Synthese (2022) 200:352
https://doi.org/10.1007/s11229-022-03837-1

ORIGINAL RESEARCH

Motivation, counterfactual predictions and constraints: normativity of predictive mechanisms

Michał Piekarski[1]

Received: 27 March 2021 / Accepted: 29 July 2022 / Published online: 19 August 2022

Summary
The aim of this paper is to present the ontic approach to the normativity of cognitive functions and mechanisms, which is directly related to the understanding of biological normativity in terms of normative mechanisms. This approach assumes the hypothesis that cognitive processes contain a certain normative component independent of external attributions and researchers' beliefs. This component consists of specific cognitive mechanisms, which I call normative. I argue that a mechanism is normative when it constitutes given actions or behaviors of a system. More precisely, it means that, on the one hand, it is their constitutive cause, and on the other hand, it determines a certain field of possibilities from which the system, guided by its own goals, preferences, environmental constraints, etc., chooses the appropriate action or behavior according to a given situation. The background for the analyses presented here is the predictive processing framework, in which it can be shown that at least some of the predictive mechanisms are in fact normative mechanisms. I refer here to the existence of a motivational relation which determines the normative dependence of the agent's actions due to specific predictions and environmental constraints.

Keywords Normativity · Counterfactual predictions · Motivation · Constraints · Mechanisms · Explanation · Predictive processing · Generative model

1 Introduction

Currently, many researchers emphasize the need to use normative concepts in the analysis of cognitive and biological functions (cf. Bickhard, 2003; 2009; Christensen,

✉ Michał Piekarski
m.piekarski@uksw.edu.pl; m.a.piekarski@gmail.com

[1] Cardinal Stefan Wyszyński University in Warsaw Institute of Philosophy, Wójcickiego 1/3 St, 01-938 Warsaw, Poland

2012; Christensen & Bickhard, 2002; Godfrey-Smith, 1993; Kitcher, 1993; Millikan, 1984; 1989). On the one hand, they argue that functions are normative and that their normativity makes it possible to explain the functioning of organisms, while showing when it is incorrect. On the other hand, the attractiveness of their approach lies in the fact that they reduce what is normative to what is descriptive. The issue of the normativity of biological functions is the subject of lively debate. However, it is an issue that is often raised "on the occasion" or "on the sidelines" of discussing other problems, even though it seems that in many cases it is of key importance.

Christensen (2012) emphasizes the need to include what he describes as evaluative normativity in scientific analyses of living and artificial systems. In his opinion, we must be able to say when such and such states of a given system can be described as better or worse. Christensen's thesis is confirmed by the influential opinion of Paul Thagard, who stated that philosophy has a non-trivial role to play, because it provides science, and in particular cognitive science, with certain normative questions: "philosophy is concerned not only with how things are but also with how they should be. Philosophical theories of knowledge and morality need to go beyond descriptive theories of how people think and act by also developing normative (prescriptive) theories of how people ought to think and act" (2009, p. 238). The view of Thagard and Christensen on the one hand points to an important place for normative language in scientific debates, and on the other expresses the unspoken idea that normativity is something that (1) has an evaluative character and that (2) it is ascribed to objects, states of affairs, processes or functions externally.

In this paper, I will argue for the opposite. I claim that there are specific mechanisms that are normative by their nature and not because of the attribution made by the observer. In this sense, normativity is a real property that can be assigned to a given mechanism due to the function it performs in an organism, cognitive system, or, for example, a decision-making process, and which cannot be reduced to some evaluation practices.[1] In other words, the condition for ascribing normativity is that the mechanism in question is normative *per se*.[2] I will argue that in order to be able to explain the success or failure of an agent's actions in the environment, attention should be paid to specific biological mechanisms (illustrated here by the example of predictive mechanisms) that determine the selection of such and such actions. This means that explaining the potential effectiveness of an agent's actions in an uncertain environment presupposes explaining mechanisms that are normative for these actions. A mechanism[3] is normative when it plays a specific causal role in the expla-

[1] We deal with them when we determine that, for example, a given organ is working properly. See the teleosemantic concept of a function (cf. § 2; Millikan. 1984).

[2] The proposed approach assumes that the ascription of normativity is independent of the strategy that Dennett (1987) links with adopting the so-called intentional stance, which consists in the fact that the observer tries to describe and predict given phenomena with the help of attributed beliefs, desires or normative concepts. In practice, this means that the correlates of these beliefs, desires and concepts do not have to exist. In this article, I argue that ascribing normativity to given mechanisms is possible because at least some of them are normative as such. In this sense, intentionality is not a constitutive component in the explanation of these mechanisms.

[3] By mechanism I mean "structure performing a function in virtue of its component parts, component operations, and their organization. The orchestrated functioning of the mechanism is responsible for one or more phenomena" (Bechtel, 2008, p. 13).

nation of such and such actions or behaviors. And it can play such a role because it plays such and such a causal role in the functioning of a given mechanism.[4]

My analyses distinguish between external normativity—that is, one that allows the observer to externally (i.e. from the perspective of the observer of a given object or mechanism) decide whether a given object or mechanism is effective or ineffective, correct or incorrect, good or bad, etc.—and internal normativity (the perspective of a system or mechanism), i.e. that which is ascribed to a given mechanism for ontic reasons (i.e. because the mechanism *is* as it is). This is the perspective that the biological system itself has about what counts as its proper functioning Winning,2020a, p. 20).[5] This approach assumes that internal normativity (understood here as a constitutive property of actions and behaviors) is a real property of a given object, mechanism or behavior, independent of the observer and its evaluative practices.[6] In this sense, the term "normative" applies to those properties that are constitutive for the structure and functioning of a given system (Bickhard, 2003)[7] as well as for their explanation. Internal normativity can therefore be called "constitutive normativity" or "explanatory normativity".

Before going any further, it is necessary to justify why the notion of normativity is important for the explanation of the predictive processing framework (PP).[8] According to this framework brains are predictive machines (entailing generative models), whose primary function is to constantly match information coming from sensory modalities to internally generated, model-based predictions explaining the nature and sources of such information (for full exposition see: Clark, 2013; 2016; Hohwy, 2013, 2014, 2020a; Piekarski, 2021; Wiese & Metzinger, 2017). The process of minimizing prediction errors, i.e. the disproportion between expectations (hypotheses about the world) based on the internal parameters of the model and the variable infor-

[4] The theoretical framework for my deliberations on mechanisms is provided by the mechanistic philosophy of science (Bechtel, 2008; Craver, 2007; Machamer et al., 2000; cf. Garson, 2013; Winning 2020b).

[5] The approaches assuming the existence of evaluative normativity can be described, following Winning (2020a), as external perspectivism. It assumes that proper functioning of a given system depends on the perspective taken by an external observer (e.g. a scientist studying the system).

[6] In practice, this means that normativity is, in a sense, reducible to the functional or structural properties of specific mechanisms. This is what I will argue for: a special functional property.

[7] By system, I mean a complex or simple biological mechanism that performs specific functions. In this sense, for example, bacteria and humans are living systems.

[8] It is important to emphasize the differences between PP, predictive coding, prediction error minimization and active inference as it is possible to confuse these approaches. This confusion is licensed by reference to prediction error minimization, which is a hallmark of predictive coding. Predictive processing is an umbrella term that inherits from the notion of predictive coding in the brain, where the (Bayesian) brain can be cast as minimizing prediction errors. However, predictive coding in and of itself has nothing to say about action. Predictive processing then appeals to active inference as extending the principles of prediction to cover action, planning and decision-making. Active inference is a corollary of the free energy principle (FEP) and admits the possibility that prediction errors can be minimized not just through perceptual inference but by actively sampling the sensorium to realize sensations that are predicted or predictable (cf. Friston, 2009). Technically, prediction error minimization survives as a normative principle in active inference (i.e., predictive processing accounts of action) in the form of free energy gradients. In other words, the gradients of the objective function that underwrites action, planning and perception can always be expressed as a form of prediction error. It should also be added that, in the literature, some authors (cf. Hohwy, 2013) use the terms PP and predictive error minimization interchangeably.

mation reaching the model through the senses, assumes hierarchical and multi-level predictive information processing and a generative model which is (generally speaking) the statistical model of how observations are generated.[9] Supporters of the PP claim that the generative model is a hierarchical Bayesian probabilistic model, which constructs and tests internal models of the external environment by implementing cognitive processes that are an approximation of Bayesian inference (Clark, 2013, p. 189). Bayesian rule helps to identify an optimal way of updating one's beliefs given new evidence under conditions of uncertainty. Thanks to this, prediction errors will be minimized by the model only when the model adopts the best possible hypothesis regarding the causes of the sensory signal source (perceptual inference) or by making an active inference when the agent interferes with the causal structure of relevant states of affairs and changing the information reaching the model (cf. § 3; Friston, 2010; Pezzulo et al., 2015). Perceptual inference process can be called 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 statistics of the input. For this reason, the generative model can be understood as self-evidencing (Hohwy, 2014). In this regard, Jakob Hohwy states that the Bayesian rule can be perceived as a paradigm of normativity: "it prescribes optimal relative weighting of evidence and prior belief. Violations of the norm occur when too much or too little weight is given to the prior or to the evidence, leading to false inference. (…) It approximates the optimal results a system would get by complying with the Bayesian norm" Hohwy, 2020b, p. 15).[10]

The belief in the normative nature of Bayesian models is shared by many researchers (cf. Anderson, 1990; Hahn, 2014; Oaksford & Chater, 2007; Oaksford, 2014). Bayesian models are meant to be normative in the sense that human thinking is measured and evaluated in the light of the rules it formulates. In other words, the Bayesian models using Bayesian rule are meant not only to describe *how* cognitive and decision-making processes take place, but also to show how they *should* proceed.

We have to say that the assumption about the normativity of Bayesian models is closely related to the assumption regarding the rationality of agents that think, make decisions and act, approximating the (normative) Bayesian rule. The position that links rationality with normativity can be defined as normative rationalism (Elqayam & Evans, 2011, p. 235). For this reason, there is a common view that Bayesian theorem describes the optimal procedure for inference under uncertainty. However, this position is not free from objection (cf. Knill & Pouget, 2004; Joyce, 2004). The belief in the normative nature of the Bayesian rule approximated by generative models is also challenged by many researchers (cf. Colombo et al., 2018; Elqayam & Evans, 2011; Jones & Love, 2011). In their opinion, in science, and above all in psychology and cognitive science, the prescriptive approach should be abandoned in favor of the empirical or descriptive one.

[9] This is the model of the conditional probability of the observable X, given a target Y (cf. Kiefer & Hohwy, 2017; Nair et al., 2008).

[10] Hohwy associates this approach with the FEP (cf. Friston, 2010; Friston, 2013; Friston & Stephan, 2007), claiming that it offers a mathematical and conceptual framework necessary for the analysis of the existence of self-organizing systems.

In this paper I will defend the view according to which Bayesian PP is normative not because it allows for the formulation of rules of action and policies or because it contains such rules, but because (some of) the predictive mechanisms themselves are normative[11](cf. Piekarski, 2019 for a basic introduction to this topic). They condition the choice of such and such actions by the agent. To substantiate this view, I will refer to the relation that takes place between a given prediction and the action conditioned by it. I will call this relation motivational. It enables the agent to act in one way and not another, depending on the belief system that the agent considers true, *resp*. accurate (cf. O'Brien, 2005). The motivational relation (on which the motivation is founded as the need to reduce uncertainty (Anselme, 2010)) is the relationship between predictions and the actions they guide in relation to certain environmental states. This means that the agent's motivation is shaped by the generated predictions that stem from the need to minimize prediction errors by taking into account the states of the environment as well as the possibilities of action this environment offers. I will argue that a given mechanism is normative as long as its operation (function) is to generate (normative) predictions. The mechanism thus understood makes it possible to explain the behavior of the agent in the environment in terms of its success or failure.

At this point, we should also mention normative theories in the life and physical sciences. This is important because: (1) the approach proposed here can be regarded to some extent as an application of these theories to explain the activities of living organisms. In the approach proposed here, this means that low-level physical normativity can be and is actually realized by high-level normative mechanisms (such as, for example, the predictive mechanisms described here); and (2) there is a clear link between these theories and the approach proposed here, especially with regard to the significant distinction between rational Bayesian inference and bounded or approximate Bayesian inference. Namely: in the technical literature, a normative theory is generally read as a formal specification of a process that is equipped with an objective function. In other words, the process can be understood as extremizing or optimizing a function of its states. Clear examples here include Lyapunov functions in dynamical systems theory through to loss functions in optimal control. Implicit in this sort of definition is a process that can be cast as optimizing some measurable function.

This reading of normative may have a special role in the present discussions. This follows because treatments of PP implicitly invoke an objective function – namely - model evidence or marginal likelihood. Mathematically, rational Bayesian decision-making then corresponds to exact Bayesian inference. However, exact Bayesian inference is mathematically intractable, which is why approximate Bayesian inference is the only realizable kind of Bayesian inference. The objective function is then known as a variational free energy (VFE) or an evidence bound (cf. Winn & Bishop, 2005). This converts exact Bayesian inference into approximate Bayesian interest

[11] This position may also be referred to as "metaphysical normativism" according to which predictive mechanisms have a normative essence, which means that their real definition includes a normative property as a constituent (cf. Tiehen, 2022). According to Justin Tiehen, with whom I fully agree, "metaphysical normativism might be part of the best explanation for why methodological normativism is empirically successful". Methodological normativism assumes that "normative accounts should guide psychological theorizing" (Tiehen, 2022, p. 5). See § 7.

inference.[12] By introducing VFE, an in intractable integration problem was converted into a tractable optimization problem; namely minimizing VFE (Dayan et al., 1995). In short, the definitive aspect of PP—at least under FEP—is its normative aspect. It is an important observation because, on the one hand, it justifies the need to analyze predictive mechanisms, and on the other, it constitutes an additional argument against those approaches that deny the legitimacy of research on normativity: from the point of view of normativity understood in this way, postulated by the aforementioned authors, a descriptive or empirical approach is also a prescriptive approach.

I have structured the paper as follows. In Sect. 2, I discuss the concept of the normative function and justify why I give up the teleosemantic approach to it. Instead, I propose an approach based on the actual causal role played by a given function. In Sect. 3, I describe the motivational relation that defines the normative relation between the generated predictions, the actions taken, and the (expected) states of the world. Section 4 explains the relation between generated prediction and policy selection. I claim that the choice of a given policy is related to a specific environmental situation, the generated high-precision prediction that allows inferring the expected outcome; and objective function, which specifies a general (normative) requirement for policies. Section 5 distinguishes counterfactual predictions from semifactual predictions. The former normalize the (relatively) effective actions, the latter the ineffective ones. The possibility of choosing between them proves the normativity of the motivational relation. In fact, the agent does not need to know which of the actions taken by it are based on counterfactual predictions and which are based on semifactual predictions, but is always obliged to choose an action normalized by this or that prediction. In Sect. 6, I discuss the role of environmental constraints in the motivational relation and their importance in explaining the normativity of predictive mechanisms. In Sect. 7, I argue that mechanisms that perform normative functions can be the causes of certain actions and behaviors. Importantly, these are not the only causes for these actions, but they are the causes that explain (in mechanistic manner) the success or failure of the action taken by a given organism in a specific environment and situation. They are therefore what I call "constitutive" causes. In the *Conclusion*, I summarize the analyses carried out and emphasize their importance.

2 The concept of the normative function

The concept of the normative function appears in the works of Millikan (1984; 1989). She links it with the concept of the etiological proper function. An object has a proper function if it derives from a line that owes its survival to the existence of a correlation between its distinguishing features and the effects that can be defined as functions of these features (Millikan, 1989, pp. 288–289). This means that a proper function has properties that have been selected through the mechanisms of evolutionary natural selection. In this way, functions can be assigned to such systems, organisms or artifacts which, although having their functions, are not capable of performing them.

[12] Interestingly, VFE was introduced by Richard Feynman to solve an intractable inference problem in quantum electrodynamics (Feynman, 1998).

They fail to fulfill their inherent functions because of some "damage" or because of certain background conditions, for example, those that helped their predecessors keep the function in operation and are now absent. This approach to function relates firstly to how a given thing has been designed or how it acts on purpose (as opposed to what it does accidentally) and secondly, to the existence of some kind of "pattern" that can be found wherever there is a natural attribution of purpose and/or intentionality. For this reason, this approach is called normative, with the proviso that this understanding of normativity has nothing to do with some "evaluation" or "assessment". In such an approach, the concept of a proper function should be treated as a specific measure or norm for deciding whether something is a function or not, or, crucially, whether it is a dysfunction of a given system, organism or artifact (Millikan, 1984). Normativity in this framework is understood purely historically and is associated with normality as a certain regularity of action or performance (Millikan, 1989, p. 284).

In this approach, the function of representation is the proper function of perceptual and cognitive systems. The dysfunction here consists in not fulfilling the proper function, that is, in incorrect, or ultimately erroneous representation. Misrepresentation is an important and constitutive element of the normal functioning of the organism, because "from an evolutionary perspective, it is more profitable to overrepresent certain features of the environment, rather than not representing them at all" (Bielecka, 2018, p. 185). What is significant for this view is (1) that the possible misrepresentation cannot be explained without reference to the evolutionary history of an organ or part of it; and (2) that in order to explain the function of representation by specific organs or, more broadly, organisms, it is necessary to explain the dysfunction or to define the normative conditions appropriate to the representation function.

However, what makes Millikan's proposal attractive also determines its problematic nature. Firstly, assuming that the explanation of the dysfunction of representational systems is based on the reference to the history of the selection of their natural predecessors, leads to objections formulated by Davies (2001).[13] Secondly, Millikan's approach offers a description of the normative function of representation only from the perspective of the species, and not individuals. In practice, this means that we can only provide minimal conditions that the representative system should meet in order to be able to represent correctly or incorrectly, but we cannot explain why a given organism or system behaved in one way or another or produced such and such representation. The point is that talking about normativity in teleosemantics serves only to justify this and no other type of description of biological and artificial systems. In practice, this means that if we wanted to use the notion of proper function to explain the normativity of the Bayesian generative model and its predictive mechanisms, we should reduce this explanation to the conclusion that it consists in showing the normal conditions[14] that it the model must satisfy in order to (for example) incorrectly minimize prediction errors. However, such an approach is unsatisfactory,

[13] Davies (2001) argues that while most natural features do have specific functional properties, those properties are not produced and favored by natural selection processes. The latter, according to Davies, are not essential for the explanation of any function.

[14] These conditions relate to the proper functioning of a given organ and regular relations with other parts of the organism and the environment (Millikan, 1989).

because generative models, by definition, work in such a way that they are always prone to misrepresentation. In the technical language of PP: a statistical generative model determines the posterior probabilities of the predictions it generates and constantly compares them with the incoming data. It thus updates its internal parameters, while determining new probability distributions for the specific values of the variables it takes. The whole process consists in bringing the probability distributions of internal parameters and the assigned probability distributions of the external states of the environment as close as possible to each other. The discrepancy between the two distributions is referred to in the PP as the "Kullback-Leibler divergence" (KL divergence). It concerns the difference between posterior generative and approximate recognition distributions (cf. Bogacz, 2017; Kiefer & Hohwy, 2017). This means that misrepresentation is inherent in the very nature of the generative model.

It should be stated that ascribing the possibility of misrepresentation to given organs, systems or mechanisms or recognizing their dysfunction (which is constitutive for the normativity of a function or mechanism) is epistemic. Therefore, it is relativized to the description formulated by an external observer in relation to its research interests. I argue that the causal conditions that the organism or system "armed" with the generative model must meet in order to be able to act in the uncertain environment and guarantee cognitive and non-cognitive successes to a specific organism, are based on internally generated probability distributions about the world. This means that a given function is normative not only because it makes it possible to explain the possibility of dysfunction, but also because it plays a specific causal role, i.e. it influences and maintains the stability of a given system (cf. Bickhard, 2003; 2009). It is normative because the system, in order to exist and self-organize itself, *must* fulfill certain normative functions. For example, the function of the heart to pump blood serves the purpose of delivering oxygen to the brain. Therefore, it is normative, because, if the function is not performed, the organism would cease to exist. For this reason, we will say that the function of the heart to produce acoustic effects is only a causal function and not a normative function, since it contributes neither to the realization of other processes, nor to the stability and survival of the organism. It should also be emphasized that the normative function in the sense of Millikan is possible thanks to the performance of the normative function as described here. In this sense, I will refer to explanatory normativity because of its important role in explaining the behavior of a given system, as I argue in § 6. This explanation comes down to indicating a specific normative mechanism (i.e. one that performs a normative function), which is the cause of this behavior or, in other words, makes the behavior possible. In the approach that I defend, the constitutive causes of actions and behaviors are specific mechanisms that perform normative functions. By constitutive cause I mean one that determines the logical conditions for the appearance and realization of a given action or behavior.[15]

[15] One might object to the approach proposed here by saying that the normativity of the heart pumping blood is more crucial to the survival of the organism than the normativity of actions. I agree, but it should be emphasized that both the normativity of the heart pumping blood and the normativity of predictions or actions are explanatory. The explanation of the work of the heart implies a reference to the normative mechanism of blood pumping just like the explanation of the action of a given agent is assumed with reference to certain normative mechanisms, which are associated here with the predictive mechanisms.

Referring these analyses to PP, it should be stated that the content of the generated prediction is normative for the selection of specific actions. Predictions are normative because they are conditions for selecting appropriate actions, *resp*. policies of action. This means that the actions taken by the agent have their own logical conditions that define the criteria for selecting these actions. Thanks to this, specific predictions can justify selected policies of action. Normativity understood in this way is internal, because the agent's obligation to act in a certain way results from the very fact of the existence of a given prediction, just as, for example, the fact of the existence of a moral norm implies an obligation to observe it. The normativity of prediction is directly related to the requirement of long-term minimization of the prediction error, *resp*. VFE. An agent that does not minimize prediction errors in the long-term will cease to exist, so it must do so because of all the possible states it may be in. It must find an appropriate subset of states (determined by some priors and the organism's phenotype) that will allow it to survive and effectively exchange energy with the environment (Friston & Stephan, 2007).

The above remarks should be made more precise. First of all, when I speak of the "normativity of predictions", I refer to the predictions that are related to active inference, i.e. the domain of actions and decision making. I claim that a specific prediction or a set of them (I will describe them later in terms of counterfactual predictions) normalizes the selection of such and such actions, *resp*. policies of action. It is not difficult to notice that such an approach presupposes weak normativity. Weak normativity—as I understand it—can be associated with the metaethical position of motivational internalism (cf. Rosati, 2016). Motivational internalism assumes that having a motive for a given action, in our case appropriate prediction, is sufficient to justify this action. In other words: prediction is the norm for a given action, *resp*. policy of action, in the sense that this action, *resp*. policy of action, is consistent or inconsistent with this prediction (cf. Brandom, 1994, p. 18–20) (by consistence I understand the possibility of justifying a given action, *resp*. policy of action, by reference to a given prediction).[16] The weakness of this approach is that a normative prediction (or sets of normative predictions) allows one to reconcile certain actions, in the sense that the agent acts according to a given norm (prediction) or not. This means normative prediction (or sets of normative predictions) can specify the sequence of sensory states that must be brought about in order to achieve some outcome, which means that one might pursue multiple policies designed to bring about the same outcome. This state of affairs reveals a weak normativity, since "must" is downgraded to a "could", while duty is replaced by arbitrariness.[17]

It should also be added that from the point of view of the system, *resp*. the organism, there is no "more" or "less" normative function. Of course, most natural features do have specific functional properties, but these properties are not created and favored by natural selection processes. The latter is not essential for the elevation of any function (Davies, 2001). This means, therefore, that to say that some functions are more critical to the survival of an organism, and others less so, is to move from the position of the system to the position of an observer (functions are obviously a significant contribution to certain system capabilities, but natural selection can only preserve, modify or eliminate them).

[16] Here I take into account the difficulties of this position, which concern the possibility of agreeing rules and their applications (cf. Kripke, 1982).

[17] I would like to thank one of the reviewers for drawing my attention to this fact.

In this paper, however, I defend a stronger understanding of the normativity of prediction, which can be associated with the metaethical position of externalism. According to this approach, and contrary to the claims of internalism, having a motive, or generating a prediction for a given action, is insufficient to justify it (cf. Rosati, 2016). A motive is therefore something separate from justification (reason), although it may sometimes coincide with it. In our case, this means that a prediction is the norm for a given action, *resp*. policy of action, not only in the sense that this action, *resp*. policy of action, is consistent or inconsistent with this prediction, but primarily in the sense that this action, *resp*. policy of action, is realized precisely *because of* such and such prediction. In other words: the agent not only acts according to the prediction, but acts *because of* it. The normativity of predictions in this strong sense implies that are related to a certain claim to have them fulfilled (cf. Korsgaard, 1996, pp. 8–9). Normative predictions in this sense not only describe the way in which the agent regulates its actions (weak normativity), but also demand what the agent should do to effectively minimize prediction errors, *resp*. VFE[18] (strong normativity). I devote the analysis in § 5 to the justification of the strong normativity of predictions[19]

This normative property of predictions makes them good guides of actions. It favors actions in situations that increase the probability of certain predictions, thereby rejecting actions that relate to situations predicted by low probability predictions. Therefore, predictions are normative also because we can assign certain logical values to them (cf. Bickhard, 2003). In this way, they can influence the content and structure of the generative model and guide action.

Let's illustrate these analyzes with a simple example: I have an important job meeting far from my house. I don't own a car and the nearest bus stop is an hour's walk away. The weather is important to me, because I have to take this condition into account when choosing my clothes, and in the long run, perhaps also the means of transport. If it is going to rain, for example, I will have to put on a coat and take an umbrella, and if it is going to be sunny, I can leave home without these items. So I look out the window and see the cloudy sky. I can also look at the barometer and thermometer. My previous experiences with changing weather and the information my senses provide (e.g. cloud cover, atmospheric pressure, air temperature) allow me to predict that it is about to rain and maybe there will be a storm. The accuracy of my predictions is crucial for the actions will I take (choice of clothes, means of transport, time of leaving home, etc.). I know that a break in the weather may make it difficult for me to be on time for my appointment and my clothes may get wet. In this sense, this prediction is normative for the actions which it conditions. It determines which actions should be taken if I assume the high probability of this prediction. It

[18] This requirement can again be interpreted as a biologically founded normative requirement: "any self-organizing system that is at equilibrium with its environment *must* minimize its free energy" (Friston, 2010, p. 1. - my emphasis; cf. Hohwy, 2021b).

[19] It should be added that the generated prediction does not of course constitute a specific norm of actions (ethical, legal or otherwise), but it is a factor that should be taken into account when taking such and such actions that are to lead to (long-term) minimization of the prediction error. In this sense, the prediction is constitutive for the actions taken. Put differently, when creating such and such predictions, the agent is obliged to take such and such action (cf. Friston, 2009).

also points to some such actions that are completely unrelated to this prediction (see § 4). If I predict that there will be a storm and want to minimize its negative effects on the achievement of the goal of reaching my destination at the appointed time in neat clothes, my optimal choice of action will be to put on an overcoat, order a taxi or postpone the meeting to another date. However, I will not minimize these negative effects if, for example, I turn on the TV, order a pizza or go to sleep, because I will not fulfill my purpose in this way (see § 5).[20] The normativity of my prediction[21] not only normalizes what I can do, but also excludes many actions that are not relevant to that prediction if it is considered accurate or true. It is also normative in the sense that if I act in accordance with my prediction (i.e. its content), I increase my chance of success for my actions, and if I ignore them, my actions may end in failure.[22]

3 Motivational relation

I claim that a given prediction significantly influences the choice of certain actions, favoring some of them and rejecting others. Consider an example: I am driving on an unlit road at night. I can see two approaching points of light. I predict that these are the headlights of a car coming from the opposite direction. I also assume this car is in the correct lane. There is also a risk[23] that it goes against the tide. However, I am not sure. So how can I make a decision about what to do? It is a situation in which there are numerous discrepancies between the model's priors and the information coming from hidden causes in the environment. These prediction errors will only be minimized once the model adopts the best possible hypothesis about the causes of the sensory source with respect to the corresponding space-time vectors. For example, one level of the model (higher) will concern the possibility of recognizing light points as car lights, another (lower lying) will refer to e.g. the detection of the edge of the perceived object, the next level will generate predictions regarding e.g. the time when both vehicles will collide. At each level, the model estimates how precise a given

[20] "I know" this from beliefs and past experiences.

[21] Normativity is understood here conditionally, and not as an unconditional concept that defines a standard or norm. It should be emphasized, however, that the statistical effectiveness of actions based on such and such prediction and undertaken under such and such conditions may lead to the formation of specific behavior patterns that can be associated with unconditionally understood normativity (see Conclusion).

[22] I am simplifying the situation here, because it may also be that I have a certain prediction, but for some reason I do not consider it true, *resp*. highly probable. Then the choice of actions will be different. The example I have given is therefore a kind of idealization that highlights a specific problem.I want to supplement these analyses by the observation that the language of the debate on social rules distinguishes between norms and conventions (Peter & Spiekermann, 2010). Conventions are certain regularities of behavior, established and maintained by the system of preferences and expectations in a given social group. Norms, on the other hand, are more prescriptive because they formulate a commitment to behavior. Failure to meet the obligations and violation of social norms is always subject to some kind of sanction that can be formal, material, etc. In this sense, they are the reason for following them. I argue, by analogy, that the same is true of predictions. They are norms in the sense that an agent that would not act in accordance with its predictions would be exposed to certain adaptive sanctions. For example, its actions in the environment could turn out to be ineffective.

[23] By risk I mean uncertainty about predicted outcomes relative to preferred outcomes (cf. Kahneman & Tversky, 1979; Bach & Dolan, 2012).

prediction error is, so that it is possible to revise the previously adopted hypotheses (cf. Friston, 2009, p. 299).[24] Appropriate perceptual hypotheses are a basis for predictions that condition my behavior as a driver. For example, the model generates a prediction according to which if the car I am driving keeps the current direction and track, there will be a collision with the vehicle coming towards it. The error regarding the discrepancy in determining the position of the car in front of me and the possibility of a collision can be minimized in two ways: either the model will revise the priors under the influence of the prediction generated, or it will perform active inference, i.e. it will interfere with the causal structure of the world in order to minimize the error about a potential collision (cf. Friston, 2010, p. 129). By active inference I mean selective sampling of sense data so that it can be adapted to the generated predictions. In practice, this means that the agent's previous priors include the assumption that it must take actions that will minimize surprise, or a given prediction error. This means that the agent must represent itself in the expected future states by performing specific actions (Friston et al., 2014; Schwartenbeck et al., 2013, p. 2).[25] What is important in these analyses is the relation between the generated prediction and the action taken (as part of active inference).

In the analyzed example with a car, I can take both actions that will minimize the prediction error effectively and those whose effectiveness is questionable. Predicting a collision, I can, for example, pull over to the roadside or stop the car. Each of these actions interferes with the causal structure of the world in the sense that they can trigger a specific reaction in the driver of the other vehicle. However, there are actions that, despite interfering with certain states of affairs, will not minimize the prediction error. Such actions are, for example, turning on the music or air conditioning, activating the windshield wipers or talking to a fellow passenger. This means that the appropriate prediction normalizes the choice of possible actions to be taken (limited by the knowledge of the model about probabilistic relationships in the world and by specific priors regarding, for example, driving behavior). A given prediction may therefore normalize both actions that will allow for effective minimization of the prediction errors and those that do not lead to minimization, although the agent may believe that e.g. listening to relaxing music in the car will allow him or her to react faster in a situation of emergency.

Schematically, this „normalization" can be written as the following conditional:

If *'prediction'* (condition), then *'action'* (outcome).

I argue that the relation between predictions and actions, however, is not a typical causal relation that can be written symbolically as "If *B*, then *A*", but a relations that I will refer to as motivational. It can be represented as a conditional with a specific form:

[24] If the system expects a higher precision from a prediction error then the signal will be weighted higher, so the error will not be smaller.

[25] In this paper, I do not further analyze the concept of active inference, the research framework associated with this concept and the free energy principle (cf. Friston et al., 2003; Ramstead et al., 2017; Ramstead et al., 2018; Schwartenbeck et al., al. 2013).

If *B* or *C* or *D* (etc.) then *A*, but not if *E*, *F* or *G* (etc.).

Predictions condition the emergence of certain actions, thus excluding others. This thesis is justified by one of the main assumptions of PP, according to which the purpose of the model is (long-term) minimization of average prediction errors. It follows that only those actions that help achieve this goal are favored. In practice, this means that a given organism acts in a way that increases its chance of survival, i.e. reduces the degree of surprise associated with the need to act in an uncertain and an unknown environment.

The justification of the normative nature of predictive mechanisms also implies the need to refer to the interaction between the organism and the environment, which, as many researchers claim (cf. Gibson, 1979; Norman, 2013), is already pre-structured. Therefore, it should be said that the normativity of prediction is determined by specific functional roles in the generative model (including selection and guiding actions) and by reference to specific properties of the natural environment as well as socio-cultural circumstances. Due to the fact that the world is "previously" structured, it can present for the organism certain values of reward or evidence (cf. Friston et al., 2012). I will focus on this thread in § 6.

Predictive mechanisms are normative because they refer not only to the requirement of minimizing prediction errors and uncertainty, *resp*. VFE, but also to prior beliefs, preferences, or motivations that arise in relation to certain environmental states.[26] From this perspective, a possible error or misrepresentation has a normative significance for the organism not only as a potential result of specific causal processes, but above all because they significantly shape the causal transitions between states with specific content and a structured environment, thus determining the selection of appropriate actions in such a way that they correspond to the normative Bayesian rule (cf. Kiefer, 2017; Shams, Ma & Beierholm, 2005). Thus, what an organism does depends on the requirement to minimize prediction errors, individual preferences and beliefs as well as on specific properties of the environment. This means that actions are selected on the basis of some conditional potential (linked to predictions) and the relationship that the organism enters into with its environment. It is therefore appropriate to agree with Bickhard that „ Such conditional relationships can branch — a single interaction outcome can function to indicate multiple further interactive potentialities — and they can iterate — completion of interaction *A* may indicate the potentiality of *B*, which, if completed, would indicate the potentiality of *C*, and so on" (Bickhard, 2009, p. 78). This means that the cognitive system continues

[26] The full analysis of the dependencies between prediction, prior preferences from phenotype and so on goes far beyond the scope of this analysis. However, it should be emphasized that, for the argument about the strong normativity of prediction, these are not overriding issues—according to the position of motivational externalism, the very process of generating a prediction and its foundation in the generative model and / or phenotype is in some sense a separate issue from its normative function. In other words: the analysis of the causal mechanism of generating a prediction is independent of the analysis of the normative or logical function of this generated prediction. Heuristically speaking, this also means that the normativity of prediction can be investigated independently of the normativity of the Bayesian updating rule used to transform the prior into a posterior.

to predict the form and content of sensory signals, not only by using the priors and Bayesian rule, but also by actively acting in its environment.

Now let's take a closer look at the motivational relation. It is not a factual conditional (if *A* happened, then *B* happened), but rather a counterfactual conditional (if *A* had happened, then *B* would have happened). Why? Because (in PP version of this story) ignorance of the actual outcome (caused by an action) causes that the probability of a given prediction (according to Bayesian rule) is estimated only in relation to the likelihood of observations. For this reason, the model selects the counterfactual prediction that is the most likely, in the light of the data, to explain the prediction error, i.e. one for which high precision is expected. This means that the model favors such actions (as part of active inference) that are conditioned by the most probable counterfactual prediction. It speaks to the fact that uncertainty reducing policies have to be selected via a process of Bayesian model selection. This in turn rests upon the capacity to entertain counterfactual hypotheses like "what would happen, if I did that". It must also be added that ignorance of the outcome is constitutive of the necessity to make a choice between actions that are conditioned by an appropriate prediction. Thus, the expected outcome is only more or less probable. For this reason, the entire process described here specifies conditionals that are not factual but counterfactual. "Generative models underlying perception incorporate explicitly counterfactual elements related to how sensory inputs would change on the basis of a broad repertoire of possible actions, even if those actions are not performer" (Seth, 2014, p. 97). This means that generative models encode not only the likely causes of sensory signals, but also process these signals according to the a repertoire of possible actions. Based on this *knowledge*, they then make a choice between alternative policies. Vague predictions of alternative actions which minimize prediction errors and their expected consequences are ignored by the model in the light of reliable information based on the model's priors and uncertain information from input.

The policy selection process can be described in more general terms as one in which agents look backward for the best explanation for the antecedent and then forward to see whether the explanation would imply the output (Rips, 2010, p. 212). The best explanation in PP is strictly related to the abductive inference: the model abductively "infers" about the possible outcome of actions with high expected precision in such a way that it presents hypotheses that best explain information coming from the environment. For example: from the action "I will pull over to the roadside" someone can derive the outcome "There will be no collision with a driving car coming from the opposite direction". Some authors (cf. Mackie, 1974) argue that conditionals are neither true nor false, but that they only serve to highlight inferences that are permissible in a given cognitive situation, and not to state that something is true or false. Therefore, they do not speak of an ontological situation. It seems, however, that the conditionals embodied in the motivational relation occur not only in an epistemic but also an ontic situation. Counterfactual predictions, which are among the constitutive elements of the motivational relation, are inferences based on sensory information from an environment that is already pre-structured by certain constraints. According to PP, the model does not know these constraints (it can, of course, infer about them on the basis of information from input), but they are conditions for structuring this information. For example, in the process of seeing, they guarantee the matching of

appropriate elements to most natural scenes (cf. Marr & Poggio, 1976; 1979). One such natural constraint is, for example, spatial location (Marr, 1982, pp. 68–70). This means that objects in the world that cause changes in the intensity of light are spatially located. Thus, these constraints can be understood as specific facts about the real world (Shagrir, 2010, p. 489).

4 Counterfactual predictions and policy selection

The statement that the motivational relation is ontic in nature does not yet determine its normative character. The normativity of this relation is directly related to a certain arbitrariness of the choice of actions conditioned by a given prediction. Returning to the example of the weather, if I predict that there will be a storm and at the same time I want to reach the agreed place in dry clothes, my prediction is justified by the fact that, for example, I will put on a coat, order a taxi, etc. However, it does not justify turning on the TV or ordering take-away food. This means that such actions are unlikely to be undertaken by me, as they do not maximize the model evidence. This normative requirement to maximize the model evidence and the specific environmental conditions justify the normativity of prediction (in such a way that such a prediction is highly precise), i.e. they justify why the agent may choose among such and such possible actions, and why it will rather not choose from some other actions.

So it means that the agent does not have to choose that particular action (then we would be able to say that the choice is caused by a prediction), but is obligated to choose some action that is normalized by such and such counterfactual prediction.[27] Therefore, the agent is obliged to choose one of the actions offered by the generated prediction in the sense that this prediction (abductively) justifies the agent's choice. In other words, the agent somehow interferes in the world so that its actions result in the realization of one of the expected counterfactual outcomes. This obligation results from the need to minimize prediction errors, *resp*. VFE, i.e. the need to remain in existence. (cf. Hohwy, 2020b). In this way, a prediction error will be minimized

[27] It should be emphasized that according to the mathematical game theory, in most cases involving a choice of actions, but depending on the utility (risk) distribution, there may be several choices, as shown by various dilemmas related to the Nash equilibrium. Nash equilibrium assumes pairs of action strategies that provide the best (i.e. optimal) responses to each other. Once this balance is achieved in the game, neither player can improve his score by unilaterally changing the chosen strategy. There are various paradoxes associated with such a balance (e.g. the prisoner's dilemma). In such an approach, the phrase "is obligated" is strongly conditioned by the discussed equilibrium, but also by its adopted criterion.However, this remark, as I will only indicate, may raise doubts if we want to associate it with PP. First, some researchers have shown that humans, for example, unlike chimpanzees, have a marked tendency to decline offers when division deviates from equality. This means that they often act according to the principle of the "psychology of justice" which makes them extremely suboptimal in economic games where maximizing individual utility is the measure of success (cf. Jensen et al., 2007, pp. 107–109). Second, there are models (cf. Constant et al., 2019) that can explain the cooperation between human players on the basis of other factors, without making the Nash equilibrium a significant element of the explanation. This is directly related to the rejection of many assumptions of classical decision theories (cf. Friston et al., 2012; Schwartenbeck et al., 2013).

by implementing counterfactual (active) inference.[28] These observations require a clarification.

Due to the fact that the agent does not have any specified input data, it must independently find the appropriate patterns, dependencies and relationships against which it will plan its actions. It is a common view that in PP we are dealing with unsupervised learning algorithms. In practice, this means that, given a policy adopted by an agent and a set of environmental constraints, a generated counterfactual prediction will prescribe a course of action that the agent should take, assuming that they want to advance the policy as best they can, given their constraints. And here's the problem: given that agents can hold multiple policies at once, the actions related to each potentially being pragmatically in conflict, it's hard to understand how any predictions prescribing such actions could be said to be normative to the agent (at best they can be normative to the agent qua a given policy—but that doesn't settle which action should be performed, since the agent still needs to choose between policies). This problem can be solved by referring to the normative nature of the prediction: on the one hand, they are normative because they justify the choice of a given policy, which means that at the same time they can suggest a change of the currently implemented policy, if it is different from the one that regulates, or justifies the generated prediction. Predictions are therefore necessary because they drive the selection of action policies and at the same time they oblige the agent to change the policy, because the outcome expected by them maximizes the Bayesian model evidence.[29] On the other hand, they are normative because they can "ease" conflicts between existing policies. This issue requires a closer look.

Conflicts between policies can be related to the explore/exploit trade-off. The concept is taken from machine learning, but has a much wider application (cf. Cohen et al., 2007). Generally speaking, this trade-off concerns situations in which one chooses between what is known and can be foreseen (it may meet the agent's expectations)—exploitation—and what is not certain, but there is a strong assumption that it may offer some novelty in the form of information, experience or skills—exploration. Given the goals set in these analyses, I assume that the exploration-exploitation trade-off concerns (1) the choice between acquiring new knowledge and using the already existing knowledge; and (2) the choice between new non-obvious action options and proven and known action strategies. The choice between exploration and exploitation can be understood as a choice between certain behavioral tendencies. The challenge is to provide a formal account of goal-directed exploration, where agents are guided by minimizing uncertainty, *resp*. prediction error and actively learning about

[28] The agent's policy may be conditioned by "retrospective" inference (analysis of a previous decision that precipitated a negative outcome, and consideration of how events might have unfolded differently had he selected an alternative course of action) and "prospective" inference (imaginatio how various policies might play out under given circumstances) (Corcoran et al., 2020, p. 32).

[29] Normativity of predictions understood in this way can be treated as the basis for the normativity par excellence, which appears in the socially and culturally structured environment in which human agents live. Rich in material artifacts, conventions and social rules such an environment is a natural constraint for generating high-level predictions. In other words: an agent becomes a social being, i.e. a norm-sensitive entity, by being armed with a generative model that generates normative predictions that are sensitive to environmental constraints. Of course, I am only emphasizing this problem, and its full explanation goes far beyond the scope of this paper.

the world (Schwartenbeck et al., 2019). Imagine going out to the restaurant in the evening. Do we want to choose a well-known and proven place that bore us a bit, or go to another one that may positively or negatively surprise us? It is a choice between options that may have a positive or a negative outcome, which is directly related to the unexpected and expected uncertainty (Yu & Dayan, 2005).

At this point, I should refer to the concept of the expected free energy (EFE). EFE quantifies the VFE of various actions based on expected future outcomes (cf. Friston et al., 2015; Millidge et al., 2021). Why is this concept relevant to the issue of the normativity of prediction? Future actions, i.e. those to be conditioned by normative predictions, trigger future outcomes that have not yet been observed. Actions must therefore be selected in such a way that they can minimize the EFE. The already mentioned exploration-exploitation trade-off returns here, because minimizing the EFE leads to both maximization of reward and minimization of uncertainty.[30] By minimizing the EFE, the agent maximizes the expected outcomes in the exploitation of the environment. At the same time, the agent minimizes the uncertainty about the state of the world by obtaining information from the environment (exploration). This means, to use the language of active inference framework, that most actions have both pragmatic and epistemic aspects that can be associated with the already mentioned exploration-exploitation dilemma (Friston et al., 2015, p. 2). The solution to this dilemma by the agent is connected with the implementation of the normative requirement according to which the agent must minimize the EFE if it wants to solve the exploration-exploitation dilemma.

Thus, the minimization of the EFE reveals another aspect of the normativity of prediction. How the agent will minimize the EFE, i.e. whether by realizing pragmatic actions (exploitation) or by realizing epistemic actions (exploration), depends on the predictions about future actions and their expected (future) consequences (Smith, Friston & White, 2022, p. 10) i.e. predictions about the EFE. EFE is referred to both prior beliefs (that play the role of preferences), higher-order representations (e.g. the agent's model of itself) and the phenotype of a given organism, which defines a set of states that are the least "surprising" for a given agent and are consistent with its survival. In this sense, what is normative is the ability to arbitrate policies with respect to predictions, certain sets of prior beliefs, higher-order representations and an organism's phenotype[31].

Summing up, it should therefore be stated that the choice of a specific policy is related to (1) a specific environmental situation, which can be analyzed in terms of constraints for mechanisms (see § 7); (2) the generated high-precision prediction that guides future actions by minimizing the EFE; and (3) VFE (objective function), which specifies a general (normative) requirement for policies, which is minimizing expected surprise in the long-term average (Friston et al., 2017). In this understanding of the choice of policies of action, unsupervised learning becomes in PP self-

[30] This is because EFE can be decomposed into extrinsic and epistemic (or intrinsic) value. Extrinsic value refers to the utility of an expected outcome under the posterior predictive distribution (in other words, extrinsic value means outcomes preferred by the agent). Epistemic value means the expected information gain under predicted outcome or—in other words—it reports the reduction in uncertainty about hidden states based on the observations (Friston et al., 2015).

[31] See footnote 26.

supervised learning that is normatively constrained on the one hand by normative predictions and on the other hand by specific states of the world.[32]

5 Counterfactual predictions vs. semifactual predictions

According to the previous analysis the motivational relation should be written as follows:

If *B*, *C* or *D* (etc.) then *A*, but not if *E*, *F* or *G* (etc.).

This notation shows that as part of active inference the agent is obliged to choose one counterfactual prediction from among many available. For example:

"If I change lane (*B*), there will probably be no collision (*A*)"; or.
"If I change the direction of travel (*C*), there is probably no collision (*A*)"; or.
"If I turn on the radio (*E*), a collision will probably occur (*A*)"; or.
"If I start a conversation with a fellow passenger (*F*), there will probably be a collision (*A*)," etc.

Why would the model generate the prediction "If I change the lane, there will probably be no collision", rather than the prediction "If I turn on the radio, there will probably be no collision"? After all, the agent does not need to know the statistics of car accidents, their causes, or the unsuccessful attempts to avoid them (it is difficult to imagine such a situation in the modern world, but it is not impossible). The latter prediction seems unreliable. However, the agent is unable to state that with certainty. Estimates of counterfactual causal relationships between events (e.g. radio on) and their outcomes (car collision) may influence the subjective impression that some alternative variants are close to reality and others are not (Kahneman & Varey, 1990). In the context of PP, this means that the agent actually chooses between counterfactual predictions (more effective) and semifactual predictions (less effective or ineffective). Counterfactual predictions can be specified as "if conditionals"" and semifactual ones as "even if conditionals". The latter are so defined by philosophers (cf. Chisholm, 1946; Goodman, 1973) because they combine a counterfactual antecedent and a factual consequent. This means that, unlike counterfactuals about what might have been, semifactual alternatives seem to suggest that the outcome is inevitable (Byrne, 2005, p. 129), when in fact different antecedents of behaviors can lead to different outcomes/consequents. This is directly related to the features of counterfactual predictions aimed at identifying actions that will actually influence the course of events in ways that matter to the agent. For example: "Even if the driver turned on the radio, it would not avoid a collision". The situation is different with counterfactual predictions: "If the driver changed the lane, it would avoid a collision". Here the outcome suggests that it is the result of a specific action. This means that counterfac-

[32] The analysis presented here is therefore to some extent an extension of the ideas presented in Hohwy 2020b>.

tual predictions relate to how sensory stimuli would change if the agent interacted with the world in the manner suggested by these predictions, taking into account the expected consequences of these interactions. In the case of semifactual predictions, the relation between interactions or consequences is weak or absent. Again, in an effort to avoid a collision, the agent expects that it will accomplish this goal by changing the lane. Turning on the radio will certainly not help, because the avoidance of a collision is not the expected consequence or outcome of such an action.

Let's take a closer look at it. The agent, who wants to achieve the expected outcome (e.g. to avoid a collision on the road), takes actions that are conditioned by specific counterfactual predictions. Thus, if the agent is to act in such a way that selects the best policy for bringing about some outcome (in other words—that the EFE is minimized), it must act according to the predictions that suggest actions leading to the realization of these expected outcomes. It therefore means that the agent not only acts according to these counterfactual predictions but also *because of* them. Why? Because the requirement to minimize the EFE implies not only specific actions, but also recognizing minimization of the EFE as leading to actions. The driver will interact with the environment in such a way that, in his or her opinion, his or her actions that will result in avoiding a collision. Thus, the actions will not only be in line with his or her predictions, but will also be taken precisely because of these predictions. In this sense, these counterfactual predictions are *de facto* norms of what the driver has or is not supposed to do in a given situation.

Generally speaking: the agent's actions are therefore not only in line with the predictions, but are also explained by them. They provide answer to the question *why* the agent should take such and such actions if it wants to achieve specific goals (in our case this will be to avoid a collision) (cf. § 7). If this argument is correct, it justifies the fact that the predictions are normative in both the weak and the strong sense, which implies the position of motivational externalism.

It should be added that counterfactual predictions assume the minimization of the prediction error by means of active inference, which in this case means the actual interference of the agent with the causal structure of the world: lane change is the cause for avoiding a collision. Semifactual prediction assume a change in the hypothesis regarding the causes of the sensory signal source, which means that the agent minimizes the prediction error by changing, for example, its beliefs about the causal relationships between radio activation and car collisions. What I mean is that choosing a policy of action based on this semifactual prediction presupposes the change of certain parameters of the generative model first, so that "collision avoidance" is the expected outcome of the "radio on" action. More precisely, the acceptance of the truth of a semifactual prediction assumes a change in the Bayesian network, which is the generative model. For example, the prediction "If the driver turns on the radio, it will avoid a collision" assumes the introduction of a belief that turning on the radio may have an impact on avoiding a collision on the road, which in turn changes the coherence of the model. This is because such a belief is generally not substantiated by other beliefs present in the Bayesian network. This is not the case with the counterfactual prediction "If the driver changed the lane, he or she would avoid a collision"

which may be coherent with the agent's other beliefs,[33] which in practice means that the expected consequences assumed by this prediction have a specific degree of corroboration (cf. Popper, 2005, Chap. 10) in the light of the agent's actual knowledge.[34]

Therefore, it can be concluded that counterfactual predictions normalize (relatively) effective actions, and semifactual predictions normalize ineffective actions. The possibility of choosing between counterfactual predictions (with related actions) and semifactual predictions indicates the normativity of the motivational relation. In fact, the agent does not need to know which of the actions taken by it are based on a counterfactual prediction and which are based on a semifactual prediction, but is always obliged to choose an action normalized by a given prediction. And this action, let's repeat it again, is not only consistent with a given (counterfactual) prediction, but more importantly, it is taken precisely because of this prediction. The necessity to choose a given action, *resp*. policy of action, is conditioned by the necessity to minimize the emerging prediction errors. In other words, the normativity of the motivational relation is grounded in the normativity of the generative model: the generative model does not simply optimize itself (increases its coherence) in terms of both actions and perception, but needs to be optimized. Otherwise it will not exist (cf. Hohwy, 2020b; Ramstead et al., 2020, p. 233).[35]

6 Motivation and constraints

The agent is in a causal relation with the environment. Nevertheless, as has been shown, it may also be in a motivational relation with it. The motivational relation is normative, which was justified in § 3. It is also normative in the sense that it constitutes a reference of the agent to a given object in the environment that allows it to be perceived as valuable, i.e. one that evokes desire, will, aversion or disgust in the agent. In other words, when faced with certain objects, the agent may feel motivated or obliged to take a certain action or not.[36] The existence of motivational states would not be possible if a normative motivational relation between the agent and the environment had not been established. The object would also not be perceived as valuable if it did not become the pole of the motivational relation. Certain objects or states of affairs in the environment may have a special meaning for the agent, precisely as something valuable. In this sense, the agent perceives its environment not only as a

[33] About the coherence in the generative model see Kiefer, 2017. It should be added that beliefs in PP can be justified by both perceptual and other beliefs (cf. Gładziejewski, 2021a).

[34] For the above reasons, the agent assigns greater precision to counterfactual predictions and therefore they can guide its actions (cf. Seth, 2014).

[35] It can be added here that the minimization of the prediction error is normative in the sense that it is necessary to maintain the homeostatic balance of the organism, and because it obliges the agent to create such predictions about the states of the world that will allow him to choose appropriate actions by optimizing statistical information from sensory inputs (Friston et al., 2010, p. 233).

[36] This means that the choice of principles for action depends on: (1) the potential for acquiring information about future world states (i.e. the epistemic value – "where should I be if this and that"), and (2) the potential for achieving preferred sense outcomes (i.e. pragmatic value – "what should I perceive if this and that") (Constant et al., 2019).

source of information, but also as a place where its interests, desires or intentions are realized. However, these claims require further justification.

I found that there is a motivational relation between a prediction and the action taken by the agent. Without its existence, it is impossible to explain why the agent undertook such and such action in a specific situation. Its motivational nature is related to the fact that a given prediction or (Bayesian) belief, "by itself and relative to a fixed background of desires, disposes the subject to behave in ways that would promote the satisfaction of his desires if its content were true" (O'Brien, 2005, p. 56). Therefore, what makes a prediction or belief normative for a specific action or behavior is whether it plays any significant motivational role (Sullivan-Bissett, 2017, p. 95). However, I argue that what is motivational is not only the prediction itself, but most of all the relation between the prediction and the specific action given such and such environmental conditions. I claim this because the environmental conditions (understood by me in terms of constraints) co-constitute the actions directed by the predictions. What I mean is that, in the motivational relation, certain predictions, understood as probability distributions of such and such states of the world, normalize the appearance of such and such actions, thus excluding other actions (cf. § 3). Therefore, priority is given to those actions that result from the agent's motivation in relation to specific, expected states of the world. For this very reason, I argue that attention should be paid to the key role of the environment as a cause of motivational signals.

I suggest to describe the environment in terms of constraints. The concept of constraint was proposed by Pattee (1968; 1972). In his opinion, to identify constraints in a given system is to ensure a better understanding of its functioning. Pattee distinguished between constraints and laws. The latter are necessary and cannot be avoided. On the contrary, constraints are often random and relative. Constraints, unlike the laws of nature, must be a consequence of specific material structures, such as particles, membranes or, for example, machines. These structures are static, that is, to some extent dependent on the laws of nature, but their behavior can only be explained by pointing to their time-dependent constraints. It is for this reason that Pattee refers to them as "rules" (cf. Marr, 1982, pp. 22–23). In general, constraints reduce the degree of freedom of a given system with regard to the variability or the possibility of changing its parameters, components and behavior (Umerez & Mossio, 2013).

So let's consider how environmental constraints affect the functions of an organism. What an organism encounters in an environment structured by constraints constitutively influences its motivation, allowing it to reduce its uncertainty under certain conditions. Importantly, uncertainty not only has a potentially detrimental effect on the agent, but is also a motivational property (Anselme, 2010). Thus, motivation should not be treated simply as an expression of the needs of a given organism, but as a factor constituting policies aimed at seeking novelty. This is because it is directly related to the processing of information about objects in order to optimize actions (Anselme, 2010, p. 292). It consists in the fact that certain properties of the world constraint the pool of possible actions of the organism, excluding some options and pointing to others. For example: if we enclose a frog in an aquarium, the properties of such an "artificial" ecosystem will reduce the possible behavior pool of the amphibian. It will only be able to move within the boundaries of the aquarium and

"hunt" only what is in it. A trapped frog's potential behavior is governed by the environmental constraints of the aquarium in which it resides. It has been showed that, when introduced into such an ecosystem, a predator tadpole can change the shape of its body under the influence of the stress hormone, so that it is better prepared for a potential attack (Maher et al., 2013). A threat signal, i.e. an increase in uncertainty in the ecosystem, triggers a corresponding hormonal response in tadpoles.

In my interpretation of environmental constraints, information about an emerging threat motivates the organism to react in a specific way. This reaction may be, as in the case of tadpoles, a morphological change, escaping to a safe place or remaining motionless in order to prevent a predator from tracking the victim (cf. Tolledo, Sazima & Haddad, 2011). The reaction may also be a change of the car's lane when the signal of a prediction error regarding a potential collision motivates me to act in this and no other way.

It should be emphasized that the approach to motivation presented here does not define it as a mental state in the sense of folk psychology (cf. Ravenscroft, 2019), but rather as a functional role of mechanisms generating predictions (cf. Miller Tate, 2019). This claim needs clarification. Alex Miller Tate points out that "theory of motivation can only succeed if it shows how a single mental state (typically, a proximal intention or similar) can play the roles of action initiation, guidance, and control. And it's far from obvious that such a supposition is reasonable; though we might begin our theorising with a folk-psychological notion of intention, there may turn out to be no one-to-one mapping between this category and the computational / neural components of motivational architecture" (Miller Tate, 2019, p. 4). This is a significant issue to consider. This author suggests adopting a framework in which motivation is understood as states or combinations of states that play the functional roles of initiation, guidance, and control. The states thus understood "[cause] the prediction of, and selective redeployment of attention towards, action-relevant proprioceptive and exteroceptive sensory signals" (Miller Tate, 2019, p. 5). At this point, our take is broadly in line with Miller Tate's view. The main difference, however, concerns (1) paying attention to the normative nature of the motivational relation, which indicates those properties of the environment thanks to which the objects of perception become valuable, i.e. important due to the fact that they trigger the motivation of the agent. In this sense, one can speak of the agent's normative dependence on the environment; and (2) the ontic nature of the motivational relation.[37] I argue that its ontological character is determined by the fact that this relation, conditioned by certain environmental constraints, is a component of the mechanism by which we can explain the actions of the agent in the environment. I will devote my further analyses to this issue.[38]

[37] Miller Tate states: "Functionally speaking, motivational mental states, as we are working with the term here, just are the states or combinations of states that play the functional roles of initiation, guidance, and control (and perhaps some others). There is no particular reason why we should worry about whether in fact these roles are fundamentally played by a single state or many" (Miller Tate, 2019, p. 4). This approach brings him closer to what I define as the epistemic approach, i.e. one that assumes that the ascription of (e.g.) functions depends on the observer's perspective.

[38] The arguments presented here may be criticized for overestimating the role of the environment, *resp*. environmental constraints. In line with the standard interpretation of PP, generative model recapitulates the

7 Normative predictive mechanisms

Many researchers (cf. Gładziejewski, 2019; Harkness & Keshava, 2017; Hohwy, 2015) believe that the appropriate explanatory framework for PP is provided by mechanisms (cf. Craver, 2007; Bechtel, 2008; Kaplan, 2011). In this approach, PP is to be a sketch of a mechanism that will allow researchers to formulate a mechanistic explanation of specific cognitive phenomena. This means that although the brain is composed of many distinct mechanisms, these mechanisms may be unified by the fact that they fall under a common blueprint in their functional organization (Gładziejewski, 2019, p. 659). The thesis about normativity, which for me is a strong premise for adopting a realistic position in relation to PP, is grounded in the belief that mechanisms are normative as long as they allow one to explain the normativity of specific functions (cf. Garson, 2013).[39] In this sense, I claim that they can be referred to as normative mechanisms. Garson emphasizes that if we do not refer to the normative functions performed by mechanisms, it becomes difficult to explain their dysfunctions, which may lead to talking about the mechanisms responsible for dysfunctions, e.g. mechanisms that are responsible for heart attacks, malfunction of mixers or misrepresentations. Garson's thesis suggests that normativity is not only a pragmatically useful concept because of a specific research strategy, but it is a concept that seems to have a specific explanatory power. Nevertheless, one can make an objection to Garson similar to that of the teleosemantic approach: the concepts of function and dysfunction are constitutively assuming and mutually defining each

relevant causal structure of its environment and it provides all the explanatory resources one needs insofar as the scoring and selection of competing policies is concerned. In this sense, the fact that environmental states somehow discipline model predictions means that active inference doesn't just depend on predictions—it also depends on prediction errors. The account defended here clarifies what is meant by the fact that active inference depends on prediction errors. The point is that the model is related to hidden states in the world, which can be understood as specific environmental constraints or real patterns (cf. Dennett, 1991; Gładziejewski, 2021b>, p. 23), which generative models track. In order to explain the actions taken by the agent in the environment, it is important to note that it undertakes such and such actions precisely in relation to such and such constraints, *resp*. real patterns that are tracked by its generative model. In other words, the agent's actions are not suspended in a vacuum, but they are also directed by a specific, and not arbitrary, structuring of the environment (for example, the fact that a driver wants to avoid a collision can of course be treated as a top-down quasi-imperative of the statistical majority of drivers, but this specific driver avoids a collision in one way and no other, in this and no other situation. It does not depend only on the fact that the driver simply wants to avoid a collision—specific environmental conditions are also important here, which "force" whether he or she should slow down, pull over to the side of the road etc.). The explanatory role of environmental constraints lies precisely in the fact that explaining the agent's actions presupposes explaining the motivational relation, i.e. indicating the conditions that co-define the agent's success and failure. Only recently have philosophers recognized that ignoring the role of constraints in cognitive processes leads to their inadequate explanation, because "a completely unconstrained system will have no behaviors" (Winning & Bechtel, 2018, 7; cf. Winning, 2020b). I justify these remarks in the last section.

[39] One can object to the approach to normative mechanisms defended here by saying that the problem of the normativity of mechanisms comes down to the problem of normative functions, so it is enough to talk about functional mechanisms. However, I argue that the concept of normative mechanisms that I defend is justified, because every normative mechanism is a functional mechanism, but not every functional mechanism is a normative mechanism. Why? Because the normative mechanism, in my sense, is associated with the indication of constitutive causes for a given phenomenon, and not only with the indication of specific functions.

other, for there is no function without dysfunction. For this reason, it is difficult to say that this concept of function actually explains anything.

However, it can be argued, and I will defend this approach, that what is meant by describing a given mechanism or function as normative is that it plays certain causal roles (and not only functional roles, as Miller Tate claims). I argue that when speaking of such roles in relation to the normative properties of predictive mechanisms and functions, they are referred to as the causes of specific actions of an organism in the environment. In other words, normativity is a predicate with which we can explain certain phenomena (i.e. actions and behaviors) in terms of the mechanisms and functions that cause them.

Craver (2012) provides a strong argument for linking the explanation of mechanisms with their functions (see also Piccinini & Craver 2011). He points out that despite the rejection of teleological explanations by many sciences, both the physiological sciences and the neurosciences often refer to functional descriptions and these often lead to the search for mechanisms. Functional descriptions contribute to mechanistic explanations in three ways: (1) as a means of carefully pointing to appropriate etiological explanations;[40] (2) as ways of framing constitutive explanations; and (3) as ways of explaining specific items by locating them in higher order mechanisms. What is important for us is that functional descriptions are ontic in nature, which means that they are not based on the observer's decisions and research strategies, but on the actual regularities present in the phenomena (cf. Craver, 2013).

Here, I come to the conclusion of our considerations, according to which the mechanisms that serve normative functions may be the causes of specific actions and behaviors. We recognize them by identifying these and no other functions. The important point is that not all mechanisms can constitute constitutive causes of actions and behaviors, but only those that can be defined as normative. This means that I distinguish between the normative, or constitutive, causal mechanisms, and the simply causal mechanisms. For this reason, it cannot be said that the mechanism responsible for moving the hand is the constitutive cause for the fact that, for example, I wanted to drink coffee from a cup. It is of course the mechanism that co-constitutes the act of grabbing the cup and pointing it towards the mouth, but it would be a great abuse to say that it is the cause of all that might be described as drinking coffee. This is not the case with the (normative) predictive mechanism (which determines the constitutive cause of an action): the feeling of thirst and fatigue increases uncertainty in the environment. If I am sleepy and thirsty, my possibilities of action are much smaller, which may result in the emergence of situations with a greater degree of uncertainty. The counterfactual prediction that I will minimize potential uncertainty if I drink a caffeinated drink could lead to coffee consumption. Let's go further. The appropriate high-level predictive mechanism responsible for the appearance of predictions minimizing the prediction errors associated with low performance of the organism is the cause of such and such action, in this case drinking coffee. However, I distinguish the cause of a given action from its reason. Confusing these concepts may result in a false treatment of PP in terms of folk psychology. In the approach that I defend, the causes

[40] Constitutive explanations consist in explaining a given phenomenon by indicating its structural (internal) cause. In this sense, such explanations are opposed to etiological explanations (Craver, 2013, p. 151).

of actions are specific predictive mechanisms that perform such and such normative functions with regard to the requirement for (long-term) minimization of prediction errors. These are, of course, not the only causes for these actions, but they are the causes that explain the success or failure of an action taken by a given organism in a specific environment and situation. They are therefore what I call "constitutive".

It should also be added that the explanation of normative mechanisms, including predictive mechanisms, should include a description of the components and their relations, their actions and physical constraints that are jointly responsible for the appearance of a given phenomenon. It is important that not all phenomena can be explained in terms of neural mechanisms (cf. Weiskopf, 2016). Sometimes it is necessary to refer to appropriate components and operations that are also co-constituted by social and cultural constraints (Miłkowski et al., 2018, p. 9; Norman 2013). Following Marr, I understand the constraints as causal and effective, that is, those which provide the necessary and sufficient conditions for the functioning of specific processes and mechanisms (Marr, 1982, pp. 111–116). In this approach, constraints are, in a sense, norms or principles that define the boundaries and principles of realizing such and such processes. Constraints can be physical, biological, social or cultural.[41] Their analysis is crucial to explaining the constitution of a given mechanism. For example: it is impossible to satisfactorily explain the mechanism of driving a car without taking into account the physical and symbolic restrictions related to road traffic (specific regulations, knowledge of road signs, etc.).

On the basis of the above analyses, it should be concluded that constraints are an important component of mechanisms that can be used to explain relevant situations, actions, phenomena or processes. The example of driving a car in urban space shows this clearly. If you do not take into account the many possible and existing constraints of driving, then the explanation for this phenomenon is either trivial or schematic, so ultimately it is not a good explanation (cf. Winning & Bechtel, 2018). I can now explain how constraints affect motivation. What the agent encounters in an environment structured by constraints (constitutively and not merely causally) influences its motivation to reduce uncertainty. Certain physical, social, symbolic or cultural properties of the world constrain the pool of possible actions of the agent, excluding some options and pointing to others.

The predictive function of the model is obviously constrained, on the one hand, by its internal parameters (the structure and content of the Bayesian network and the Bayesian inference), and on the other hand, by certain physiological constraints (e.g. the capacity of the organ of vision, the efficiency of neural processes) or chemical constraints (e.g. chemical reactions in pyramidal cells), etc., as well as by the motivational relation. The relation is motivational due to the constraints imposed on the selection of actions by the model's predictive function—or, more precisely, the generated prediction—and certain environmental constraints. The motivational relation, which is somehow embodied in the relation of prediction and certain states of the world, is the basis for the emergence of such and such motivation in the agent. Thus, the explanation of normative predictive mechanisms should refer not only to the structure of the generative model and its parameters, but also to the already struc-

[41] On the types and applications of the concept of constraint in mechanisms, see Winning 2020b>.

tured environment as a natural constraint for the agent. Thus, the mere requirement to maximize the evidence of the model cannot constitute a sufficient justification for such and such actions or behaviors of the agent. What does constitute such justification is the existence of certain specific properties in the environment.

8 Conclusion

The choice of a given action depends on the generated prediction. The prediction selection process is based on (1) Bayesian abductive inference; (2) a policy adopted by the agent (subordinated to the requirement to minimize prediction errors, *resp*. VFE or EFE); and (3) internally (high-level beliefs and priors of the generative model) and externally (specific environmental constraints) regulated motivation. This means that action normalizing predictions are selected on the basis of a certain potential implemented in the generative model of the Bayesian network and specific relations with the environment. In practice, this means that the agent chooses among several counterfactual hypotheses about the form "what will happen, if I do this and that" (cf. Seth, 2015). Modeling probabilistic action scenarios allows for planning actions and long-term minimization of prediction errors (Pezzulo et al., 2015, p. 24; cf. also Clark 2019, p. 10). However, what distinguishes the approach proposed here from others in the literature is the emphasis on the explanatory role of environmental factors. The action is effective if the prediction generated by the model takes into account such and such states of the environment. This means that a precise, highly weighted counterfactual prediction must correspond to certain factual or counterfactual properties of the world.[42] In this way, high-level, precision-weighted predictions determine how agents act in the world. It must therefore be said, as I have already emphasized several times, that the actions taken by the agent are regulated not only by such and such predictions, but also by specific states or properties of the environment, which are understood here as constraints for the mechanism. Thanks to this, predictions not only guide actions, but also shape causal transitions between states that have specific content and satisfaction conditions (e.g. mental states). The position defended here should therefore be described as "externalist", by which I mean that the exemplification of certain mental states is conditioned both by the parameters of the generative model (e.g. specific excitation of neural populations in the case of low-level states) as well as the environmental states and socio-cultural situatedness of agents.[43] Thus, it is the normative relation of predictions with the environment that determines or specifies the content of certain states. Why normative? As I have shown, this relation, which is constitutive of the agent's interaction with the environment, cannot be reduced to the structure of the model, specific learning algorithms or the very requirements of minimizing prediction errors or maximizing the coherence of the model. Additionally, it enables explaining the meaning of cognitive errors (e.g. representational error) from the organism's own perspective, and not, as in the case

[42] Godfrey-Smith claims that the existence of a correspondence between the states of the agent and certain states of the world is a guarantee of the effectiveness of actions (cf. Godfrey-Smith, 1996, Chap. 6).

[43] This type of externalism should be distinguished from the externalism of motivation described earlier.

of e.g. teleosemantics, from the perspective of an external observer. This means that a possible error or misrepresentation has a normative significance for the organism, i.e. it affects the selection and guiding of actions, and not only a potential effect of specific causal processes.

The analyses carried out here strengthen the thesis about the normativity of predictive mechanisms. The motivational relation between a selected prediction and an action or sequence of actions taken because of that prediction can be fixed over time. Namely: the statistical effectiveness of actions taken under such and such conditions may lead to the emergence of a specific pattern of behavior. The point is that if, in certain circumstances, a certain prediction that normalizes a certain actions or their sequence leads to the goals intended by the agent, the agent may learn a certain pattern of action. This pattern can be called a "pattern of behavior" in the sense that it constitutes a matrix that determines how the agent should behave in such and such environmental conditions.[44] Thus, statistical effectiveness can lead to the emergence of certain rules of action, which can be reduced to the following form:

If under the environmental conditions X, the optimal action is S, then if agent A wants to act optimally, then A in X should perform S,,.[45]

According to the approach to normative mechanisms defended here, action S is explained by indicating its constitutive cause at the level of the mechanism, i.e. a specific prediction generated by the predictive mechanism. The justification for such action is the rule given above. Such justification can take the following form: someone performed S because under the conditions of X, such action was optimal. In this sense, indicating a rule is tantamount to giving the reason for such and such action of the agent.

Acknowledgements Previous versions of this paper were presented at the "4th Avant Conference 2019" meeting at University of Porto and during the workshop "Scaling up the Bayesian brain" in 2020 at Nova University of Lisbon. I would like to thank the organizers and participants of these events for inspiring discussions. I also thank the three anonymous reviewers for this journal for helpful discussion and for their comments on previous versions of this paper. Finally, I would also like to thank Paweł Gładziejewski, Piotr Litwin, Marcin Miłkowski, Maxwell J. D. Ramstead and Witold Wachowski who discussed with me the ideas presented here at different stages of writing this paper for their fruitful and inspiring remarks.

Declarations I have no conflicts of interest to disclose.

[44] This "should" must be understood in the sense of pragmatic effectiveness (that is, it is better to choose such a pattern of action in these circumstances than another), and not in the sense of a moral obligation.

[45] The rules formulated in this way can be transformed into unconditional norms: "Someone should do this under such and such conditions".

References

Anderson, J. R. (1990). *The adaptive character of thought*. London: Hillsdale

Anselme, P. (2010). The uncertainty processing theory of motivation. *Behavioural Brain Research*, *208*, 291–310

Bach, D., & Dolan, R. (2012). Knowing how much you don't know: a neural organization of uncertainty estimates. *Nature Reviews Neuroscience*, *13*, 572–586. https://doi.org/10.1038/nrn3289

Bechtel, W. (2008). Mental mechanisms. *Philosophical perspectives on cognitive neuroscience*. New York: Routledge

Bickhard, M. H. (2003). Process and emergence: Normative function and representation. In J. Seibt (Ed.), *Process theories. Cross disciplinary studies in dynamic (121–155)*. Dordrecht: Springer. https://doi.org/10.1007/978-94-007-1044-3_6

Bickhard, M. H. (2009). The biological foundations of cognitive science. *New Ideas in Psychology*, *27*, 75–84. https://doi.org/10.1016/j.newideapsych.2008.04.001

Bielecka, K. (2018). *Błądzę, więc myślę. Co to jest błędna reprezentacja? (I Err, Therefore I Think. What is Misrepresentation?*.) Warszawa: WUW

Brandom, R. B. (1994). *Making it Explicit*. Harvard University Press

Bogacz, R. (2017). A tutorial on the free-energy framework for modelling perception and learning. *Journal of Mathematical Psychology*, *76*, 198–211. https://doi.org/10.1016/j.jmp.2015.11.003

Byrne, R. M. J. (2005). *The Rational Imagination: How People Create Alternatives to Reality*. Cambridge MA: The MIT Press

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

Clark, A. (2016). *Surfing uncertainty. Prediction, action and the embodied mind*. Oxford: Oxford University Press

Clark, A. (2019). Beyond desire? Agency, choice, and the predictive mind. *Australasian Journal of Philosophy*, first online, 1–15. https://doi.org/10.1080/00048402.2019.1602661

Cohen, J. D., McClure, S. M., & Yu, A. J. (2007). Should I stay or should I go. How the human brain manages the trade-off between exploitation and exploration. Philos. *Trans R Soc Lond B Biol Sci*, *362*, 933–942. DOI: https://doi.org/10.1098/rstb.2007.2098

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*, 1–32. https://doi.org/10.1093/bjps/axy059

Chisholm, R. M. (1946). The contrary-to-fact conditional. *Mind*, *55*, 289–307

Christensen, W. D., & Bickhard, M. H. (2002). The process dynamics of normative function. *Monist*, *85*(1), 3–28

Christensen, W. D. (2012). Natural sources of normativity. *Studies in History and Philosophy of Biological and Biomedical Sciences*, *43*, 104–112

Constant, A., Ramstead, M. J. D., Veissiere, S. P. L., & Friston, K. J. (2019). Regimes of expectations: An active inference model of social conformity and human decision making. *Frontiers in Psychology*, *10*(679), 1–15. https://doi.org/10.3389/fpsyg.2019.00679

Corcoran, A. W., Pezzulo, G., & Hohwy, J. (2020). From allostatic agents to counterfactual cognisers: Active inference, biological regulation, and the origins of cognition. *Biology And Philosophy*, *35*, 32, 1–45. https://doi.org/10.1007/s10539-020-09746-2

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

Craver, C. F. (2012). Functions and mechanisms: A perspectivalist account. W: P. Huneman (ed.), *Functions*. Dordrecht: Springer

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

Davies, P. S. (2001). *Norms of nature: Naturalism and the nature of functions*. Cambridge: MIT Press

Dennett, D. (1987). *The Intentional Stance*. Cambridge, MA: The MIT Press

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

Elqayam, S., & Evans, J. S. (2011). Subtracting‚ought" from, ‚is": Descriptivism versus normativism in the study of human thinking. *Behavioral and Brain Sciences*, *34*(5), 233–248. https://doi.org/10.1017/S0140525X1100001X

Feynman, R. P. (1998). *Statistical Mechanics: A Set Of Lectures*. Avalon Publishing

Friston, K. J. (2009). The free-energy principle: A rough guide to the brain? *Trends in Cognitive Sciences*, *13*(7), 293–301
Friston, K. J. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*, 127–138
Friston, K. J. (2013c). Life as we know it. *Journal of The Royal Society Interface*, 10. https://doi.org/10.1098/rsif.2013.0475
Friston, K. J., Adams, R. A., & Montague, R. (2012). What is value – accumulated reward or evidence? *Frontiers in Neurorobotics*, *6*(11), 1–25. https://doi.org/10.3389/fnbot.2012.00011
Friston, K. J., Daunizeau, J., Kilner, J., & Kiebel, S. J. (2010). Action and behavior: A free energy formulation. *Biological Cybernetics*, *102*, 227–260. https://doi.org/10.1007/s00422010-0364-z
Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2017). Active inference: A proces theory. *Neural Computation*, *29*(1), 1–49
Friston, K. J., Harrison, L., & Penny, W. (2003). Dynamic causal modelling. *Neuroimage*, *19*, 1273–1302. https://doi.org/10.1016/S1053-8119(03)00202-7pmid:12948688
Friston, K., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T., & Pezzulo, G. (2015). Active inference and epistemic value. *Cognitive neuroscience*, *6*(4), 187–214
Friston, K. J., Schwartenbeck, P., FitzGerald, T., Moutoussis, M., Behrens, T., & Dolan, R. J. (2014). The anatomy of choice: Dopamine and decision-making. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1655), 1–12
Friston, K. J., & Stephan, K. E. (2007). Free energy and the brain. *Synthese*, *159*, 417–458.
Garson, J. (2013). Functional sense of mechanism. *Philosophy of Science*, *80*, 317–333. https://doi.org/10.1086/671173
Gibson, J. J. (1979). *The ecological approach to visual perception*. New York: Psychology Press
Gładziejewski, P. (2019). Mechanistic unity and the predictive mind. *Theory & Psychology*, *29*(5), 657–675. https://doi.org/10.1177/0959354319866258
Gładziejewski, P. (2021a). Perceptual justification in the Bayesian brain: A foundherentist account. *Synthese*, *199*, 11397–11421. https://doi.org/10.1007/s11229-021-03295-1
Gładziejewski, P. (2021b). Un-debunking ordinary objects with the help of Predictive Processing. *The British Journal for the Philosophy of Science*. https://doi.org/10.1086/715105
Godfrey-Smith, P. (1993). Functions: Consensus without unity. *Pacific Philosophical Quarterly*, *74*, 196–208
Godfrey-Smith, P. (1996). *Complexity and the function of mind in nature*. Cambridge: University Press
Goodman, N. (1973). *Fact, Fiction and Forecast*. Cambridge, MA: Harvard University Press
Hahn, U. (2014). The Bayesian boom: Good thing or bad? *Frontiers in Psychology*, *5*(765), 1–12. https://doi.org/10.3389/fpsyg.2014.00765
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* (16 vol., pp. 1–10). Frankfurt am Main: MI ND Group. https://doi.org/10.15502/9783958573178
Hohwy, J. (2013). *The predictive mind*. Oxford: Oxford University Press
Hohwy, J. (2014). The self-evidencing brain. *Nous*, *50*(2), 259–285
Hohwy, J. (2015). The neural organ explains the mind. In: T. Metzinger & J. M. Windt (eds.), *Open MIND*, 19(T), 1–22. Frankfurt am Main: MI ND Group. https://doi.org/10.15502/9783958570016
Hohwy, J. (2020a). New directions in predictive processing. *Mind & Language*, *2*(35), 209–223. https://doi.org/10.1111/mila.12281
Hohwy, J. (2020b). Self-supervision, normativity and the free energy principle. *Synthese*, first online, 1–25. https://doi.org/10.1007/s11229-020-02622-2
Jensen, K., Call, J., & Tomasello, M. (2007). Chimpanzees are rational maximizers in an ultimatum game. *Science*, *318*, 107–109. https://doi.org/10.1126/science.1145850
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. https://doi.org/10.1017/S0140525X10003134
Joyce, J. M. (2004). Practical Aspects of Theoretical Reasoning. In A. R. Mele, & P. Rawling (Eds.), *The Oxford Handbook of Rationality* (pp. 132–154). New York: Oxford University Press
Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *2*(47), 263–291
Kahneman, D., & Varey, C. A. (1990). Propensities and counterfactuals: The loser that almost won. *Journal of Personality and Social Psychology*, *59*, 1101–1110

Kaplan, D. M. (2011). Explanation and description in computational neuroscience. *Synthese*, *183*, 339–373. https://doi.org/10.1007/s11229-011-9970-0
Kiefer, A. (2017). Literal perceptual inference. In: T. Metzinger & W. Wiese (eds.), *Philosophy and Predictive Processing*, 17, 1–19. Frankfurt am Main: MI ND Group. https://doi.org/10.15502/9783958573185
Kiefer, A., & Hohwy, J. (2017). Content and misrepresentation in hierarchical generative models. *Synthese*, *195*, 2387–2415
Kitcher, P. (1993). Function and design. *Midwest Studies in Philosophy*, *1*(18), 379–397
Knill, D., & Pouget, A. (2004). The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends in Neurosciences*, *27*, 712–719
Korsgaard, C. M. (1996). *The Sources of Normativity*. Cambridge University Press
Kripke, S. (1982). *Wittgenstein on Rules and Private Language*. Harvard University Press
Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science*, *67*(1), 1–25
Mackie, J. L. (1974). *The Cement of the Universe: A Study of Causation*. London: Oxford University Press
Maher, J. M., Werner, E. E., & Denver, R. J. (2013). Stress hormones mediate predator-induced phenotypic plasticity in amphibian tadpoles. *Proceedings of the Royal Society B: Biological Sciences*, 280(1758), 20123075. DOI: https://doi.org/10.1098/rspb.2012.3075
Marr, D. (1982). *Vision: A computational approach*. San Francisco: Freeman & Co
Marr, D., & Poggio, T. (1976). Cooperative computation of stereo disparity. *Science*, *194*, 283–287
Marr, D., & Poggio, T. (1979). A computational theory of human stereo vision. *Proceedings of the Royal Society of London B*, 204, 301–328
Miller Tate, A. J. (2019). _A predictive processing theory of motivation. *Synthese first online*, 1–29. https://doi.org/10.1007/s11229-019-02354-y
Millidge, B., Tschantz, A., & Buckley, C. L. (2021). Whence the expected free energy? *Neural Computation*, *33*(2), 447–482.
Millikan, R. G. (1984). *Language, thought, and other biological categories: New foundations for realism*. Cambridge: MIT Press
Millikan, R. G. (1989). In defense of proper functions. *Philosophy of Science*, *2*(56), 288–302.
Miłkowski, M., Clowes, R., Rucińska, Z., Przegalińska, A., Zawidzki, T., Krueger, J., Gies, A., McGann, M., Afeltowicz, Ł., Wachowski, W., Stjernberg, F., Loughlin, V. & Hohol, M. (2018). From wide cognition to mechanisms: A silent revolution. *Frontiers in Psychology*, *9*(2393), 1–17. https://doi.org/10.3389/fpsyg.2018.02393
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 (1–10). Berlin, Heidelberg: Springer
Norman, D. (2013). *The Psychology of Everyday Things*. New York: Basic Books
Oaksford, M. (2014). Normativity, interpretation and Bayesian models. *Frontiers in Psychology*, *5*(332), 1–5. https://doi.org/10.3389/fpsyg.2014.00332
Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford: Oxford University Press
O'Brien, L. (2005). Imagination and the Motivational View of Belief. *Analysis*, *65*(1), 55–62
Pattee, H. H. (1968). The physical basis of coding and reliability in biological evolution. In C. H. Waddington (Ed.), *Towards a theoretical biology* (1 vol., pp. 33–54). Edinburgh: Edinburgh University Press
Pattee, H. H. (1972). Laws and constraints, symbols and languages. In C. H. Waddington (Ed.), *Towards a theoretical biology* (4 vol., pp. 248–258). Edinburgh: Edinburgh University Press
Peter, F., & Spiekermann, K. (2010). Rules, Norms and Commitments. In I. C. Jarvie, & J. Zamora-Bonilla (Eds.), *Handbook of The Philosophy of Social Science* (pp. 216–232). London: Sage
Pezzulo, G., Rigoli, F., & Friston, K. J. (2015). Active inference, homeostatic regulation and adaptive behavioural control. *Progress in Neurobiology*, *134*, 17–35
Piccinini, G., & Craver, C. F. (2011). Integrating psychology and neuroscience: Functional analyses as mechanism sketches. *Synthese*, *183*(3), 283–311
Piekarski, M. (2019). Normativity of Predictions: A New Research Perspective. *Frontiers In Psychology*, *10*, 1710. doi: https://doi.org/10.3389/fpsyg.2019.01710
Piekarski, M. (2021). Understanding Predictive Processing. *A Review Avant*, *1*(12), 1–48. https://doi.org/10.26913/avant.2021.01.04
Popper, K. R. (2005). *The Logic of Scientific Discovery*. London – New York: Routledge
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., Badcock, P. B., & Friston, K. J. (2018). Variational neuroethology: Answering further questions: Reply to comments on „answering Schrödinger's question: A free-energy formulation". *Physics of Life Reviews*, *24*, 59–66

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

Ravenscroft, I. (2019). Folk psychology as a theory. In: E. N. Zalta (ed.), *The Stanford encyclopedia of philosophy* (Summer 2019 Edition)

Rips, L. J. (2010). Two Causal Theories of Counterfactual Conditionals. *Cognitive Science*, *2*(34), 175–221. https://doi.org/10.1111/j.1551-6709.2009.01080.x

Rosati, C. S. (2016). Moral Motivation. In: E. N. Zalta (ed.), *The Stanford encyclopedia of philosophy* (Winter 2016 Edition)

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

Schwartenbeck, P., Passecker, J., Hauser, T. U., FitzGerald, T. H. B., Kronbichler, M., & Friston, K. J. (2019). Computational mechanisms of curiosity and goal-directed exploration. *eLife*. 2019; 8: e41703. DOI: https://doi.org/10.7554/eLife.41703

Seth, A. K. (2014). A predictive processing theory of sensorimotor contingencies: Explaining the puzzle of perceptual presence and its absence in synesthesia. *Cogn Neurosci*, *5*(2), 97–118. DOI: https://doi.org/10.1080/17588928.2013.877880

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

Shagrir, O. (2010). Marr on computational-level theories. *Philosophy of Science*, *77*(4), 477–500

Shams, L., Ma, & Beierholm, W. J., U (2005). Sound-induced flash illusion as an optimal percept. *Neuroreport*, *16*, 1923–1927. https://doi.org/10.1097/01.wnr.0000187634.68504.bb

Smith, R., Friston, K. J., & Whyte, C. (2022). A Step-by-step Tutorial on Active Inference and Its Application to Empirical Data. *Journal of Mathematical Psychology*, *107*, 102632

Sullivan-Bissett, E. (2017). Biological Function and Epistemic Normativity. *Philosophical Explorations*, *20*(1), 94–110. DOI: https://doi.org/10.1080/13869795.2017.1287296

Thagard, P. (2009). Why cognitive science needs philosophy and vice versa. *Topics in Cognitive Science*, *1*(2), 237–254. https://doi.org/10.1111/j.1756-8765.2009.01016.x

Tiehen, J. (2022). Metaphysics of the Bayesian mind. *Mind & Language*. https://doi.org/10.1111/mila.12411

Toledo, L. F., Sazima, I., & Haddad, C. F. B. (2011). Behavioural defences of anurans: an overview. *Ethology Ecology & Evolution*, *23*, 1–25

Umerez, J., & Mossio, M. (2013). Constraint. In W. Dubitzky, O. Wolkenhauer, K. H. Cho, & H. Yokota (Eds.), *Encyclopedia of systems biology* (pp. 490–493). Berlin: Springer. https://doi.org/10.1007/978-1-4419-9863-7

Weiskopf, D. A. (2016). Integrative modeling and the role of neural constraints. *Philosophy of Science*, *83*, 674–685. https://doi.org/10.1086/687854

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 vol., pp. 1–18). Frankfurt am Main: MI ND Group. https://doi.org/10.15502/9783958573024

Winn, J., & Bishop, C. M. (2005). Variational message passing. *Journal of Machine Learning Research*, *6*, 661–694

Winning, J. (2020a). Internal perspectivalism: the solution to generality problems about proper function and natural norms. *Biology And Philosophy*, *35*, 33, 1–22. https://doi.org/10.1007/s10539-020-09749-z

Winning, J. (2020b). Mechanistic causation and constraints: Perspectival parts and powers, non-perspectival modal patterns. *The British Journal for the Philosophy of Science*, *71*, 1385–1409. https://doi.org/10.1093/bjps/axy042

Winning, J., & Bechtel, W. (2018). Rethinking causality in biological and neural mechanisms: Constraints and control. *Minds and Machines*, *2*(28), 287–310. https://doi.org/10.1007/s11023-018-9458-5

Yu, A. J., & Dayan, P. (2005). Uncertainty, neuromodulation, and attention. *Neuron*, *46*, 681–692. DOI: https://doi.org/10.1016/j.neuron.2005.04.026