Review of Philosophy and Psychology
https://doi.org/10.1007/s13164-018-0416-1

Self-Knowledge in a Predictive Processing Framework

Lukas Schwengerer[1]

Published online: 23 August 2018

Abstract
In this paper I propose an account of self-knowledge based on a framework of predictive processing. Predictive processing understands the brain as a prediction-action machine that tries to minimize error in its predictions about the world. For this view to evolve into a complete account of human cognition we ought to provide an idea how it can account for self-knowledge – knowledge of one's own mental states. I provide an attempt for such an account starting from remarks on introspection made by Hohwy (2013). I develop Hohwy's picture into a general model for knowledge of one's mental states, discussing how predictions about oneself can be used to capture self-knowledge. I further explore empirical predictions, and thereby argue that the model provides a good explanation for failure of self-knowledge in cases involving motor aftereffects, such as the broken escalator phenomenon. I conclude that the proposed account is incomplete, but provides a valuable first step to connect research on predictive processing with the epistemology of self-knowledge.

Keywords Self-knowledge · Introspection · Self-ascription · Predictive processing

1 Introduction

In this paper I propose an account of self-knowledge based on a framework of predictive processing. Predictive processing understands the brain as a prediction-action machine that tries to minimize error in its predictions about the world. For this view to evolve into a complete account of human cognition we ought to provide an idea how it can account for self-knowledge – knowledge of one's own mental states. I provide an attempt for such an account starting from remarks on introspection made by Hohwy (2013). I begin with a short discussion of what an account of self-knowledge is looking for. In part 3, I provide an overview of the predictive processing framework.

✉ Lukas Schwengerer
L.Schwengerer@sms.ed.ac.uk

[1] School of Philosophy, Psychology and Language Sciences, University of Edinburgh, Dugald Stewart Building, 3 Charles Street, Edinburgh EH8 9AD, UK

Part 4 explains Hohwy's attempt to embed introspection in the framework. Part 5 develops his picture into a general model for knowledge of one's mental states. I thereby also provide some remarks on the relation of propositional attitudes to the predictive processing picture. Part 6 discusses empirical predictions and provides one case which fits with my proposed account. In part 7 I summarize the account and relate it to questions in the wider philosophical discourse on self-knowledge. I conclude that the account is promising, but incomplete.

2 Self-Knowledge

My aim is to provide an account of self-knowledge in the predictive processing framework. However, before I can do that, I need to make sure what exactly such an account aims for. What are the features of self-knowledge that ought to be explained? I do not have the space to argue for the individual features at length. Nevertheless, I presuppose three features which need to be explained. We need an account of knowledge of one's own mental states that explains:

- ***Asymmetry***: There seems to be a difference between getting to know our own mental states and the mental states of other human beings.
- ***Reliability*****:** Introspection[1] usually produces true beliefs about our mental states.
- ***Fallibility*****:** Beliefs about one's mental states formed by introspection can be wrong.

The asymmetry is often spelled out in two different aspects: privileged access and peculiar method.[2] The former states that beliefs about one's own mental states are more likely to constitute knowledge than beliefs about other people's mental states. The latter that the process generating self-knowledge is somewhat peculiar or special. In the literature this also shows up under different terminology as self-knowledge being 'immediate' (Moran, 2001; Cassam, 2011), 'transparent'[3] (Carruthers, 2011), or as being the result of a 'special access' (Fernández, 2013).[4] The extent of the asymmetry is highly debated. On one side of the spectrum self-knowledge is supposed to be luminous and infallible. That is, when one has a mental state one necessarily is in a position to know that one is in that state, and if one believes that one is in a mental state that belief is always true. A proponent on this side of the spectrum is Smithies (2016) who endorses luminosity and infallibility at least for phenomenal states. On the other end are philosophers who either disagree that we are especially reliable at assessing our own mental states (e.g. Schwitzgebel (2008)) or that the method we use to self-ascribe mental states is not as special as is often assumed (e.g. Carruthers (2011), Cassam (2014)). However, even Carruthers and Cassam endorse a minimal asymmetry. For instance, even though Carruthers (2011) argues that self-knowledge

[1] I use 'introspection' as the process of belief formation relevant to self-knowledge, not as referring to any kind of inner perception.

[2] The terminology is taken from Byrne (2005).

[3] This is not how I, or other proponents of transparency accounts use the term 'transparent', but only Carruthers' terminology.

[4] The list is not exhaustive. Moreover, Wright (1998) and Bar-On (2004) point to a similar difference on the level of speech acts.

is mostly based on the same mechanisms and resources as knowledge of other people's mental states, we are in a better position to observe ourselves. Hence, we are in a minimal sense privileged.

Reliability and fallibility are both related to the truth of a self-ascription. In our folk psychology we assume that we can know our own mental states. We often act and talk in a way that presupposes that we have self-knowledge. One of my aims is to vindicate this possibility. Hence, I aim to provide an account that includes an explanation why we can, at least often enough, correctly assess our mental states. However, I also acknowledge that we seem to be fallible. One can be wrong about one's own mental state. This is supported by empirical data such as studies on the unreliability of verbal reports shown by Nisbett and Wilson (1977) and the studies of split-brain patients analyzed by Gazzaniga (1995). Moreover, it is supported by thought experiments such as the classic case presented by Peacocke (1998):

> Someone may judge that undergraduate degrees from countries other than her own are of an equal standard to her own, and excellent reasons may be operative in her assertions to that effect. All the same, it may be quite clear, in decisions she makes on hiring, or in making recommendations, that she does not really have this belief at all (Peacocke, 1998, p. 90).

Both empirical studies and thought experiment are challenged. Wilson (2002) himself argues against the significance of the empirical studies, and Parent (2016) provides further rebuttals to empirical cases. Moreover, restricted infallibilism based on compositionality principles has been defended by Parent (2007) and Burge (1988). For this paper I presuppose that fallibilism is true, even though I do not provide any conclusive evidence for it.

The final requirement for an account of self-knowledge is its compatibility with a plausible cognitive story of the production of self-knowledge. It is not enough for an account to explain the asymmetry between knowledge of our own mental states and knowledge of other people's mental states. The account also has to give us an idea of how it relates to our cognitive architecture – or at least to our current state of research on the cognitive architecture. This is the point at which an account based on the predictive processing framework has an advantage over alternative accounts. Predictive processing is a proposal of how our cognitive makeup looks like that has a lot going in its favor. Hence, any view of self-knowledge that is compatible with a predictive processing picture of the mind is in a good position compared to its rivals. In the next section I will further elaborate what exactly this predictive processing proposal is, and why we should take it seriously.

3 Predictive Processing

The predictive processing framework provides a novel account of at least perception and action, but perhaps even the workings of the brain in general. It promises an understanding of the brain as a prediction-action machine that constantly predicts sensory input and aims to minimize error in its predictions, thereby reinventing

Helmholtz's (1860 (1962)) idea that the function of the brain is best summarized by the slogan of error correction. Importantly, prediction does not indicate a person predicting something. Rather, the notion of prediction in play is a subpersonal, automatic, probabilistic guessing as part of neural processes. Prediction in this sense is something that brains do and which enables embodied, environmentally situated agents to carry out various tasks (Clark, 2016, p. 2).

The initial idea can be explained as the brain trying to guess the causes behind sensory inputs. States of affairs in the world have an effect on the brain via our senses. The difficult task for the brain is then to figure out what these states of affairs are on the basis of the effects they have on the senses. It is easy to see that the task is difficult, because a single effect can have numerous different causes. A tree and a picture of a tree might have the same effect on our senses, but are clearly different things. Moreover, a single state of affairs in the world can have various effects. We can be related to the state in different ways. A tree looks different from far away, for instance. Predictive processing explains how our brains solve these problems. The idea is that the brain provides hypotheses about the world. It uses an internal model based on past encounters to predict what is out there. Moreover, the prediction about the world is then tested by predicting what the brain's next sensory inputs will be. Based on whether this prediction is correct, the model is updated.

Suppose I hear a sound in the middle of a weekend night. Based on previous, similar instances my brain takes the most probable cause of this noise to be my flatmate coming home and closing the front door. In the past when this happened she would go up the squeaking stairs shortly after. Hence, my brain predicts this squeaking sound to occur. And sure enough, a minute later the sound hits my ears. The prediction gets confirmed and its assigned probability increases. However, suppose the squeaking sounds had not followed. In this case the prediction would not fit and the probability of this being my flatmate would decrease. An alternative, more probable hypothesis would be put forward. In this fashion probabilities of hypotheses can be updated based on the error of the predictions. The less error in a prediction, the more the hypothesis gets confirmed. Based on this probability updating the predictive brain can learn by itself without going beyond the perspective of the skull-bound brain (Eliasmith, 2005). The brain uses its own 'bootstrapping' mechanism (Hohwy, 2013, p. 16).

The revolutionary feature of this framework is its top-down approach. The brain does not simply represent whatever input it gets, but rather builds an internal model of the world, predicts the sensory input and then modifies its own model based on the extent the predictions were incorrect. This top-down approach can provide explanations that bottom-up accounts cannot. For instance, binocular rivalry – the case in which each individual eye provides vastly different sensory input and the brain picks one over the other – can be explained by predictive processing models. The top-down approach can make sense of this case insofar as the most probable hypothesis is that there are two things out there, rather than one thing that looks so different from one eye to the other. Hence, the brain represents only one object at a time and neglects the sensory input from the other eye (Hohwy, et al., 2008; Hohwy, 2013). A recent fMRI study supports this explanation of binocular rivalry by showing that error times derived from predictive processing models correlate with neural signal time (Weilnhammer, et al., 2017). Further evidence for the predictive processing model can be found in its explanation of some mental illnesses as discussed by Clark (2016) and van Schalkwyk et al. (2017), or the account of the functioning of the retina by Hosoya et al. (2005). Friston (2005) also shows that the predictive processing story is compatible with a range of anatomical facts.

In addition to switching to a top-down generative model the predictive processing framework employs the same move in frequent iteration. A vast number of top-down predictions function together. Predictive processing models are for the most part hierarchical. Different layers of smaller internal modelling processes are related downwards, upwards and sidewards, where usually the higher layers are more general and relate to a longer timeframe. Hence, predictive processing uses a complex network of structured models related by predictions and error-signals (Rao & Ballard, 1999; Lee & Mumford, 2003; Friston, 2008). In case of perception these are tracking the external causes based on unsupervised learning (Kawaro, et al., 1993; Hinton & Zemel, 1994; Hinton, et al., 1995) (Fig. 1).

This picture of hierarchical top-down and bottom-up interaction of modeling, predicting and receiving error-signals has been enhanced recently by including action in the same framework (Brown, et al., 2011; Hohwy, 2013; Clark, 2013, 2016; Friston, 2010). The main idea is that predictions can be true in two vastly different ways: either by getting the world right, or by changing the world to fit the prediction. The former one is perception, the latter one action. I can impact the world in a way that reduces my prediction error. Understanding actions as predictions one is making true is a big step towards predictive processing as a way to explain cognition in a unified way. Hence, Hohwy (2013) and Clark (2016) provide developed accounts that combine explanations of predictions with explanations of action in a single framework. It would be another step in this direction if introspection can be understood in the same framework. The rest of this paper aims to explore one way of doing this based on Hohwy (2013).

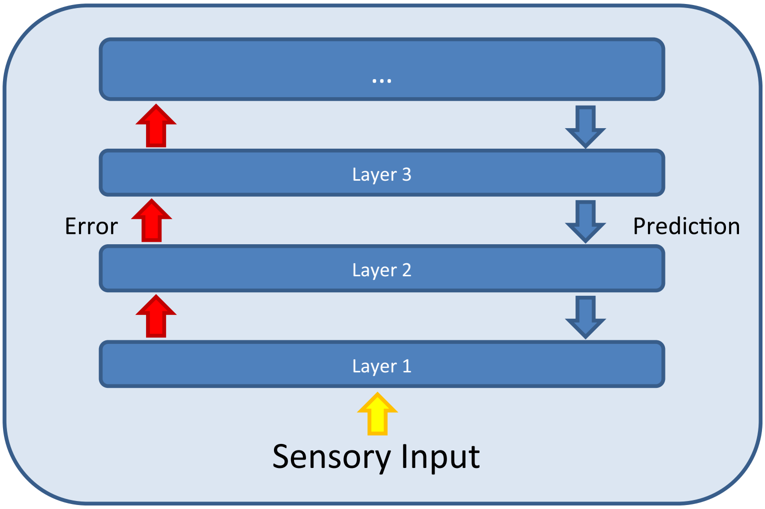

Fig. 1 An illustration of hierarchical predictive processing for perception

4 Double Bookkeeping

Hohwy (2013) proposes a way to integrate an account of introspection into the predictive processing framework. The idea is aimed at capturing self-knowledge of experiential states and for that reason rather limited. However, one can adapt the general structure as a basis to design a general framework for knowledge of one's own mental states, including propositional attitudes. To start let us sketch the model for experiential states. The motivation for the model is the phenomenon of being surprised by experiential events. A pain might be sharper than expected, or a colour experience may differ from what the brain predicted. To capture these phenomena Hohwy proposes a second hierarchical structure that works parallel to the predictive processing story on perception (2013, pp. 245–250). The brain's model of the world includes a model of experiences that creates experiential expectations. These expectations are then met with prediction error, which ultimately stems from phenomenal experience. Just as the model for perception, the model for experiences is built hierarchically. Every layer makes predictions regarding the next lower layer and receives prediction errors if these predictions are not met. And just as in the perception case these layers differ in abstractness and temporality. So whereas layer 1 might be about immediate changes in visual experiences when I turn my head, layer 3 might be about the gradual changes in experience due to natural lighting conditions (for instance gradual change in lighting in the afternoon). Hohwy is rather vague on the exact mechanisms. However, he gives us a good analogy: double bookkeeping.[5]

Experience is thought of as a byproduct that keeps shadowing perception. Shadowing here is meant to indicate two points. First, that usually perception comes with experience, so perceptual predictions plausibly also come with experiential predictions; and second, that this might go unnoticed most of the time. Both 'books' are ultimately based on sensory input. The predictions and error signal of the perception side generate the input that is supposed to match the prediction on the experiential side. We should understand that perception generally goes together with experience as the predictive process of perception being connected to the predictive process of experience. Layers of perception predictions are connected to layers of experience predictions. This is treating the "[…] deliverances of perceptual inference as causes of the input to a model" (Hohwy, 2013, p. 246) (Fig. 2).

Predictions of experiences are not only based on other layers representing experiential states. They are also connected to predictions about sensory input, and hence predictions about the world. This seems obvious when we consider the role that states of affairs in the world play in prompting experiential states. When I predict a football hitting me on my head shortly I will also predict an unpleasant experience.

Given these connections between predictions about the world and predictions about experiential states, we can explain why experiential prediction errors seem rare. Most of the time the experiential prediction errors go unnoticed, because the error is located on the perception side of processing. We are only aware that something is wrong on the experiential side of things when the prediction errors

[5] Hohwy (2013, p. 246) considers this first as an objection, but I take it to be a useful metaphor to understand the generation of self-knowledge in his framework.

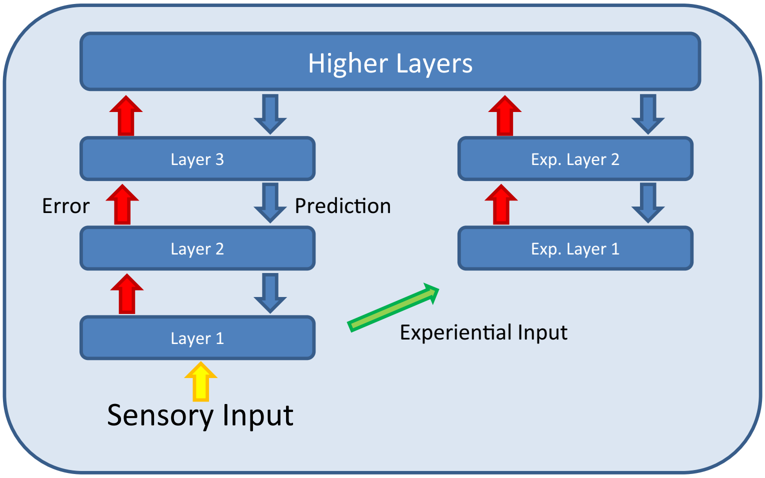

Fig. 2 An illustration of predictions of experiences in a predictive processing model

cannot be explained by inferences on environmental causes (Hohwy, 2013, pp. 247–248). This happens when our expectations of the sensory inputs are correct, but our predictions of the phenomenal experiences are not. For instance, in the case that one is surprised by a very bright light that one turned on. The brain predicts the incoming bright light, but the experiences surpass the experiential predictions. In this case any change in the perceptual prediction will create a worse fit to the sensory input, so this change is off the table. Moreover, any action that changes the input will also create a further mismatch between the sensory input and the prediction of the sensory input. So this is no option either. The appropriate adaption for the brain is to change the model of the experience that is connected to this particular sensory prediction.

In terms of Fig. 3, we have a prediction of sensory input in Layer 2, based on Layer 3. We also have an experiential prediction on Exp. Layer 1 based on Layer 2 and Exp. Layer 2. The Layer 1 prediction fits the incoming sensory input, so no relevant error is forwarded to Layer 2. However, the phenomenal experience caused by Layer 1 does not match the prediction in Exp. Layer 1. We have a prediction error at Exp. Layer 1. This error is forwarded to Exp. Layer 2 and (potentially) up to one's general model of the world and oneself. The result is an adjustment of the predictions of one's experiences. Ideally, next time one turns on the same light the expected phenomenal experience fits the actual brightness one experiences.

5 From Experience to Mental States

Hohwy (2013) introduces introspection with regard to perceptual experiences. However, because active predictive processing combines perception and action he tends to generalize at various points. He writes:

> There is a more fundamental reason to believe that creatures like us introspect. This relates to active inference and our ability to control the environment through

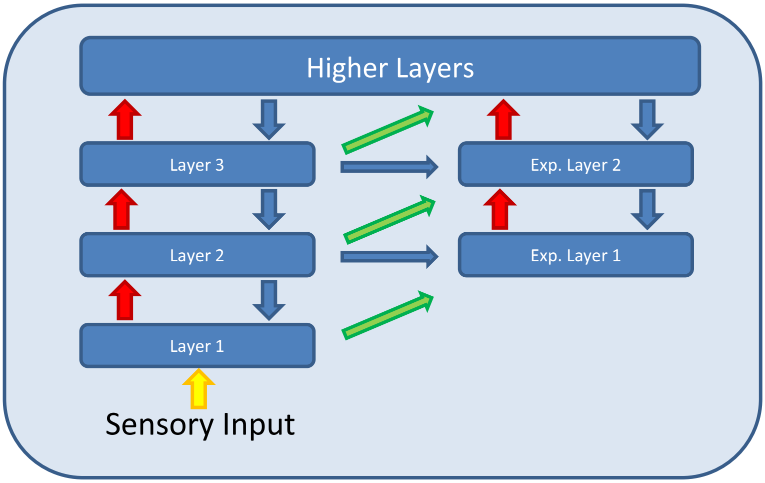

Fig. 3 An illustration of the connections between first-order predictions and second-order predictions

> action. In particular, any agent who represents its own actions must in some sense introspect. Representation of action or control of the environment is a subtle point and may only be a faculty of higher organisms; representing one's own action is not necessary for simple reflexes or homoeostasis—and yet it becomes imperative for creatures like us who engage in planning and entertaining fictive outcomes. (Hohwy, 2013, p. 247)

This seems obviously correct, but not accounted for. We can introspect all kinds of mental states and they relate very differently to action. Most importantly we can know what we desire and believe, and we can base our actions on these states. Currently the model only features predictions of sensory input (which captures the state of the world) and experiential input (which captures some states of the mind). We need to generalize the double bookkeeping picture to mental states in general to explain self-knowledge completely.

The first issue here is the question how to translate different mental states into a predictive processing picture. The straightforward case is the notion of 'belief.' Predictions share some of the relevant features of belief. Most importantly, predictions aim at truth. However, predictions can be true by different means. If you predict that a glass that you see falling towards the floor will shatter you are correct because your predictions fit the world. It is just a matter of the properties of the glass, the floor and the laws of physics that it shatters. However, you can also predict that the glass standing safely on the table will shatter in the next minute and be correct in virtue of grabbing the glass and throwing it against a wall. In this case the prediction is not true because it fits the world, but because you made the world fit to the prediction. The former prediction shows a mind-to-world direction of fit, whereas the latter shows a world-to-mind direction of fit.[6] It is central to the

[6] Anscombe (1957) provided a fantastic analogy for this difference. It is akin to the difference of consulting a shopping list to select which items to purchase (the list determines the contents of the shopping basket) and listing some actually purchased items (the contents of the shopping basket determine the list).

predictive processing picture that predictions can be true in virtue of either direction of fit. Mental states with either direction of fit can be identified with predictions. Hohwy uses this approach to describe desires:

> What drives action is prediction error minimization and the hypothesis that induces the prediction error is a hypothesis about what the agent expects to perceive rather than what the agent wants to do. If this idea is expanded to standard examples of desires, then desiring a muffin is having an expectation of a certain flow of sensory input that centrally involves eating a muffin. This means the concept of desire becomes very broad: any hypothesis associated with a prediction error can induce a want or an intention or a desire, simply because such prediction error in principle can be quenched by action.
>
> What makes the desire for a muffin a desire and not a belief is just its direction of fit. Both are expectations concerning sensory input, and the "motivator" is the same in both cases […] (Hohwy, 2013, p. 89)

There are two different ideas here. First, that a hypothesis together with a prediction error can induce a desire. And second, that a desire is an expectation concerning sensory input with a certain direction of fit. These are two very different things mentioned back to back, but they can go together in the hierarchical structure. Consider a situation in which the brain predicts sensory input from water on the tongue on Layer 1 based on Layer 2. This prediction comes with a mind-to-world direction of fit. Initially, the sensory input does not line up with this expectation, so a prediction error is fed forward. Now the brain has various options to react. A plausible reaction is a new prediction with the same content, but different direction of fit together with an action (grabbing a glass and drinking water from it). Now the sensory input lines up with the prediction. In this case a hypothesis associated with a prediction error induced a desire. The desire is also the new hypothesis, but only as a prediction with different direction-of-fit. The first hypothesis was a belief, and the second, replacing the first, is a desire. The desire was fulfilled in virtue of an action based on the desire and background beliefs (other predictions about the environment).

This picture still has challenges ahead. Direction of fit might be good enough to distinguish between belief and desire, but it is not enough to distinguish between different attitudes with the same direction of fit. Velleman (1992) argues that direction of fit is not enough to define beliefs, because "Hypothesizing that p, assuming that p, hoping that p, and the like are all attitudes in which p is regarded, not as a representation of what is to be brought about, but rather as a representation of what is" (p. 12). If this is correct – and it seems correct to me – we need more than just direction of fit.

However, there is a more pressing problem in front of us. A difficulty combining Hohwy's picture with traditional ideas on introspection is that there are different notions of 'belief' in play. Hohwy talks of beliefs and desires, but these are not quite the same things the philosophers of self-knowledge usually talk about. A belief in the predictive processing picture is not necessarily a propositional attitude. One could perhaps make the case that predictions on a high level are structured propositionally, but certainly on lower levels they need not be propositional. Rather, they are only partial representations that require the other layers to model a state of the world together. They are partial representations insofar as they make predictions and report error signals that are on their own insufficient to represent a state of the world. Call the non-propositional state ‘belief$_{\text{pred}}$’ and the propositional state ‘belief$_{\text{prop}}$.’ ‘Belief’ without qualifier includes both.

How exactly belief$_{\text{prop}}$ relates to belief$_{\text{pred}}$ is unclear. Dewhurst (2017) argues that these two conceptions are actually incompatible. The main reason for this is that the folk notion of belief$_{\text{prop}}$ is a concept that does not come in degrees, whereas beliefs$_{\text{pred}}$ are inherently probabilistic. A belief$_{\text{pred}}$ reflects the probability of a certain state of affairs. Dewhurst argues that we need to either reconsider the relation between folk psychology and the brain, or need to revise our folk notion of belief to be probabilistic. He chooses the former while attributing the latter option to Pettigrew (2015). Dewhurst’s approach aims to show that we should not use folk notions of propositional attitudes at all when trying to understand cognition on a scientific level that predictive processing does, but instead we should understand the folk notion as a broader behavior interpretational tool. He argues that this is a good way to hold on to the explanatory functions that folk psychological attitudes have in our everyday life. After all, we use folk psychology to explain and predict behavior, and to form narratives for ourselves and others (Dewhurst, 2017, pp. 6–8). Moreover, for my purpose of connecting the predictive processing framework to philosophical accounts of self-knowledge it seems at least instrumentally useful to hold on to propositional attitude talk to some degree.

Fortunately, my proposal does not require a straightforward realist conception of propositional attitudes, even though it would be compatible with that. All I need to provide is an idea how our folk notion of propositional attitudes that is at work in philosophical accounts of self-knowledge relates to predictions in the predictive processing framework. I will opt for an undemanding thesis for the relation of propositional attitudes to prediction and take on a fictionalist stance. The fictionalist stance towards propositional attitudes accepts that even though the ontological status of propositional attitudes might be dubious, it is a useful concept to take on board.[7] This is similar to treating mathematical objects as real, even though it is not fully clear what their ontological status actually is. That predictive processing leads to this fictionalist view of representation in general has been recently argued by Downey (2017). Moreover, there is good reason to opt for the fictionalist stance. In the predictive processing picture the only causally effective states are predictive states.[8] Propositional attitudes have no place in the causal explanation of predictions and actions. What attributions of propositional attitudes are mostly used for is to explain behavior on a reasonably sized scale in our ordinary linguistic practice. We tend to explain behavior in broad strokes: We explain why someone gets a drink, rather than why someone moves his hand 1 centimeter after the other. The behavior we usually want to explain is in itself a complex action built up from smaller, more basic actions. This sort of behavior is based on a multitude of predictive states (with both directions of fit), but explaining it with reference to all these different states would be rather inefficient (if not impossible)

[7] A reviewer remarked that we might question the assumption that fine-grained mental states are ontologically prior to coarse-grained ones. In this case we could consider the other option that coarse-grained attitudes like beliefs are ontologically more secure and be non-realist about fine-grained predictive states. The result would be an inverse picture of the fictionalist idea. This is an interesting option that I cannot properly look into here.

[8] It is a further topic of current debate whether these predictive states themselves count as representational. For instance, Downey (2017) denies this, whereas Gładziejewski (2016), and Wiese (2017) affirm it.

in our daily communication. Hence, we talk about propositional states that capture these various predictive states working together. The fictionalist stance claims that this is all the propositional states are: tools that let us talk about mental states more easily, even though less accurately. For my purpose it makes no difference whether propositional attitudes are states on ontologically safe ground, or whether they are states that we invented to talk as if they supervene on predictive states. Hence, I can concede to Dewhurst (2017) that our propositional attitude talk has no place on the explanatory level that cognitive scientists aim at and take his view on board. Moreover, with a commitment to fictionalism I can employ a notion of propositional attitudes that is very close to Dewhurst's own proposal. Propositional attitudes still play a role in explaining behavior in our everyday life. Their ontological status is not secure, but that does not make them any less relevant in our folk explanations. Moreover, as long as any account of self-knowledge wants to capture folk intuitions about the asymmetry of the attribution of folk psychological attitudes, we ought to explain how self-ascriptions of these folk attitudes relate to the cognitive processes going on in one's brain. This is still possible if we accept a fictionalist stance on propositional attitudes.

The question of how predictions relate to folk psychological attitudes is still not fully answered. How are they exactly connected? For instance, what beliefs$_{\text{pred}}$ are relevant for ascribing a belief$_{\text{prop}}$? I doubt that we are in a position to pick out the exact beliefs$_{\text{pred}}$ in question. However, we can observe what determines our use of belief$_{\text{prop}}$ in ordinary language and reasoning, and connect this observation to the beliefs$_{\text{pred}}$. Following Dewhurst (2017) we can identify the use of folk attitudes in predicting and explaining behavior and in our corresponding talk. Moreover, he points out that we find folk attitudes in our construction of narratives that further help to predict and explain behavior (Cf. Bruner (1990), Hutto (2008)). The important point here is that our use of propositional attitudes in folk psychology seems to be primarily about the identification, ascription, and predictive use of behavioral dispositions. Therefore Dewhurst rightly claims that "propositional attitudes were never meant to refer to fine-grained mental states, but are instead intended to pick out and predict coarse-grained behavioural patterns and dispositions" (Dewhurst, 2017, p. 10). If we now look at a description of our cognitive going-ons in the predictive processing framework we can pick out what is primarily relevant for behavioral patterns and dispositions: sets of interconnected first-order predictions on various levels. The predictions and their relations to each other taken together are responsible for behavioral patterns and dispositions. What folk ascriptions of propositional attitudes are trying to get at is the predictions and their structure that lead to actions. And in the same vein, what self-ascriptions of propositional attitudes are trying to capture is one's own structure of predictions that lead to actions.[9] An ascription of a folk psychological attitude will not allow one to identify the exact predictions in place. However, it will allow one to pick out well enough what a bunch of predictions together do. 'Well enough' here simply means that an ascription of a folk psychological attitude captures a set of predictions at work sufficiently such that it allows us to rely on the folk atittudes for behavioral predictions, explanations, and the building of narratives in our everyday life.

We can now use the idea that propositional attitudes are connected to multiple representational layers in the predictive processing framework to get a better grasp

[9] A close connection to the prediction of one's own action that fits into this picture will be discussed later.

on the problem of different propositional attitudes. Latching onto an idea of Dennett (2013) we can try to explain more complex attitudes and social features in terms of predicted processing. Dennett proposed to understand cuteness in virtue of expected expectations.

> When we expect to see a baby in the crib, we also expect to "find it cute" – that is, we expect to expect to feel the urge to cuddle it and so forth. When our expectations are fulfilled, the absence of prediction error signals is interpreted as confirmation that, indeed, the thing in the world we are interacting with has the properties we expected it to have (Dennett, 2013, p. 30).

Dennett aims this idea towards explaining features of things in the world that are not, or at least not obviously, translated into sensory input. Finding something cute is not to be identified with any individual prediction, but rather with a set of predictions. These include both predictions of sense input and experiential predictions. I want to go a step further and use the same idea to get a grasp on propositional attitudes in general.

The central idea is that different propositional attitudes can be identified with different predictive profiles that lead to different behavioral patterns and dispositions. If there is a straightforward predictive processing explanation to finding something cute, as Dennett proposes, then we can adapt the story for explaining why we find something adorable, fear something, hope for something, etc. Start with two basic prediction modes – predictions distinguished solely by their direction of fit. These predictions have content, but not propositional content. The two modes are supplemented by adjustments on precision-weightings, differences in margin of error for predictions. We can then understand all other attitudes in terms of these two modes and precision-weighting. For instance, we can understand belief$_{prop}$ as if it were a mind-to-world directed state that captures a pattern or profile of mind-to-world directed predictions.[10] Moreover, beliefs$_{prop}$ involve predictions that are quite sensitive to error signal, compared to other attitudes. Imagining that p, for instance, also captures a pattern of mind-to-world predictions, but these predictions are not, or at most barely, sensitive to error signals originating from sensory input on low levels of representation. In other words: Imagining that p is not threatened by a mismatch with the outside world, whereas belief$_{prop}$ is. Sensory input can prompt revision of beliefs$_{prop}$, but not of imaginings. A belief$_{prop}$ can also involve counterfactual predictions, as have been proposed in the predictive processing models by Friston et al. (2012), Seth (2014) and Pezzulo et al. (2015). The idea here is that counterfactual predictions capture how the sensory input (and more generally the predictive profile) would change, were we to interact with the world in a possible way. Moreover, this can be supplemented by adding further conditions that relate predictions that are part of what we take to be a belief$_{prop}$ to other types of mental states in a functionalist fashion.

Take another example. Desires$_{prop}$ can be understood as capturing a group of world-to-mind directed predictions. And desires can be differentiated from intentions by their responses to error signal, that is, by precision weighting. Similar to the belief$_{prop}$ versus imagining distinction, having an intention involves a higher sensitivity to error signal. I can desire$_{prop}$ something while not doing anything about it being the case. A desire can

[10] These formulations ought to be read with the fictionalist stance in mind.

be unfulfilled and still not prompt an action. That is, there can be significant mismatches on world-to-mind predictions that do not prompt a change on any representational layer, nor cause an action. On the other hand, if I intend something then prediction mismatches are reduced by acting. In this manner a desire$_{\text{prop}}$ shows different behavioral dispositions to an intention$_{\text{prop}}$.

Consider finally one's fearing that p. This would be understood roughly as capturing a group of world-to-mind directed predictions aiming to prevent p from occurring and, plausibly, some mind-to-world directed predictions about p being dangerous.

At this point this framework is still underdeveloped. For instance, I have not given you any reason to think that this way of identifying propositional attitudes with predictions is supposed to work for all attitudes. I also did not provide any proposals for the margins of propositional attitudes. If different predictive profiles can be identified with a folk psychological propositional attitude, then we ought to say how much the profile can change without changing the folk attitude. My description stayed on a general level, still in need of being fully spelled out. However, for the purpose of this paper it is sufficient to provide a rough idea how we need not rely on more than two different modes of predictions, both solely distinguished by their direction of fit, and precision-weighting. The important point is to provide an attempt for an account of self-knowledge in a predictive processing picture, and we now have enough resources to transform the double bookkeeping picture for perception and experience into a more general version for perception and mental states.

To label one part 'perception' and the other introspection is slightly misleading, because the same engine is at work in both. Therefore, I am going to speak of first-order processing and second-order processing for the two 'books' in Hohwy's analogy. First-order processing relates to states about the world, second-order to states about one's own mind. The basic idea is the same as in the model for expectations of experiences. The important difference is that now these are replaced with expectations of predictions on the next lower level. For instance, Layer 2 on the first-order processing strand predicts both a sensory input from Layer 1 and a prediction on MLayer 1 (together with the next higher MLayer). If everything goes right the second-order prediction on MLayer 1 matches the sensory input prediction on Layer 2. In other words: Based on Layer 3 the brain produces a belief$_{\text{pred}}$ of certain sensory input on Layer 2, and (together with the next higher layer) a belief$_{\text{pred}}$ that it believes$_{\text{pred}}$ a certain sensory input on Layer 2. Layer 2 sends prediction errors upward to both, either, or neither Layer 3 and MLayer 2 (Fig. 4).

A difficulty arises when we simply substitute expected experiences with expected predictions. The experiential strand had the neat property that it was possible to compare predicted experience with actual experience and then feed the respective prediction error upwards. However, the model does not have the capability to compare predicted prediction to actual prediction. The prediction error that is sent upwards by, say, Layer 1 is only the difference between the prediction and the sensory input received at Layer 1. Moreover, if we want to avoid any form of inner-sense model of introspection for attitudes, the model better not involve a mechanism that directly compares second-order predictions with first-order predictions, because this would be a mechanism that checks whether the second-order belief$_{\text{pred}}$ fits with the first-order state and hence requires a way to access the complete first-order state.

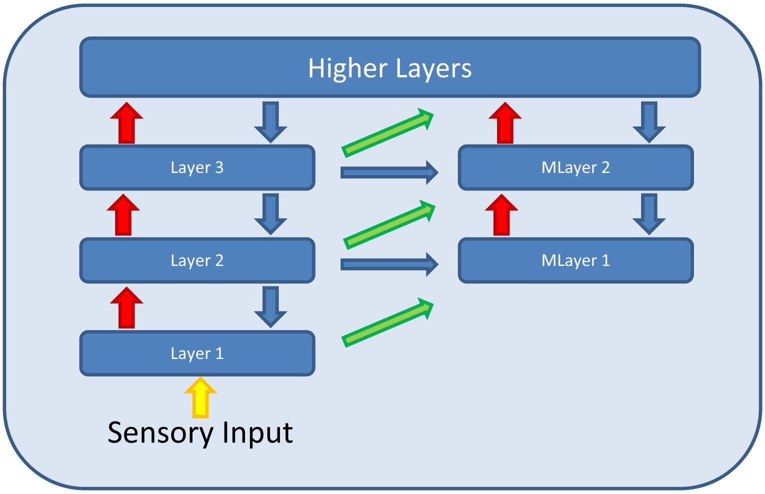

Fig. 4 An illustration of predictions of mental states in a predictive processing model

There might be a conceptual way around this problem. So far I presented the second-order predictions as predictions of predictions. However, we can think about them slightly differently in terms of predictions of an activity. That is, based on, say, Layer 2 the brain predicts that it is going to predict a certain sensory input at Layer 1. This prediction of an activity is captured in MLayer 1. This is not a switch in the notion of prediction in play. There is no relevant difference between 'prediction of prediction' and 'prediction of predicting.' Both are spelled out in exactly the same way in the predictive processing framework, i.e. as predicting certain neuronal activity. However, this different point of view gives us the conceptual option that predicting an activity might also be a case of predicting a certain movement of one's body – predicting one's action. It is a prediction of the way in which one is going to minimize the error in Layer 1. Predictions of one's activity in this sense can be compared with one's body's actual movement. Hence a prediction error can go upstream.

The big question here is then how this prediction error can relate the second-order predictions to the first-order predictions. My proposal here is the following: active predictive processing allows for the brain to secure the correct fit for its predictions about the world in two ways: Either by changing the model and thereby the prediction, or by changing the world and thereby acting. Any Layer in the first-order strand is capable of eliciting action. These acts can be compared with the predicted action based on the second-order strand to generate prediction error for the second-order strand (Fig. 5).

Comparing the predicted action and the actually performed action does not require any direct access to first-order mental states, so we can avoid falling back into an inner-sense view of introspection. Importantly, predicting an action is also nothing else but predicting a certain sensory input. Moreover, because the second-order processing ends up with a prediction of a certain sensory input, and this can be compared with the sensory input that actually arrives, we can describe how second-order processing can be trained. It can be trained because whenever there is no, or little prediction error on the first-order processing side of things, while there is still a prediction error between predicted action (in terms of predicted sensory input) and

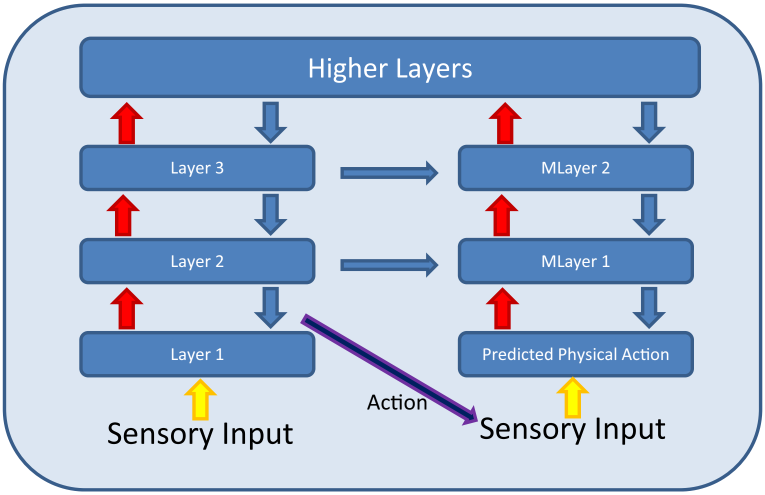

Fig. 5 An illustration of self-knowledge in a predictive processing model

actual action (in terms of sensory input) this can be fed upwards to change the second-order processing model. In case there is a prediction error in the first-order side then the prediction error on the second-order strand is irrelevant, because the problem might be inherited from the first-order processing.[11]

We can describe the brain according to this model as permanently keeping track of itself. It tries to predict the world, and itself. At both levels the brain improves its models by noticing whether the predictions have been erroneous.

6 Empirical Requirements, Predictions, and Support

Double bookkeeping does not imply doubling the channels of sensory inputs. All that is required is that the sensory inputs are processed in a way that generates two prediction errors that are fed forward. One related to the low level predictions of the incoming sensory signals in general, and one related to the sensory signals connected to one's own body in the world. To do this, sensory signals need to be understood not as merely input from outside (exteroception), but also include proprioception and interoception. Importantly, it would be a mistake to think of the prediction of actions (in terms of sensory input) only on a proprioceptive and interoceptive basis. Multiple experiments indicate that exteroceptive input is an important factor in locating one's own body and actions (Maravita, et al., 2003; Blanchard, et al., 2011; Guerraz, et al., 2012). Moreover, in some cases visual input seems to be the dominate factor (Lishman & Lee, 1973; Botvinick & Cohen, 1998; Drummer, et al., 2009). Therefore exteroception has to be included in the process generating prediction errors related to one's actions.

My proposed picture of double bookkeeping uses two prediction errors based on sensory input for the two different strands of predictions: first-order and second-order predictions. Either strand can be trained in virtue of prediction errors and

[11] However, it might be possible to train both sides at the same time. I want to remain neutral on this issue.

change of the internal models underlying the predictions. If this is correct, we can expect to find indications of these trainings. Training the first-order processing is uncontroversial, because it only requires that the internal model of the world (or part of it) changes when the predicted sensory input does not match the actual one. Training in this sense is just learning about the world. I focus on training for the second-order predictions, which ultimately end in predictions of one's action (in terms of sensory input). Training here can only come from a mismatch between the predicted action and the actual action. That is a mismatch between the predicted sensory input of one's own bodily movement and the actual one. In other words, training of the model happens at moments of surprise. We can capture surprise in a quantifiable way by ignoring the phenomenal elements of it. Surprise in this sense becomes "the negative log-probability of an outcome"(Friston, 2010, p. 128). Following this definition lower subjective probability of an outcome comes with higher surprise if it occurs. Tribus (1961) uses 'surprisal' instead of 'surprise' to mark this difference between the phenomenal surprise, and the quantifiable negative probability of an outcome.

Given that successful prediction requires minimization of surprise, we can expect that a brain that is still at an early stage of training its models (and hence its predictions) will be met with surprising events frequently, whereas a more refined, well-trained model meets less surprises. For the prediction of one's own behavior based on second-order predictions this implies that we should expect cases in which one is surprised by one's own behavior. Moreover, we should expect that how exactly these cases look will differ from infants and young children to adults, because infants and young children predict in a coarser grained manner than adults (Baillargeon, 1994a, 1994b).

It is difficult to know when infants and young children are actually surprised by their own behavior. Methodologically it seems to be easier to locate this phenomenon in older children and adults who can report on the cause of their surprise. In the following I discuss a case of being surprised by one's own behavior. This is the everyday, benign surprise of the 'Broken Escalator Phenomenon.' This also has the advantage of being an easily accessible and relatable case. Even though I will look at an experimental setting, the phenomenon is one with which most of us are familiar with.

Reynolds and Bronstein (2003) studied motor aftereffects by using a combination of a fixed platform and a mobile sled. Subjects were trained by moving from the fixed platform onto the mobile sled for 20 trials. After the moving trials subjects were given clear, verbal warnings that the sled would keep stationary for the next trials. Subjects were asked whether they heard and understood the warning before walking another 10 trials. Body positioning and walking velocity were recorded for all trials. This setup is not too different from other, previous motor aftereffect experiments tracing back at least to Held (1965). The interesting part of Reynolds' and Bronstein's experiment is that they included reports of the subjects experiencing the effect. The bodily result of their experiment is unsurprising to anyone familiar with broken escalators. After about five trials with the moving platform the subjects adapted to the movement. Their velocity increased and soon after their posture changed as well, leaning slightly forward. The trials on the stationary sled afterwards showed that they still compensated for the moving sled, even though they reported full awareness of the sled being stationary. Moreover, they reported being surprised by their own behavior. Reynolds and Bronstein write:

> Most subjects spontaneously expressed great surprise and amusement when, on walking on the stationary sled, the aftereffect occurred. When subsequently questioned, however, all confirmed that they had understood and believed the experimenters' warning that the sled would not move. Those subjects who could relate to the broken escalator phenomenon found the experimental aftereffect to be similar to the real-life experience (Reynolds & Bronstein, 2003, p. 305).

Subjects are surprised by their own behavior. They claim to know that the sled is not going to move but nevertheless they behave as if they were stepping on a moving sled. Reynolds and Bronstein interpret this as the motor system acting inappropriately even though perception and cognition are veridical (Reynolds & Bronstein, 2003, p. 306). They generalize and propose an explanation in terms of "dissociation between declarative and procedural systems in the central nervous system" (Reynolds & Bronstein, 2003, p. 308).

How does this relate to my proposed picture of double bookkeeping? First a note of caution: There is a danger of identifying the quantifiable, technical use of surprise (the surprisal) with the phenomenal experience of being surprised. There might be a disconnection between surprise (surprisal), in the sense of a mismatch between predicted, and actual sensory input, and experienced surprise. Clark (2013) gives the example of an elephant on a magician's stage. One might be surprised by the elephant, even though the brain gives this state of the world a high probability (low surprisal). However, both senses of surprise are easily reconciled. Even though the elephant might be predicted as quite probable now, it wasn't a moment ago before it showed up. Prediction error from sensory inputs caused a revision of the internal models that then in turn have the elephant as a probable hypothesis (Clark, 2013, p. 16). It is plausible that the phenomenal surprise is therefore caused by an initial surprise (surprisal) on the level of prediction. It then may last, even though the mismatch is resolved. This is enough to infer some suprisal from reported surprise, and allows us to use the relatable case of the broken escalator to illustrate how one assesses one's own mental states.

The subjects being surprised in both the phenomenal and technical senses marks a mismatch of the predicted behavior (action) and the actual behavior. However, it does not seem to be all that clear whether perception and cognition are veridical. Rather, what we can know from the reports of the subject is that reports of perceptual beliefs and cognition fit with how perception and cognition ought to be in the subject's situation. Subjects report that they believe that the sled is stationary. By these means we cannot know whether subjects really believe$_{\text{prop}}$ it is stationary. Their behavior indicates to the contrary that they do not believe$_{\text{prop}}$ it is stationary. If we rate the behavior as a better indicator than the report, then what we find here is a mismatch between belief$_{\text{prop}}$ and second-order belief$_{\text{prop}}$. Given that beliefs$_{\text{prop}}$ entail beliefs$_{\text{pred}}$, we have a mismatch between prediction and second-order prediction. This mismatch is then iterated in the next step insofar as the actual action and the predicted action do not match.

In this picture the mismatch starts when at the first-order strand a moving sled is predicted, while at the second-order strand a prediction of a stationary sled is predicted. The former is based on the previous experiences in the setting, the latter also on the testimony by the experimenter.[12] When the subject declares that she knows that the sled is stationary, she does so in virtue of the prediction that she predicts a stationary sled.[13] That is, she declares what she does in virtue of her second-order belief. This mismatch is then reiterated on the lower levels. However, the mismatch need not be only in terms of content. Both strands match insofar as one predicts input based on steady movement, and the other expects a prediction based on steady movement. The difference here has to be in the direction of fit of the prediction. While the steady movement prediction on the left side is based on a moving sled, the second-order prediction on the right is based on the second-order prediction of predicting a stationary sled. The impact is that on the left side the aim to reduce prediction error results in an action to counteract the moving sled. Steady movement on a moving sled requires changes in velocity and posture. The final left slide prediction is therefore a world-to-mind directed one, such as desire, or intention. On the right side the expected prediction of steady movement is based on the expected prediction of a stationary sled, hence no need for an action is present. Here we have a mind-to-world direction of fit – a belief (Fig. 6).

When Reynolds and Bronstein talk about perception and cognition being veridical they point to the fact that the predicted stationary sled belief (A) and the predicted steady movement belief (B) are exactly what subjects should have in the situation. However, their actual perception might not be veridical. Rather, they perceive the situation as if there was a moving sled, which causes subjects to counteract the moving sled. They are then surprised by their own action, because the action is based on a moving sled, whereas the predicted action is based on a stationary sled. This still fits well with the general explanation as “dissociation between declarative and procedural systems in the central nervous system,” (Reynolds & Bronstein, 2003, p. 308) if we identify the declarative system with the right (the second-order) strand, the procedural system with the left (the first-order) strand.

Surprise in this case is not limited to the subject’s own behavior. Surprise also occurs as a result of the overcompensation of the assumed moving sled. The brain predicts a steady movement and wants to match this prediction with the sensory input by performing a certain action (moving faster, changing posture). It turns out this action increases prediction error, so the internal model has to change to adjust the prediction. After enough iteration the internal model will be on a stationary sled model, in which case the actual action matches the predicted action. There is no need to change the second-order models if a prediction error at the first-order side was reported. However, we can expect different cases in which there is no or no significant prediction error on the first-order side, but a significant prediction error on the second-order side, based on a mismatch of action and predicted action. In this case the models of the second-order side have to be changed to minimize prediction error.

[12] A different interpretation is that both bases include the testimony, but the weighting of the testimony differs.

[13] There might not be a single prediction that captures the stationary sled. Instead this state of affairs will be represented over multiple predictions on different layers. For simplicity I treat this as the content of a single prediction here. The same goes for all other predictions in this description.

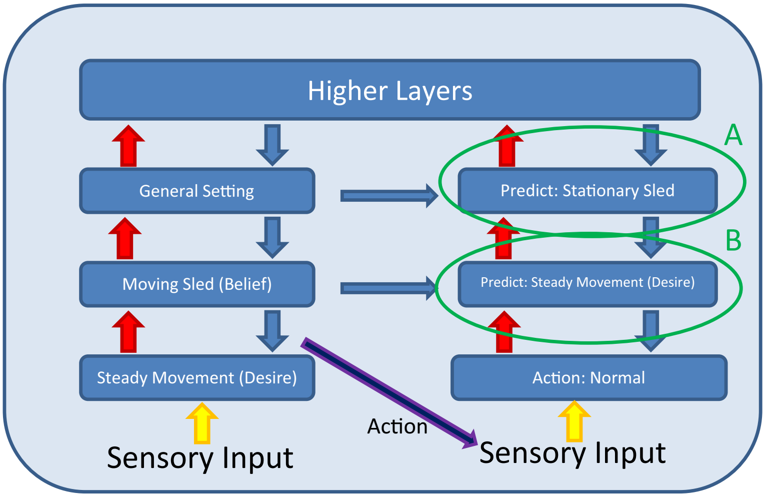

Fig. 6 Applying the model to the 'broken escalator phenomenon'

7 Explaining Self-Knowledge

I can now formulate my proposed predictive processing story of self-knowledge for propositional attitudes: self-knowledge is based on second-order predictions (mind-to-world directed). You believe$_{\text{prop}}$ that you believe$_{\text{prop}}$ that p, if you expect yourself to have a certain predictive profile. And you expect yourself to have a certain predictive profile in virtue of multiple second-order predictions, each second-order prediction a prediction of a prediction or action. Furthermore, you can only know that you believe that p, if (i) your process of generating second-order predictions is generally reliable,[14] and (ii) your current, relevant second-order predictions are largely correct.[15] Largely, because it does not seem necessary that every single second-order prediction is correct. It is conceivable that a single second-order prediction does not match any first-order prediction, but the pattern of second-order prediction nevertheless realizes a second-order belief$_{\text{prop}}$ that matches a first-order belief$_{\text{prop}}$. There seems to be some margin of error. This fits with the undemanding fictionalist stance that merely requires that our notion of belief$_{\text{prop}}$ allows us to make predictions and explanations of behavior. As long as this function is fulfilled beliefs$_{\text{prop}}$ can be imprecise. Propositional attitudes therefore need not allow us to infer the exact predictions in place. Moreover, different sets of predictions might instantiate behavioral dispositions that are similar enough for our purposes in folk psychology, which gives us further reasons not to attempt to infer specific predictions from the ascription of a propositional attitude.

The reliability of second-order predictions is established by a feedback loop based on expected actions and actual actions. This way the account gives us an explanation of why we can accurately ascribe our own attitudes. However, the account also has room for fallibility. It is possible that one's second-order predictions are wrong and that one's prediction of an action is mistaken. Moreover, the framework provides room for such mistakes in self-knowledge to be consistent over time. Suppose that one predicts a

[14] i.e. 'justified' in the reliabilist sense

[15] i.e. your second-order belief$_{\text{prop}}$ is true

specific action based on a series of second-order predictions. Further suppose that this action does not actually occur. The brain then generates an error signal based on the mismatch of predicted action and actual action. In these cases the brain always has to decide which one it should trust, the priors which generated the prediction, or the incoming prediction error. The mistake may be located in either of those. In a usual case the incoming sense data that is responsible for the prediction error will be trusted and the priors adjusted accordingly. The model recalibrates itself based on this feedback loop. However, in some cases the priors might have such a high assigned probability that it seems more likely that the prediction error itself was a mistake. In that case, the prediction error will be disregarded and the priors stay as they are. Take Peacocke's (1998) case of the biased administrator from the beginning. The administrator takes herself to be fair and unbiased. We can suppose that her brain has second-order predictions that culminate in a prediction to act in a manner that treats all applicants the same. If she now receives a mismatching sensory perception that can be interpreted as treating applicants from other countries worse, her brain has to make a choice (so to speak): was she wrong about her predictions of herself, or is the reported error a false positive? If the latter choice is deemed more probable, then there will be no change to her priors and predictions. She will still take herself to be fair and unbiased, even though her actions tell a different tale. The persistence of the undetected bias is explained insofar as actions that do not fit with her view of herself are not entering the feedback loop that impacts her predictions. Instead, these error signals themselves will be reinterpreted to fit with her priors.

For a full account we need to combine this with an explanation for knowledge of experiential states, and knowledge of non-propositional attitudes. If my proposed model can work for propositional attitudes, then there seems to be no good reason why it should not for non-propositional attitudes. On the contrary, it should be easier to achieve this, insofar as predictions are usually taken to be non-propositional. Believing that I fear a spider would be a pattern of second-order prediction, just as believing that I believe that 'this grass is green' is. For experiential states Hohwy's (2013) explanation is a starting step, but needs to be spelled out in more detail – a project I leave open here.

However, if my proposed model turns out to be correct, the main lesson we should take away from it is that forming second-order beliefs$_{\text{prop}}$ ought to be understood as part of the same process as first-order belief$_{\text{prop}}$-formation. Second-order predictions are not formed independently of first-order predictions. Rather, they are connected downward and sideward. All part of a single machinery trying to predict the world and itself. We should not think of introspection as a distinct belief-forming process that detects our mental states, but as a built-in part of our predictive processes. Self-knowledge can thereby be at least partially based on the very same thing that is the basis for first-order states. The proposed predictive processing model therefor qualifies as a version of a transparency account of self-knowledge, which proposes that a mental state and knowledge of that mental state can be formed based on the same outward phenomenon (Evans, 1982). Outward phenomena in the predictive processing case are predictions (in either direction of fit) about the world.

Furthermore, the proposal can account for the privileged nature of self-knowledge compared to knowledge of other people's mental states. Predictions of predictions and predictions of actions are available differently for oneself and for other persons. For other persons I only have access to observations of their behavior which can be a basis for predicting their mental states (mind-reading). In contrast, my brain has access to fine-grained first-order predictions as a basis for second-order predictions. This availability of sub-personal predictions accounts for the high accuracy of self-knowledge, and the peculiar nature of self-knowledge. Because these predictions are made on a sub-personal level, self-knowledge appears immediate or groundless. I do not have conscious access to the basis of my second-order belief$_{\text{prop}}$, because I have no conscious access to the basis of any involved belief$_{\text{pred}}$. Notice that the basis for the second-order belief$_{\text{prop}}$ is not the first-order belief$_{\text{prop}}$, but rather the individual first-order beliefs$_{\text{pred}}$ and relevant higher level second-order predictions. This is the result of treating the belief$_{\text{prop}}$ as a pattern of beliefs$_{\text{pred}}$. The belief$_{\text{prop}}$ is not taken to be a single entity. Instead it is a group of beliefs$_{\text{pred}}$ taken together that we identify with our folk notion of belief in a fictionalist manner.

8 Conclusion

I provided a predictive processing account for self-knowledge. I proposed that we should understand self-knowledge of mental states based on a double bookkeeping picture of predictive processing. Two connected predictive strands run in parallel, one involving predictions about the world, and the other one involving predictions about predictions or one's actions. These predictions can come in two directions of fit: mind-to-world and world-to-mind. I suggested that we should identify our folk notions of propositional attitudes with patterns or profiles of predictions in merely a fictionalist manner to keep the connection between our ordinary talk and our study of cognition. If we do so, we can accept second-order propositional beliefs as capturing a set of second-order predictions. Hence, you believe$_{\text{prop}}$ that you believe$_{\text{prop}}$ that p, if you expect yourself to have a certain predictive profile. And you expect yourself to have a certain predictive profile in virtue of multiple second-order predictions, each one a prediction of a prediction or action. Furthermore, you can only know that you believe that p, if your second-order predictions are largely correct and the process generating them is reliable. This picture explains the nature of self-knowledge as peculiar and privileged, because this basis for your second-order predictions is different from the observational basis of beliefs formed by mind-reading. The model is still incomplete. We require a better understanding whether the proposed relation between propositional attitudes and non-propositional predictions is plausible, and how second-order predictions fit into child development. Moreover, we need a better grasp on how the predictive processing model works for different mental states, if these mental states are not to be eliminated. I merely provided a first step towards a predictive processing account of self-knowledge.

Acknowledgments Special thanks to Michel Croce, Giada Fratantonio, Matt Jope, Jesper Kallestrup, Aidan McGlynn, Kegan Shaw and two anonymous reviewers for feedback on earlier versions of this paper.

Funding This article was written while funded by the UK *Arts and Humanities Research Council* doctoral studentship (Ref. 1654450).

Compliance with Ethical Standards

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

References

Anscombe, G.E.M. 1957. *Intention*. Oxford: Basil Blackwell.

Baillargeon, R. 1994a. How do infants learn about the physical world? *Current Directions in Psychological Science* 3 (5): 133–140.

Baillargeon, R. 1994b. Physical reasoning in young infants: Seeking explanations for impossible events. *British Journal of Developmental Psychology* 12 (1): 9–33.

Bar-On, D. 2004. *Speaking My Mind*. Oxford: Oxford University Press.

Blanchard, C., R. Roll, J.-P. Roll, and A. Kavounoudias. 2011. Combined contribution of tactile and propriocetive feedback to hand movement perception. *Brain Research* 1382: 219–229.

Botvinick, M., and J. Cohen. 1998. Rubber hands 'feel' touch that eye see. *Nature* 391: 756.

Brown, H., K. Friston, and S. Bestmann. 2011. Active inference, attention and motor preparation. *Frontiers in Psychology* 2 (218).

Bruner, J.S. 1990. *Acts of meaning*. Cambridge: Harvard University Press.

Burge, T. 1988. Individualism and self-knowledge. *The Journal of Philosophy* 85 (1): 649–663.

Byrne, A. 2005. Introspection. *Philosophical Topics* 33 (1): 79–104.

Carruthers, P. 2011. *The opacity of mind: An integrative theory of self-knowledge*. Oxford: Oxford University Press.

Cassam, Q. 2011. Knowing what you believe. *Proceedings of the Aristotelian Society* Volume CXI, pp. 1–23.

Cassam, Q. 2014. *Self-knowledge for humans*. Oxford: Oxford University Press.

Clark, A. 2013. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences* 36 (3): 1–24.

Clark, A. 2016. *Surfing uncertainty*. Oxford: Oxford University Press.

Dennett, D.C. 2013. Expecting ourselves to expect: The Bayesian brain as a prejector. *Behavioral and Brain Sciences* 36 (3): 29–30.

Dewhurst, J. 2017. Folk psychology and the Bayesian Brian. In: T. Metzinger & W. Wiese, eds. *Philosophy and Predictive Processing*. Frankfurt am main: MIND group.

Downey, A. 2017. Predictive processing and the representation wars: A victory for the eliminativist (via fictionalism). *Synthese*.

Drummer, T., A. Picot-Annand, T. Neal, and C. Moore. 2009. Perception. *Movement and the rubber hand illusion* 38: 271–280.

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

Evans, G. 1982. *The varieties of reference*. Oxford: Oxford University Press.

Fernández, J. 2013. *Transparent minds: A study of self-knowledge*. Oxford: Oxford University Press.

Friston, K. 2005. A theory of cortical responses. *Philosophical Transactions of the Royal Society of London. Series B. Biological Sciences* pp. 181-197.

Friston, K. 2008. Hierarchical models in the brain. *PLoS Coputational Biology* 4(11).

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

Friston, K., Adams, R. A., Perrinet, L. & Breakspear, M. 2012. Perceptions as hypotheses: Saccades as experiments. *Frontiers in Psychology* 3(151).

Gazzaniga, M. S. 1995. Consciousness and the cerebral hemispheres. In: M. Gazzaniga, ed. *The Cognitive Neurosciences*. Cambridge: MIT Press.

Gładziejewski, P. 2016. Predictive coding and representationalism. *Synthese* 193 (2): 559–582.

Guerraz, M., S. Provost, R. NARISON, A. Brugnon, S. Virolle, and J.P. Bresciani. 2012. Integration of visual and proprioceptive afferents in kinesthesia. *Neuroscience* 223: 258–268.

Held, R. 1965. Plasticity in sensory-motor systems. *Scientific American* 213: 84–94.

Hinton, G.E., and R.S. Zemel. 1994. Autoencoders, minimum description length and Helmholtz free energy. In *Advances in neural information processing systems 6*, ed. J. Cowan, G. Tesauro, and J. Alspector. San Mateo: Morgan Kaufmann.

Hinton, G.E., P. Dayan, B.J. Frey, and R.M. Neal. 1995. The wake-sleep algorithm for unsupervised neural networks. *Science* 268: 1158–1160.

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

Hohwy, J., A. Roepstorff, and K. Friston. 2008. Predictive coding explains binocular rivalry: An epistemological review. *Cognition* 108 (3): 687–701.

Hosoya, T., S.A. Baccus, and M. Meister. 2005. Dynamic predictive coding by the retina. *Nature* 436: 71–77.

Hutto, D.D. 2008. *Folk psychological narratives: The Socioculteral basis of understanding reasons*. Cambridge: MIT Press.

Kawaro, M., H. Hayakama, and T. Inui. 1993. A forward-inverse optics model of reciprocal connections between visual cortical areas. *Network* 4: 415–422.

Lee, T. & Mumford, D. 2003. Hierarchical Bayesian inference in the visual cortex. *Journal of Optical Society of America*.

Lishman, J.R., and D.N. Lee. 1973. The autonomy of visual kinaesthesis. *Perception* 2: 287–294.

Maravita, A., C. Spence, and J. Driver. 2003. Multisensory integration of the body Schema: Close to hand and within reach. *Current Biology* 13: R531–R539.

Moran, R. 2001. *Authority and estrangement*. Princeton: Princeton University Press.

Nisbett, R., and T. Wilson. 1977. Telling more than we can know: Verbal reports on mental processes. *Psychological Review* 84 (3): 231–259.

Parent, T. 2007. Infallibilism about self-knowledge. *Philosophical Studies* 133 (3): 411–424.

Parent, T. 2016. The empirical case against Infallibilism. *Review of Philosophy and Psychology* 7 (1): 223–242.

Peacocke, C. 1998. Conscious attitudes, attention, and self-knowledge. In *Knowing our own minds*, ed. C. Wright, B. Smith, and C. Macdonald, 64–97. Oxford: Oxford University Press.

Pettigrew, R. 2015. Pluralism About Belief States. *Proceedings of the Aristotelian Society Supplementary Volume* 89 (1): 187–204.

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

Rao, R.P.N., and D.H. Ballard. 1999. Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience* 2 (1): 79–87.

Reynolds, R.F., and A.M. Bronstein. 2003. The broken escalator phenomenon. *Experimental Brain Research* 151: 301–308.

Schwitzgebel, E. 2008. The unreliability of naive introspection. *Philosophical Review* 117 (2): 245–273.

Seth, A.K. 2014. A predictive processing theory of sensorimotor contingencies: Explaining the puzzle of perceptual presence and its absence in synesthesia. *Cognitive Neuroscience* 5 (2): 97–118.

Smithies, D. 2016. Belief and self-knowledge: Lessons from Moore's paradox. *Philosophical Issues* 26 (1): 393–421.

Tribus, M. 1961. *Thermodynamics and thermostatics: An introduction to energy, information and states of matter, with engineering application*. New York: D. Van Nostrand.

van Schalkwyk, G. I., Volkmar, F. R. & Corlett, P. R. 2017. A predictive coding account of psychotic symptoms in autism Spectrum disorder. *Journal of Autism and Developmental Disorders*.

Velleman, J.D. 1992. The guise of the good. *Noûs* 26 (1): 3–26.

von Helmholtz, H. 1860 (1962). *Handbuch der physiologischen Optik*. Transl. & ed. dover: J. P. C. Southall.

Weilnhammer, V., et al. 2017. A predictive coding account of Bistable perception - A model-based fMRI study. *PLoS Computational Biology* 13 (5): e1005536.

Wiese, W. 2017. What are the contents of representations. *Phenomenology and the Cognitive Sciences* 16 (4): 715–736.

Wilson, T. 2002. *Strangers to ourselves: Discovering the adaptive unconscious*. Cambridge: Harvard University Press.

Wright, C. 1998. Self-knowledge: The Wittgensteinian legacy. In *Knowing our Minds*, 13–45. Oxford: Oxford University Press.