with the explanatory function of FEP (cf. Kiverstein, Miller & Rietveld, 2019).

[66] On the concept of motivation in PP from a different perspective cf. Miller Tate, 2019.

[67] The following issue should be associated with the averaged prediction error (in other words, weighted sum of prediction errors): due to the lack of access to the causes of the sensory signal, it cannot be minimized directly, so the generative model performs its action by generating predictions that relate to averaged long-term surprise (Hohwy, 2013, p. 85).

the prediction error, i.e., it distances us from our expected states. It is analogous with the dark room. However, the lack of an indication of a phenomenon or mechanism that would explain our motivation to leave a dark room is problematic. Klein convincingly shows that a creature living in a dark room has smaller prediction errors than creatures living in other environments, so its average error is much smaller compared to those other creatures. Relying on the certainty of death does not help much, because, in the long-term perspective, each of us will die. Therefore, the question is not whether we avoid (or not) the moment of death, but why some of us delay this moment (leaving the dark room) while others do not. Predictions alone are not enough for us to take specific adaptation actions.[68]

7. 2. The postulate of integration

Many researchers claim (cf. Allen & Friston, 2018; Friston, 2009; 2010; Hohwy, 2015; Smith, Friston & White, 2021) that FEP offers an explanatory framework that makes it possible to understand why the requirement of minimizing prediction errors is justified by the fact that every living organism strives for homeostasis.[69] In this sense, FEP would be a normative framework (cf. Adams, Brown & Friston, 2014; Allen & Friston, 2018; Friston, 2013a; Friston, Fortier & Friedman, 2018). The arguments advanced by the proponents of such a solution are based on the fact that PP itself does not offer any optimal theory of decisions or learning algorithms etc. Considering such argument, (1) it can be concluded that PP's explanatory power refers only to some of brain research, or some studies on perception (cf. Sims, 2017); or (2) state that PP should be based on some biologically reliable theory that embeds the requirement of minimizing prediction error in a broader context (cf. Hohwy, 2015); or (3) claim that PP provides a general mechanistic framework or, more precisely, a sketch of the mechanism that should be supplemented by some adequate theory of representation, decision-making processes, biological functions, etc. (cf. Gładziejewski, 2019; Harkness, 2015; Hohwy, 2015).

These questions are directly related to the belief of many researchers that a full explanation of cognitive phenomena in cognitive science should be based in many cases on the mechanistic

[68] Van de Cruys, Friston and Clark (2020) argue that the objection of the inadequacy of the predictions themselves (and indirectly the problem of explaining motivation) can be answered by saying that the predictions are "optimistic." What does that mean? Generative models must be optimistically biased because the probability distribution of the expected states is realized only when we interact with the world. This means that the predictions are, on the one hand, allostatically built into the organization of a given organism, and on the other hand, that they must be as realistically world-oriented as possible, since they concern the actions of the organisms themselves and their anticipated consequences.

It should be stated that a full analysis of the responses that FEP and active inference supporters formulate to the problem of motivation and the exploration-exploitation trade-off is far beyond the scope of this discussion. It should be noted, however, that such an analysis is difficult due to the constant modification and expansion of the initial proposals under this research framework. Therefore, this presentation is largely limited and should be treated as a characteristic of some general tendencies present in this general framework.

[69] However, it is not entirely clear what the explanation is because FEP does not seem like a causal-etiological or causal-mechanistic explanation (cf. Colombo & Wright, 2021; Klein, 2018). Some studies suggest that it may be a dynamic explanation due to the fact that FEP and active inference are associated with Dynamic Causal Modelling (Friston, Harrison & Penny, 2003) (cf. Hipólito et al., 2020; Ramstead, Badcock & Friston, 2018).

integration of individual theories or models (Cf. Darden & Maull, 1977; Miłkowski, 2016a; 2016b; Mitchel, 2003; Povich, 2019). What binds individual models or theories together is the requirement to identify and describe mechanisms. The advantage of this approach over others is (1) that it provides a simple way of talking about levels; (2) offers much more insight into what integration between levels is based on specific constraints by which relationships between levels are assessed, as well as the coevolution of work at different levels, their cause-effect interactions, and spatial, temporal and hierarchical organization; and (3) is based on the need to explain phenomena both bottom-up and top-down taking into account the context in which a phenomenon is embedded. Integration is therefore an attempt to see how phenomena at many different levels are interrelated (Craver & Tabery, 2019).[70]

Craver (2007) and Miłkowski (2016a; 2016b) explicitly state that integration in neurosciences can be described in terms of space constraints of possible mechanisms. This approach can be enriched by the perspective proposed by Danks (2014, p. 31): "At a high level, one theory *S* constrains another theory *T* if the extent to which *S* has some theoretical virtue *V* (e.g., truth, predictive accuracy, explanatory power) matters for the extent to which *T* has *V*. More colloquially, *S* constrains *T* just when, if we care about *T* along some dimension, then we should also care about *S* along that same dimension." In this approach, virtue marked a specific constraint.

The explanatory power of constraints in explanations is conditioned by the actual causal role that these constraints play in *such and such* phenomena that we want to explain. This means that a particular phenomenon is explained by reference to a component, which is *such and such* constraint, because the mechanisms responsible for the implementation of this phenomenon contain *such and such* constraint themselves. I claim that this explanation necessarily involves integration with other models that may offer satisfactory explanations for the constraints for these mechanisms when investigating predictive mechanisms.

It seems that the difficulty of solving the dark room problem and exploration-exploitation trade-off is directly related to the explanatory monism adopted by many researchers, which in this case is based on the belief that PP is a sufficient framework to explain many phenomena. It should therefore be said that PP's explanatory power can be measured by the possibility of integrating it with other models and approaches (cf. Colombo & Wright, 2021; Klein, 2018; Williams, 2020).

8. Conclusions

The aim of this review was a systematic presentation of PP, taking into account its basic theoretical difficulties. It discussed the main concepts, positions, and research issues present within this framework (§1-2). I presented the Bayesian-brain thesis and the difficulty associated with it as highlighted by many researchers. The notion of internalism in PP and the role of Markov blankets were discussed in the context of the discussion between conservative and radical PP. In §6, I indicated the possibility of linking PP and FEP advocated by many authors. I also

[70] The historical examples of this approach include the discovery of protein synthesis (Darden, 2006) or cell biology (Bechtel, 2006).

discussed some problems concerning the understanding of predictive mechanisms as homeostatic. §7 presented some of PP's difficulties with solving the dark room problem and the exploration-exploitation trade-off. I emphasized the need to integrate PP with other models and research frameworks within cognitive science.

It should be noted that PP has a number of important philosophical consequences, namely supporters of PP describe cognitive mechanisms that are rich in epistemologically relevant ideas and concepts (cf. Ghijsen 2021; Gładziejewski 2017; 2021; Munton 2017) and refer to important intuitions in the philosophy of science (cf. Beni, 2018; Hohwy, 2013; 2016; Wiese, 2015). Epistemology is oriented towards the normative side of cognition, thus emphasizing the importance of justifications and seeking reason. These normative issues are still being developed within the research on PP and FEP. (cf. Constant et al., 2019; Hesp at al., 2019; Hohwy, 2020b; Piekarski, 2019). However, there remain a number of epistemological problems unexplored from the PP perspective, such as the issue of rationality, foundationalism or disputes, realism/anti-realism, or internalism/externalism. Still unsolved is the problem of the criterion that makes it possible to distinguish between personal and sub-personal states in PP, problems of consciousness and phenomenology (cf. Clark, 2017; 2019; Dołęga & Dewhurst, 2021; Hohwy & Seth, 2021; Hutto 2017; Marvan & Havlík, 2020; Nave, 2021; Schlicht & Dołęga, 2021; Wilkinson, 2014). These are promising research areas. It should be remembered, however, that non-trivial PP must be critical of its own assumptions and theoretical commitments, which, as I have briefly demonstrated, have a number of epistemic and explanatory consequences. This should be remembered at all times when developing this research ramework.

Acknowledgements: I would like to thank Krzysztof Dołęga, Paweł Gładziejewski, Marcin Miłkowski, Przemysław Nowakowski and two anonymous reviewers for their thoughtful remarks, comments and help in presenting these ideas. Special thanks are due to the editors of the "Avant" journal for inviting me to write this review.

References

Adams, R. A., Brown, H. R. & Friston K. J. (2014). Bayesian inference, predictive coding and delusions. *Avant*, 3(5), 51-88. DOI: 10.26913/50302014.0112.0004.

Adams R. A., Stephan K. E., Brown H. R., Frith C. D. & Friston K. J. (2013) The computational anatomy of psychosis. *Front. Psychiatry* 4:47. DOI: 10.3389/fpsyt.2013.00047/.

Anderson, M. L. & Chemero, T. (2013). The problem with brain GUTs: conflation of different senses of "prediction'' threatens metaphysical disaster. *Behavioral and Brain Sciences*, 36 (3), 204–205.

Andrews, M. (2021). The math is not the territory: navigating the free energy principle. *Biol Philos*, 36, 30. https://doi.org/10.1007/s10539-021-09807-0.

Allen, M. & Friston, K. J. (2018). From cognitivism to autopoiesis: towards a computational framework for the embodied mind. *Synthese*, 195(6), 2459–2482. DOI: 10.1007/s11229-016-1288-5 3.

Badcock, P. B. Friston, K. J. & Ramstead, M. J. D. (2019).The hierarchically mechanistic mind: A free-energy formulation of the human psyche. *Physics of Life Reviews*. DOI: 10.1016/j.plrev.2018.10.002.

Bechtel, W. (2006). *Discovering Cell Mechanisms: The Creation of Modern Cell Biology*. Cambridge: Cambridge University Press.

Beni, M. D. (2018). The reward of unification: A realist reading of the predictive processing theory. *New Ideas in Psychology*, 48, 21-26.

Bernard, C. (1865). *Introduction à l'étude de la médecine expérimentale*. Paris: Éditions Garnier-Flammarion.

Bowers J. S. & Davis C. J. (2012). Bayesian just-so stories in psychology and neuroscience. *Psychol Bull*. 138(3), 389-414. DOI: 10.1037/a0026450.

Brown, L. D. (1981). A Complete Class Theorem for Statistical Problems with Finite-Sample Spaces. *Ann Stat*, 9, 1289-1300.

Bruineberg, J. (2017). Active Inference and the Primacy of the 'I Can'. In T. Metzinger & W. Wiese (Eds.). *Philosophy and Predictive Processing: 5* (pp. 1-18). Frankfurt am Main: MIND Group. DOI: 10.155027/9783958573062.

Bruineberg, J., Dołęga, K., Dewhurst, J. & Baltieri, M. (2020) *The Emperor's New Markov Blankets*. [Preprint]

Bruineberg, J. & Rietveld, E. (2014). Self-organization, free energy minimization, and optimal grip on a field of affordances. *Frontiers in Human Neuroscience*, 8, 599.

Bruineberg, J., Kiverstein, J. & Rietveld, E. (2018). The anticipating brain is not a scientist: the free-energy principle from an ecological-enactive perspective. *Synthese*, 195, 2417–2444. DOI: 10.1007/s11229-016-1239-1.

Buckley, C. L., Sub Kim, C., McGregor, S. & Seth, A. K. (2017). The free energy principle for action and perception: A mathematical review. *Journal of Mathematical Psychology*, 81, 55-79.

Calvo, P. & Friston, K. J. (2017). Predicting green: really radical (plant) predictive processing. *J R Soc Interface*, 14: 20170096. https://doi.org/10.1098/rsif.2017.0096.

Chater, N., Oaksford, M., Hanh, U. & Heit, E. (2010). Bayesian models of cognition. *Wiley Interdiscip Rev Cogn Sci*., 1(6), 811-823. DOI: 10.1002/wcs.79.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36, 181–204. DOI: 10.1017/S0140525X12000477.

Clark, A. (2015a). Radical predictive processing. *The Southern Journal of Philosophy*, 53 (S1), 3–27.

Clark, A. (2015b). Predicting Peace: The End of the Representation Wars-A Reply to Michael Madary. In T. Metzinger & J. M. Windt (Eds.). *Open MIND: 7(R)* (pp. 1-7). Frankfurt am Main: MIND Group. DOI: 10.15502/9783958570979

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

Clark, A. (2017). Predictions, precision, and agentive attention. *Consciousness and Cognition*, 56, 115-119.

Clark, A. (2018). A nice surprise? Predictive processing and the active pursuit of novelty. *Phenom Cogn Sci*, 17, 521-534. DOI: 10.1007/s11097-017-9525-z.

Clark, A. (2019). Beyond desire? Agency, choice, and the predictive mind. *Australasian Journal of Philosophy*, 1-15. DOI: 10.1080/00048402.2019.1602661.

Colombo, M., Elkin, E. & Hartmann, S. (2018). Being Realist about Bayes, and the Predictive Processing Theory of Mind. *The British Journal for the Philosophy of Science*, axy059, DOI: 10.1093/bjps/axy059.

Colombo, M., Palacios, P. (2021). Non-equilibrium thermodynamics and the free energy principle in biology. *Biol Philos*, 36, 41 (2021). https://doi.org/10.1007/s10539-021-09818-x.

Colombo, M. & Seriès, P. (2012). Bayes in the brain–On Bayesian modelling in neuroscience. *British Journal for the Philosophy of Science*, 63 (3), 697–723.

Colombo, M. & Wright, C. (2021). First principles in the life sciences: The free-energy principle, organicism, and mechanism. *Synthese*, 198, 3463–3488. DOI: 10.1007/ s11229-018- 01932-w.

Conant, R. & Ashby, W. R. (1970). Every good regulator of a system must be a model of that system. *International Journal of Systems Science*, 1 (2), 89-97.

Constant, A. (2021). The free energy principle: it's not about what it takes, it's about what took you there. *Biol Philos*, 36, 10. https://doi.org/10.1007/s10539-021-09787-1

Constant, A., Ramstead, M. J. D., Veissière, S. P. L. & Friston, K. J. (2019). Regimes of Expectations: An Active Inference Model of Social Conformity and Human Decision Making. *Front. Psychol*. 10: 679. 1-15. DOI: 10.3389/fpsyg.2019.00679.

Corcoran, A. W., & Hohwy, J. (2019). Allostasis, interoception, and the free energy principle: Feeling our way forward. In M. Tsarkiris, & H. De Preester (Eds.). *The Interoceptive Mind: From homeostasis to awareness* (pp. 272-292). Oxford: Oxford University Press.

Corcoran, A.W., Pezzulo, G. & Hohwy, J. (2020). From Allostatic Agents to Counterfactual Cognisers: Active Inference, Biological Regulation, and The Origins of Cognition. *Biol Philos*, 35, 32. https://doi.org/10.1007/s10539-020-09746-2

Cosmides, L. & Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition*, 58, 1-73.

Craver C. F. (2007). *Explaining the brain*. Oxford: University Press Oxford

Craver, C. F. & Tabery, J. (2019). Mechanisms in Science. In E. N. Zalta (Ed.). *The Stanford Encyclopedia of Philosophy (Summer 2019 Edition)*. URL = <https://plato.stanford.edu/archives/sum2019/entries/science-mechanisms/>. access 20.03.2020.

Danks, D. (2014). *Unifying the mind: Cognitive representations as graphical models*. Cambridge, MA: The MIT Press.

Darden, L. (2006). *Reasoning in Biological Discoveries*, Cambridge, MA: Cambridge University Press.

Darden, L. & Maull, N. (1977). Interfield Theories. *Philosophy of Science*, 44, 43–64.

Dayan, P., Hinton, G. E., Neal, R. M. & Zemel, R. S. (1995). The Helmholtz machine. *Neural Comput*, 7, 889-904.

Dołęga, K. (2017). Moderate Predictive Processing. In T. Metzinger & W. Wiese (Eds.). *Philosophy and Predictive Processing: 10* (pp. 1-19). Frankfurt am Main: MIND Group. DOI: 10.15502/9783958573116.

Dołęga, K. & Dewhurst, J. (2021). Fame in the predictive brain: a deflationary approach to explaining consciousness in the prediction error minimization framework. *Synthese*, 198, 7781–7806. DOI: 10.1007/s11229-020-02548-9.

Easwaran, K. (2011). Bayesianism II: Applications and Criticisms. *Philosophy Compass*, 6(5), 321-332. DOI: 10.1111/j.1747-9991.2011.00398.x

Egan, F. (2014). How to think about mental content. *Philos Stud* 170:115–135. DOI: 10.1007/s11098-013- 0172-0.

Elqayam, S. & Evans J. S. (2011). Subtracting "ought" from "is": descriptivism versus normativism in the study of human thinking. *Behav Brain Sci*, 34(5), 233-248. DOI: 10.1017/S0140525X1100001X.

Eva, B. & Hartman, S. (2020). On the Origins of Old Evidence. *Australasian Journal of Philosophy*, 3(98), 481-494. https://doi.org/10.1080/00048402.2019.1658210.

Feldman, H. & Friston, K. J. (2010) Attention, uncertainty and free-energy. *Front Hum Neurosci*, 4(215), DOI: 10.3389/fnhum.2010.00215.

Felleman, D. & Van Essen, D. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cereb Cortex*, 1(1), 1-47.

Fink, S. B. & Zednik, C. (2017). Meeting in the Dark Room: Bayesian Rational Analysis and Hierarchical Predictive Coding. In T. Metzinger & W. Wiese (Eds.). *Philosophy and Predictive Processing: 14*, (pp. 1-13). Frankfurt am Main: MIND Group. DOI: 10.155027/9783958573062.DOI: 10.15502/9783958573154.

Fitelson, B. (2006). The Paradox of Confirmation. *Philosophy Compass*, 1(1), 95-113. DOI: 10.1111/j.1747-9991.2006.00011.x.

Friston, K. J. (2003). Learning and inference in the brain. *Neural Networks*, 16. 1325–1352.

Friston, K. J. (2005). A theory of cortical responses. *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci*, 360, 815–836. DOI: 10.1098/rstb.2005.1622.

Friston, K. J. (2008). Hierarchical models in the brain. *PLoS Computational Biology*, 4 (11), e1000211. https://dx.doi. org/10.1371/journal.pcbi.1000211.

Friston, K. J. (2009). The free-energy principle: a rough guide to the brain? *Trends Cogn Sci*, 13(7), 293-301.

Friston, K. J. (2010). The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci* 11, 127–138.

Friston, K. J. (2011). What is optimal about motor control? *Neuron*, 72(3), 488–498.

Friston, K. J. (2012). A Free Energy Principle for Biological Systems. *Entropy*, 14, 2100-2121. DOI:10.3390/e14112100.

Friston, K. J. (2013a). Active inference and free energy. *Behav Brain Sci*, 36(3), 212-213. DOI: 10.1017/S0140525X12002142.

Friston K. J. (2013b). Life as we. know it. *J R Soc Interface*, 10: 20130475. DOI: 10.1098/rsif.2013.0475.

Friston, K. J., Daunizeau, J. & Kiebel, S. J. (2009). Reinforcement learning or active inference? *PLoS ONE*, 4 (2009), e6421. DOI: 10.1371/journal.pone.0006421.

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2017). Active inference: A process theory. *Neural Computation*, *29*(1), 1–49.

Friston, K. J., Fortier, M. & Friedman, D. A. (2018). Of woodlice and men: A Bayesian account of cognition, life and consciousness. An interview with Karl Friston. *ALIUS Bulletin*, 2, 17–43.

Friston, K. J., Harrison, L. & Penny, W. (2003) Dynamic causal modelling. *Neuroimage*, 19. 1273–1302. DOI: 10.1016/S1053-8119(03)00202-7 pmid:12948688.

Friston, K. J. & Kiebel, S. J. (2009). Predictive coding under the free-energy principle. *Phil. Trans. R. Soc. B*, 364, 1211-1221.

Friston, K. J., Kilner, J. & Harrison, L. (2006). A free energy principle for the brain. *J Physiol Paris*, 100(1-3), 70-87.

Friston, K. J., Mattout, J., Trujillo-Barreto, N., Ashburner, J. & Penny, W. (2006). Variational free energy and the Laplace approximation. *Neuroimage*, 34 (1), 220-234.

Friston, K. J. & Stephan, K. E. (2007). Free energy and the brain. *Synthese*, 159, 417-458. DOI: 10.1007/s11229-007-9237-y.

Friston, K., Thornton, C. & Clark, A. (2012). Free-energy minimization and the dark-room problem. *Front. Psychol.* 3:130. DOI: 10.3389/fpsyg.2012.00130.

Friston, K. J., Wiese, W. & Hobson, J. A. (2020). Sentience and the origins of consciousness: From cartesian duality to Markovian monism. *Entropy*, (22), 516–516. https://doi.org/10.3390/E22050516.

Ghijsen, H. (2021). Predictive processing and foundationalism about perception. *Synthese*, 198, 1751–1769. DOI: 10.1007/s11229-018-1715-x.

Gibson, J. J. (1966). *The Senses Considered as Perceptual Systems*. Boston: Houghton Mifflin.

Gibson, J. J. (1979). *The Ecological Approach to Visual Perception*. New York: Psychology Press.

Garfinkel, S. N., Seth, A. K., Barrett, A. B., Suzuki, K. & Critchley, H. D. (2015). Knowing your own heart: Distinguishing interoceptive accuracy from interoceptive awareness. *Biological Psychology*, 104, 65–74. https://doi.org/10.1016/j.biopsycho.2014.11.004.

Gigerenzer, G. & Brighton, H. (2009). Homo Heuristicus: Why Biased Minds Make Better Inferences. *Topics in Cognitive Science*, 1(1), 107-143. DOI: 10.1111/j.1756-8765.2008.01006.x.

Glymour, C. (1980). *Theory and Evidence*. Princeton: Princeton University Press.

Glymour, C. (2001). *The Mind's Arrows: Bayes Nets and Graphical Causal Models in Psychology*. Cambridge, MA: MIT Press.

Gładziejewski P (2016) Predictive coding and representationalism. *Synthese*, 193, 559–582. DOI: 10.1007/ s11229-015-0762-9

Gładziejewski, P. (2017a). The Evidence of the Senses - A Predictive Processing-Based Take on the Sellarsian Dilemma. In T. Metzinger & W. Wiese (Eds.). *Philosophy and Predictive Processing: 15* (pp. 1-15). Frankfurt am Main: MIND Group. DOI: 10.15502/9783958573161.

Gładziejewski, P. (2017b). Just how conservative is conservative predictive processing? *Hybris* 38, 98-122.

Gładziejewski, P (2019). Mechanistic unity and the predictive mind. *Theory & Psychology*, 29(5), 657–675. DOI: 10.1177/0959354319866258.

Gładziejewski, P. (2021) Perceptual justification in the Bayesian brain: a foundherentist account. *Synthese*. 1-25. First online. https://doi.org/10.1007/s11229-021-03295-1.

Gładziejewski, P. & Miłkowski, M. (2017) Structural representations: causally relevant and different from detectors. *Biol Philos*, 32, 337–355. DOI 10.1007/s10539-017-9562-6.

Godfrey-Smith, P. (1996). *Complexity and the Function of Mind in Nature*. Cambridge University Press, 1996.

Gottlieb, J. (2012). Attention, learning, and the value of information. *Neuron*, 76, 281–295. DOI: 10.1016/j.neuron.2012.09.034.

Gregory, R. L. (1980). Perceptions as hypotheses. *Phil. Trans. R. Soc. Lond., Series B, Biological Sciences*, 290, 181-197.

Griffiths, T. L., Chater, N., Norris, N. & Pouget, A. (2012). How the Bayesians got their beliefs (and what those beliefs actually are): comment on Bowers and Davis (2012). *Psychol Bull*. 138(3), 415-422. DOI: 10.1037/a0026884.

Gordon, N., Tsuchiya, N., Koenig-Robert, R. & Hohwy, J. (2018). Expectation and attention increase the integration of top-down and bottom-up signals in perception through different pathways. b*ioRxiv*, 446948.

Harman, G. (1973). *Thought*. Princeton: Princeton University Press.

Harkness, D. L. (2015). From Explanatory Ambition to Explanatory Power - A Commentary on Jakob Hohwy. In T. Metzinger & J. M. Windt (Eds.). *Open MIND: 19(C)* (pp. 1-9). Frankfurt am Main: MIND Group. DOI: 10.15502/9783958570153.

Harkness, D. L. & Keshava, A. (2017). Moving from the What to the How and Where – Bayesian Models and Predictive Processing. In T. Metzinger & W. Wiese (Eds.). *Philosophy and Predictive Processing: 5* (pp. 1-10). Frankfurt am Main: MIND Group. DOI: 10.15502/9783958573178.

Harrison, C. W. (1952). Experiments with linear prediction in television. *Bell System Technical Journal*, 31(4), 764–783.

Helmholtz, H. v. (1867). *Handbuch der Physiologishen Optik*. Leipzig: Leopold Voss.

Hempel, C. (1965). *Aspects of Scientific Explanation*. New York, NY: Free Press.

Hesp, C., Ramstead, M. J. D., Constant, A., Badcock, P., Kirchhoff, M. & Friston, K. J. (2019). A Multi-scale View of the Emergent Complexity of Life: A Free-Energy Proposal. In G. Georgiev,

J. Smart, C. Flores Martinez & M. Price (Eds.). *Evolution, Development, and Complexity: Multiscale Models in Complex Adaptive Systems* (pp. 195-227). Springer Proceedings in Complexity. DOI: 10.1007/978-3-030-00075-2_7.

Hinton, G. E., Sejnowski, T. J. (1983) Optimal perceptual inference. *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. (Washington DC).

Hipólito, I., Ramstead, M. J. D., Convertino, L., Bhat, A., Friston, K. J. & Parr, T. (2021). Markov Blankets in the Brain. *Neuroscience & Biobehavioral Reviews*, 125, 88-97. https://doi.org/10.1016/j.neubiorev.2021.02.003.

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

Hohwy, J. (2014). The self-evidencing brain. *Noûs*, 50(2), 259-285.

Hohwy, J. (2015a). The Neural Organ Explains the Mind. In T. Metzinger & J. M. Windt (Eds.). *Open MIND: 19(T)* (pp. 1-22). Frankfurt am Main: MIND Group. DOI: 10.15502/9783958570016.

Hohwy, J. (2015b). The Diversity of Bayesian Explanation - A Reply to Dominic L. Harkness. In T. Metzinger & J. M. Windt (Eds.). *Open MIND: 19(R)* (pp. 1-6). Frankfurt am Main: MIND Group. DOI: 10.15502/9783958570870.

Hohwy, J. (2018). The Predictive Processing Hypothesis. In A. Newen, L. S. de Bruin & S. Gallagher (Eds.). *The Oxford Handbook of 4E Cognition* (pp. 129-145). Oxford: Oxford University Press.

Hohwy, J. (2020a). New directions in predictive processing. *Mind & Language*. 2(35), 209-223. DOI: 10.1111/mila.12281.

Hohwy, J. (2020b). Self-supervision, normativity and the free energy principle. *Synthese*, 1-25. First online. DOI: 10.1007/s11229-020-02622-2.

Hohwy, J. & Seth, A. K. (2020). Predictive processing as a systematic basis for identifying the neural correlates of consciousness. *Philosophy and the Mind Sciences*, 1(II). https://doi.org/10.33735/phimisci.2020.II.64.

Howson, C. (1991). The 'Old Evidence' Problem. *Brit. J. Phil. Sci*., 42, 547-555.

Hutchins, E. (2014). The cultural ecosystem of human cognition. *Philos. Psychol*. 27, 34–49. DOI: 10.1080/09515089.2013.830548.

Hutto, D. D. (2017). Getting into predictive processing's great guessing game: Bootstrap heaven or hell?. *Synthese*, 195, 2445–2458, DOI: 10.1007/s11229-017-1385-0.

Jasrasaria, D. & Pyzer-Knapp, E. (2018). Dynamic Control of Explore/Exploit Trade-Off In Bayesian Optimization. *arXiv*:1807.01279v1.

Jensen F. V. (2006) *An Introduction to Bayesian Networks*. London: UCL Press.

Jones, M. & Love, B. C. (2011). Bayesian Fundamentalism or Enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and brain sciences*, 34, 169–231. DOI: 10.1017/S0140525X10003134.

Kaplan, D. (2012). How to demarcate the boundaries of cognition. *Biology and Philosophy*, 27(4), 545–570. DOI: 10.1016/j.shpsa.2018.05.013.

Kawalec, P. (2003). Zagadnienia metodologiczne w bayesowskiej teorii konfirmacji (*Methodological issues in Bayesian confirmation theory*). *Roczniki filozoficzne*, 1(51), 113-142.

Kelso, J. S. (2012). Multistability and metastability: understanding dynamic coordination in the brain. *Philos. Trans. R. Soc. Lond. B Biol. Sci*. 367, 906–918. DOI: 10.1098/rstb.2011.0351

Kiebel, S. & Friston, K. J. (2011). Free energy and dendritic self-organization. *Frontiers in Systems Neuroscience*, 5(80). DOI: 10.3389/fnsys.2011.00080.

Kiefer, A. (2017). Literal Perceptual Inference. In T. Metzinger & W. Wiese (Eds.). *Philosophy and Predictive Processing: 17* (pp. 1-19). Frankfurt am Main: MIND Group. DOI: 10.15502/9783958573185.

Kiefer, A. & Hohwy, J. (2018). Content and misrepresentation in hierarchical generative models. *Synthese*, 195, 2387–2415. DOI: 10.1007/s11229-017-1435-7.

Kirchhoff, M., Parr, T., Palacios, E., Friston, K. J. & Kiverstein, J. (2018). The Markov blankets of life: autonomy, active inference and the free energy principle. *J. R. Soc. Interface* 15:20170792. DOI: 10.1098/rsif.2017.0792.

Kirchhoff, M. & Robertson, I. (2018). Enactivism and predictive processing: a non-representational view. *Philosophical explorations* 2(21). DOI: 10.1080/13869795.2018.1477983.

Kiverstein, J., Miller, M. & Rietveld, E. (2019). The feeling of grip: novelty, error dynamics, and the predictive brain. *Synthese*, 196, 2847–2869. DOI: 10.1007/s11229-017-1583-9.

Kiverstein, J., Sims, M. (2021). Is free-energy minimisation the mark of the cognitive?. *Biol Philos*, 36, 25. https://doi.org/10.1007/s10539-021-09788-0.

Klein, C. (2018). What do predictive coders want? *Synthese*, 195, 2541–2557. DOI: 10.1007/s11229-016-1250-6.

Knill, D. C. & Pouget, A. (2004). The Bayesian Brain: The Role of Uncertainty in Neural Coding and Computation. *Trends in Neurosciences*, 27, 712–719.

Knill, D. C. & Richards, W. (1996). *Perception as Bayesian Inference*. Cambridge: Cambridge University Press.

Korbak, T. (2021). Computational enactivism under the free energy principle. *Synthese*, 198, 2743–2763. DOI: 10.1007/s11229-019-02243-4.

Kretzmer, E. R. (1952). Statistics of Television Signals. *Bell System Technical Journal*, 4(31), 751-763.

Kwisthout, J., Bekkering, H. & van 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.

Kwisthout, J. & van Rooij, I. (2019). Computational resource demands of a predictive Bayesian brain. *Computational Brain and Behaviour*, 3(2), 174-188.

Lau, J. & Deutsch, M. (2002). Externalism About Mental Content. In E. N. Zalta (Ed.). *Stanford Encyclopedia of Philosophy*, URL = <https://plato.stanford.edu/archives/fall2019/entries/content-externalism/>, 1-13. access: 02.08.2019.

Lelito, J. P. & Brown, W. D. (2006). Complicity or Conflict over Sexual Cannibalism? Male Risk Taking in the Praying Mantis Tenodera aridifolia sinensis. *The American Naturalist*, 2(168), 263-269.

Litwin, P. & Miłkowski, M. (2020). Unification by fiat: arrested development of predictive processing. *Cognitive Science*, 7(44), 12867. https://doi.org/10.1111/cogs.12867.

MacKay, D. M. (1956). The epistemological problem for automata. In C. E. Shannon & J. McCarthy (Eds.). *Automata studies* (pp. 235-251). Princeton: Princeton University Press.

Marr, D. (1982). *Vision: A computational approach*. San Francisco: Freeman & Co.

Marvan, T. & Havlík, M. (2020). Is Predictive Processing a Theory of Consciousness? *New Ideas in Psychology*, 61(8) DOI: 10.1016/j.newideapsych.2020.100837.

McGregor, S. (2017). The Bayesian stance: Equations for 'as-if' sensorimotor agency. *Adaptive Behavior*, 2(25), 72–82. DOI: 10.1177/1059712317700501.

Milidge, B., Tschantz, A., Seth, A. K. & Buckley, C. L. (2020). Relaxing the Constraints on Predictive Coding Models. *arXiv*: 2010.01047v2 [q-bio.NC]. 1-17.

Miller Tate, A. J. (2019) A predictive processing theory of motivation. *Synthese*, 1-29. First online. https://doi.org/10.1007/s11229-019-02354-y.

Miłkowski, M. (2016a). Integrating cognitive (neuro)science using mechanisms. *Avant*, VI(2), 45–67. DOI: 10.26913/70202016.0112.0003.

Miłkowski, M. (2016b). Unification Strategies in Cognitive Science. *Studies in Logic, Grammar and Rhetoric*, 48(1), 13–33. DOI: 10.1515/slgr-2016-0053.

Mitchel, S. (2003). *Biological Complexity and Integrative Pluralism*, Cambridge: Cambridge University Press.

Mumford, D. (1992). On the computational architecture of the neocortex. II. The role of cortico-cortical loops. *Biological Cybernetics*, 66 (3), 241–251.

Nair V., Susskind J. & Hinton G. E. (2008). Analysis-by-Synthesis by Learning to Invert Generative Black Boxes. In V. Kůrková, R. Neruda & J. Koutník (Eds.). *Artificial Neural Networks - ICANN 2008. ICANN 2008. Lecture Notes in Computer Science, vol 5163* (pp. 971-981). Berlin, Heidelberg: Springer.

Nave, K. (2021). Visual experience in the predictive brain is univocal, but indeterminate. *Phenom Cogn Sci*. 1-25. First online. https://doi.org/10.1007/s11097-021-09747-w.

Neisser, U. (1967). *Cognitive Psychology*. Englewood Cliffs, NJ: Prentice Hall.

Ng, Y. N. & Jordan, M. I. (2001). On Discriminative vs. Generative Classifiers: A comparison of logistic regression and naive Bayes. In *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic* (pp. 841–848).

Morville, T., Friston, K., Burdakov, D., Siebner, H. R. & Hulme, O. J. (2018). The homeostatic logic of reward (1–43). *bioRxiv*, 242974. https://doi.org/10.1101/242974.

Laudan, L. (1977). *Progress and Its Problems: Towards a Theory of Scientific Growth*. Berkeley: University of California Press.

Oaksford, M. & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford: Oxford University Press.

O'Brien, G. & Opie, J. (2004). Notes towards a structuralist theory of mental representation. In H. Clapin, P. Staines & P. Slezak (Eds.). *Representation in mind: new approaches to mental representation*, (pp. 1-20). Amsterdam: Elsevier.

Oliver, B. (1952). Efficient coding. *Bell System Technical Journal*, 31(4), 724–750.

Orlandi, N. (2016). Bayesian perception is ecological perception. *Philosophical Topics*, 44(2), 327-351.

Orlandi, N. (2018). Predictive perceptual systems. *Synthese*, 195, 2367–2386.. DOI: 10.1007/s11229-017-1373-4.

Orlandi, N. & Lee, G. (2018). How Radical is Predictive Processing? In M. Colombo, E. Irvine & M. Stapleton (Eds.). *Andy Clark & Critics* (pp. 206-221). Oxford: Oxford University Press. DOI: 10.1093/oso/9780190662813.003.0016.

Park, H.-J. & Friston, K. J. (2013). Structural and Functional Brain Networks: From Connections to Cognition. *Science*, 342, 579-589. DOI: 10.1126/science.1238411.

Parpart, P., Jones, M. & Love, B. C. (2018). Heuristics as Bayesian inference under extreme priors. *Cognitive Psychology*, 102, 127-144. DOI: 10.1016/j.cogpsych.2017.11.006.

Parr, T. & Friston, K. J. (2018). The Anatomy of Inference: Generative Models and Brain Structure. *Front. Comput. Neurosci*. 12:90. DOI: 10.3389/fncom.2018.00090.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Francisco: Morgan Kaufmann Publishers.

Pezzulo, G., Rigoli, F. & Friston, K. J. (2015). Active inference, homeostatic regulation and adaptive behavioural control. *Progress in Neurobiology*, 134, 17–35.

Pfeifer, R. & Bongard, J. (2006). *How the Body Shapes the Way We Think: A New View of Intelligence*. Cambridge, MA: MIT Press.

Pickering, M. J. & Clark, A. (2014). Getting ahead: forward models and their place in cognitive architecture. *Trends in Cognitive Science*, 8(9), 451-456. DOI: 10.1016/j.tics.2014.05.006.

Piekarski, M. (2019). Normativity of Predictions: A New Research Perspective. *Front. Psychol.* 10:1710. DOI: 10.3389/fpsyg.2019.01710.

Povich, M. (2019). Model-based cognitive neuroscience: Multifield mechanistic integration in practice. *Theory & Psychology*, 29(5), 640–656. DOI: 10.1177/0959354319863880.

Ramstead, M. J. D., Badcock, P. B. & Friston, K. J. (2017). Answering Schrödinger's question: A free-energy formulation. *Physics of Life Reviews*, 24, 1–16. https://doi.org/10.1016/j.plrev.2017.09.001.

Ramstead, M. J. D., Kirchhoff, M., Constant, A. & Friston, K. J. (2021). Multiscale integration: beyond internalism and externalism. *Synthese*, 198, 41–70. DOI: 10.1007/s11229-019-02115-x.

Ramstead, M. J. D., Kirchhoff, M. D. & Friston, K. J. (2020). A tale of two densities: Active inference is enactive inference. *Adaptive Behavior*, 28(4), 225-239. https://doi.org/10.1177/1059712319862774.

Rao, R. P. & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra- classical receptive-field effects. *Nat Neurosci*, 2 (1), 79–87. DOI: 10.1038/4580.

Rescorla, M. (2015). Bayesian perceptual psychology. In M. Matthen (Ed.). *The Oxford handbook of philosophy of perception* (pp. 694-716). Oxford: Oxford University Press.

Sanborn, A. M. (2017). Types of approximation for probabilistic cognition: Sampling and variational. *Brain and Cognition*, 112, 98-101. https://doi.org/10.1016/j.bandc.2015.06.008.

Schlicht, T. & Dołęga, K. (2021). You can't always get what you want: Predictive processing and consciousness. *Philosophy and the Mind Sciences*, 2. https://doi.org/10.33735/phimisci.2021.80.

Schwartenbeck, P., FitzGerald, T., Dolan, R. J. & Friston, K. (2013). Exploration, novelty, surprise, and free energy minimization. *Frontiers in Psychology*, 4(710), 1–5. https://doi.org/10.3389/fpsyg.2013.00710.

Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends Cogn Sci*. 17(11), 565-73. DOI: 10.1016/j.tics.2013.09.007.

Seth, A. K. (2015). Inference to the best prediction. In T. Metzinger & J. M. Windt (Eds.). *Open MIND*, 35R (pp. 1–8). Frankfurt am Main: MIND Group. DOI: 10.15502/9783958570986.

Seth, A. K. & Friston, K. J. (2016). Active interoceptive inference and the emotional brain. *Philosophical Transactions Of The Royal Society B: Biological Sciences*, 371(1708), 20160007. DOI: 10.1098/rstb.2016.0007.

Shea, N. (2014) Exploitable isomorphism and structural representation. *Proc Aristot Soc* CXIV, 77–92. DOI: 10.1111/j.1467-9264.2014.00367.x

Shi, Y. Q. & Sun, H. (1999). *Image and video compression for multimedia engineering: Fundamentals, algorithms, and standards*. Boca Raton, FL: CRC Press.

Sims, A. (2016). A problem of scope for the free energy principle as a theory of cognition. *Philosophical Psychology*, 29, 967–980.

Sims, A. (2017). The problems with prediction – the dark room problem and the scope dispute. In: T. Metzinger & W. Wiese (Eds.). *Philosophy and Predictive Processing: 2* (pp.1–18). Frankfurt am Main: MIND Group. DOI: 10.15502/9783958573246.

Smith, R., Friston, K. J. & Whyte, C. (2021). A Step-by-step Tutorial on Active Inference and Its Application to Empirical Data. *PsyArXiv*. January 2. DOI:10.31234/osf.io/b4jm6.

Spratling, M. W. (2012). Unsupervised learning of generative and discriminative weights encoding elementary image components in a predictive coding model of cortical function. *Neural Comput*. 24(1), 60-103. DOI: 10.1162/NECO_a_00222.

Spratling, M. W. (2017). A review of predictive coding algorithms. *Brain and Cognition*, 112, 92-97. https://doi.org/10.1016/j.bandc.2015.11.003.

Srinivasan, M. V., Laughlin, S. B. & Dubs, A. (1982) Predictive coding: a fresh view of inhibition in the retina. *Proc R Soc Lond B Biol Sci*. 216(1205), 427-59.

Stanford, P. K. (2019). Unconceived Alternatives and Conservatism in Science: The Impact of Professionalization, Peer-review, and Big Science., *Synthese*, 196, 3915–3932. DOI: 10.1007/s11229-015-0856-4.

Sterling, P. & Eyer, J. (1988). Allostasis: A new paradigm to explain arousal pathology. In S. Fisher & J. Reason (Eds.). *Handbook of life stress, cognition, and health* (pp. 629-649). Chichester: John Wiley & Sons.

Thornton, C. (2016). Predictive processing is Turing complete: A new view of computation in the brain. Preprint.

Tversky, A. & Kahneman. D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, 4157(185), 1124-1131 DOI: 10.1126/science.185.4157.1124.

Tversky, A. & Kahneman. D. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, 2(47), 263-291.

van Es, T. (2020a). Minimizing prediction errors in predictive processing: from inconsistency to non-representationalism. *Phenom Cogn Sci*, 19, 997–1017. https://doi.org/10.1007/s11097-019-09649-y.

van Es, T. (2020b). Living models or life modelled? On the use of models in the free energy principle. *Adaptive Behavior*, 3(29), 315-329. https://doi.org/10.1177/1059712320918678.

van Rooij, I., Wright, C. D., Kwisthout, J. & Wareham, T. (2018). Rational analysis, intractability, and the prospects of 'as if'-explanations. *Synthese*, 195, 491–510. DOI: 10.1007/s11229-014-0532-0.

Varela, F., Maturana, H. & Uribe, R. (1974). Autopoiesis: The organization of living systems, its characterization and a model. *Biosystems*, 5 (4), 187–196.

Varela, F., Thompson, E. & Rosch, E. (1991). *The embodied mind: Cognitive science and human experience*. Cambridge, MA, London, UK: The MIT Press.

Veissière, S., Constant, A., Ramstead, M. J. D., Firston, K. J. & Kirmayer, L. J. (2020). Thinking through other minds: A variational approach to cognition and culture. *Behavioral and Brain Sciences*, 43, e90. DOI: 10.1017/S0140525X19001213.

Wachowski, W. M. (2018). Commentary: Distributed Cognition and Distributed Morality: Agency, Artifacts and Systems. *Frontiers in Psychology*, 9, 490. DOI: 10.3389/fpsyg.2018.00490.

Wald, A. (1947). An Essentially Complete Class of Admissible Decision Functions. *The Annals of Mathematical Statistics*, 549-555.

Wiese, W. (2015). Perceptual Presence in the Kuhnian-Popperian Bayesian Brain – A Commentary on Anil K. Seth. In T. Metzinger & J. M. Windt (Eds). *Open MIND: 35(C)* (pp. 1-19). Frankfurt am Main: MIND Group. DOI: 10.15502/9783958570207.

Wiese, W. (2017). What are the contents of representations in predictive processing? *Phenomenology and the Cognitive Sciences*, 16, 715–736. DOI: 10.1007/s11097-016- 9472-0.

Wiese, W. & Metzinger, T. (2017). Vanilla PP for Philosophers: A Primer on Predictive Processing. In T. Metzinger & W. Wiese (Eds.). *Philosophy and Predictive Processing: 1* (pp. 1-18). Frankfurt am Main: MIND Group. DOI: 10.15502/9783958573024.

Wilkinson, S. (2014). Accounting for the phenomenology and varieties of auditory verbal hallucination within a predictive processing framework. *Consciousness and Cognition*, 30, 142-155.

Williams, D. (2019a). Hierarchical minds and the perception/cognition distinction. *Inquiry*. 1-24. DOI: 10.1080/0020174X.2019.1610045.

Williams, D. (2019b). Epistemic Irrationality in the Bayesian Brain. *The British Journal for the Philosophy of Science*, 1-45, axz044, DOI: 10.1093/bjps/axz044.

Williams, D. (2020). Predictive coding and thought. *Synthese*, 197, 1749–1775. DOI: 10.1007/s11229-018-1768-x.

Zahavi, D. (2018). Brain, Mind, World: Predictive Coding, Neo-Kantianism, and Transcendental Idealism. *Husserl Studies*, 34, 47-61. DOI: 10.1007/s10743-017-9218-z.