Underlying delusion. Predictive processing, looping effects, and the personal/sub-personal distinction

Matteo Colombo & Regina Fabry

Abstract What is the relationship between the predictive processing theory of brain functioning and the everyday concepts with which people explain and conduct their mental lives? To answer this question, we focus on predictive processing explanations of mental disorders that appeal to false inference. We distinguish two concepts of false inference. Next, we survey various ways of understanding the relationship between personal and sub-personal level explanations. Finally, we argue that if predictive processing actually governs psychological and socio-cultural dynamics, then personal and sub-personal concepts must display a looping, co-constitutive relationship. In particular, explaining delusion in terms of the sub-personal concept of false inference has a constitutive impact on the nature of delusional experiences: It would be a constitutive feature of delusion that sub-personal systems tend to draw false inferences about the external world.

Keywords: Personal/Sub-personal distinction; Predictive Processing; Delusion; Looping Effects

1 Introduction

It has been claimed that the predictive processing theory of brain functioning (PP) fails to specify how sub-personal concepts referring to brain states and activities relate to personal-level concepts referring to conscious mental states and activities (Holton, 2016). While PP appeals to belief, prediction and inference to explain brain functioning, these concepts differ radically from the everyday notions of belief, prediction and inference, which people employ to make sense of their mental life and behaviour. If the concepts in PP explanations are aptly employed only when they refer to sub-personal states and activities, then it is hard to see how PP can illuminate our understanding of people as rational agents and subjects of experience (on the personal/sub-personal distinction, see Colombo, 2013; Dennett, 1969; Drayson, 2014).

Clark (2016a) clarifies the “various potential conflicts between what I, the conscious agent, might explicitly predict and the underlying whirl of non-conscious (sub-personal) probabilistic processing [are in fact] golden opportunities.” This is because PP “may one day deliver a better understanding even of our own agent-level experience than that afforded by the basic framework of ‘folk psychology’” (Clark, 2016b, 82). Discussing PP accounts of mental disorders, Clark emphasises these accounts “begin to suggest, in broad outline, ways to connect our neurophysiological understandings, via computational and ‘systems-level’ theorizing, to the shape and nature of human experience” (Ibid, 237). Clark, unfortunately, does not specify the nature of this connection.
In this paper, we fill this gap by examining how PP explanations of delusion relate to personal-level explanations of delusion. After surveying the commitments of PP and its explanation of delusion in terms of *false inference* (Section 2), we distinguish two ways of understanding *false inference* and ask the question: How does the concept of *false inference* in PP relate to the concept of *inference* employed in personal-level explanations of delusion? Put differently, how can PP discourse, and in particular its characterisation of delusion in terms of *false inference*, bear on the nature of delusion itself and on the way delusional psychopathology is experienced, both by patients and other people? (Section 3). We consider four answers: first, PP and personal-level explanations are mutually *autonomous*; second, PP explanations *enable* personal-level explanations; third, personal-level explanations *reduce* to PP explanations; fourth, PP and personal-level explanations are *co-constitutive* with respect to at least some phenomena (Section 4). We argue that if PP actually governs conceptual change and cultural dynamics, then personal and sub-personal explanatory concepts of mental disorders must display a looping, co-constitutive relationship. Explaining delusion in terms of the sub-personal concept of *false inference* should have a constitutive impact on the personal-level idea of *false, tenacious, strange belief based on incorrect inference*, which is used to classify delusional psychopathologies and to make sense of delusional experience (Section 5). In a conclusion, we summarize our contribution and suggest avenues for future research.

**2 PP explanations of mental disorder**

While PP is a big tent, a broad consensus has emerged about its core ideas (Clark, 2013; Friston, 2010; Hohwy, 2013; Seth, 2015; Wiese & Metzinger, 2017). The central commitment is that brains are kinds of probabilistic prediction mechanisms. Their function is to generate predictions about the likely causes of the signals they receive from the environment, to compute prediction errors, to update their predictions on the basis of these errors, and to use their predictions for the control of action. While brains aim at minimizing their prediction error, mental abilities like perception and action would emerge as ways organisms use to regulate and stabilize their internal milieu.

PP has two explanatory goals: first, to identify a small set of kinds of computational processes that are implemented by many different types of neural and non-neural structures at different spatial and temporal scales. Second, to provide functional explanations of most (if not all) mental phenomena, including mental disorders, in terms of those computational processes.

The main explanatory concepts in PP are three: *prediction, prediction error*, and *precision*. Because PP draws on the theoretical resources of information theory, statistics, and machine learning, these concepts refer to quantities defined over statistical models of the causal structure of the environment. These models are ‘generative’, because they specify how observed data might have been generated from some underlying probability distribution. Given an observable variable $X$ (for
example, sensory signals) and a target variable $Y$ (for example, the environmental causes of sensory signals), a generative model is a statistical model of the joint probability distribution $\text{Prob}(X, Y)$.

According to PP, brains implement generative models, which are poised to make predictions about sensory signals, where ‘prediction’ refers to the weighted mean of a random variable. Prior predictions made on the basis of generative models would be transmitted downwards, through the cortical and subcortical hierarchy in the brain. Each level in the hierarchy carries sensory signals for the level above, and prior predictions for the level below.

‘Prediction error’ refers to the magnitude of the discrepancies between predictions about the value of a certain variable and its actually observed value. Prediction errors would be transmitted up the cortical and subcortical hierarchy, for updating the probabilistic models of the causal structure of the environment, and ultimately for controlling action.

Each prediction error is weighted by an expectation regarding the precision of a sensory signal and of the prediction of that signal. The term ‘precision’ refers to the inverse variance of a random variable (i.e., a measure of how far each value of, for example, a type of sensory signal is from the mean value of that type of signal). The impact of a prediction error on model updating and action control is smaller, the more precise the prediction; and larger, the more precise the prediction error. If estimates of the precision of sensory signals compared with prior predictions are incorrect, this will result in the allocation of an insufficient or excessive weight to prediction errors. Allocating insufficient or excessive weight to prediction errors means brains would give too much weight to their sensory signals compared with prior predictions in updating the models they implement.

According to PP, prediction error is minimized in two complementary ways: courtesy of perceptual and active inference. While perceptual inference aims at minimizing exteroceptive prediction errors (i.e., prediction errors relating to sensory signals causally produced by states external to an organism), active inference aims at minimizing proprioceptive and interoceptive prediction errors (i.e., prediction errors relating to sensory signals causally produced within an organism, respectively in the gut and other internal organs, and in association with the position and movement of the body).

These two modes of inference in the brain consist of non-conscious, sub-personal operations defined over rules (i.e., input-output mappings) and computational vehicles (e.g., neural activations). Yet, several philosophers believe the greatest epistemic value of PP explanations is that they bring the scientific image of sub-personal neural computation, and the manifest image of people’s lived experience closer together. PP approaches to mental disorders would make this payoff salient. Hohwy (2013, 157), for example, claims “[i]f we assume that mental disorder is characterized at least partly by more or less systematic misperception, then we could perhaps learn more about such disorder by considering prediction error minimization.” Clark (2016b, 204) is more explicit, and
suggests we need “to begin to recognize ourselves in the swirl of ongoing, multilevel prediction,... [which] may determine the shape and nature of both normal and atypical forms of human experience.”

The basic idea of PP explanations of mental disorders is that they consist in false inferences performed by the brain (e.g., Fletcher & Frith, 2009; Friston, 2005). False inferences depend on aberrant prior predictions, aberrant prediction errors, or aberrant precision weighting. PP advocates explicate false inference in terms of aberrant estimates of the precision of prediction errors relative to top-down predictions. Inappropriately precise prior predictions may bring about perceptions in the absence of external causes of those perceptions (hallucinations); and inappropriately precise prediction error may bring about beliefs that are rigidly irresponsible to sensory evidence (delusions) (Adams et al., 2013; Clark, 2016b, Ch. 7; Corlett, Frith, & Fletcher, 2009; Deneve & Jardri, 2016; Hohwy, 2013, Ch. 7). To appreciate how PP would bring the scientific image of sub-personal neural computation, and the manifest image of people’s lived experience closer together, let’s use delusion as a case study.

2.1 PP models of delusion

In Allgemeine Psychopathologie, Karl Jaspers singles out three features of delusion: 1. subjective certainty incomparable to other convictions; 2. imperviousness to counterarguments; 3. implausibility of content (Jaspers, 1913/1963, 95-6). Consistently with Jaspers’ criteria, the Diagnostic and Statistical Manual of Mental Disorders defines delusion as a “false belief based on incorrect inference about external reality that is firmly sustained despite what almost everyone else believes and despite what constitutes incontrovertible and obvious proof or evidence to the contrary. The belief is not one ordinarily accepted by other members of the person’s culture or subculture (e.g. it is not an article of religious faith)” (DSM-5, 2013, 819).

Jaspers emphasised the formation of delusion is often associated with personal experiences of something radically new or alien: “Patients feel uncanny... that there is something suspicious afoot. Everything gets a new meaning. The environment is somehow different... this general delusional atmosphere with all its vagueness of content must be unbearable. Patients obviously suffer terribly under it and to reach some definite idea at long last is like being relieved from some enormous burden” (Jaspers, 1913/1963, 98). This description coheres with the characterisation the DSM-5 offers of delusional disorder, which is said to typically arise in the context of personal distress. Its formation is often associated with highly strange perceptual experiences, or with anxious efforts at explaining certain affectively-charged life events (DSM-5, 2013, 90-93).

Both Jaspers’ observations and the DSM definition suggest delusions are species of belief, namely: a personal-level mental state that depicts the external reality as being in a certain way, that
is evaluable in terms of its truth or falsity, that is formed and revised courtesy of personal-level inferences on the basis of evidence, and that can guide intentional action. To the extent delusions are obviously false, systematically insensitive to the evidence, and insulated from explanatory, logical and evidential standards shared by members of an epistemic community, they would represent a “paradigm case of irrationality” (Gold & Hohwy, 2000, 147) or of “impaired epistemic rationality” (Bermúdez, 2001, 470). Their irrationality would depend not on their content, on what they are about, but on their intra-personal and inter-personal dynamics and relationships with other mental states. Although this doxastic understanding of delusion is not without critics (e.g., Coltheart, 2007; Gallagher, 2009), this characterisation is the dominant one among psychiatrists, and among laypeople in Europe and North America, at least since the 18th century (Berrios, 1996; Murphy, 2013).

PP explains delusions in terms of an aberrant balancing of the relative weight of prior predictions and prediction errors in computational processes of perceptual and active inference (Adams et al., 2013; Bortolotti & Miyazono, 2015; Corlett et al., 2010; Corlett & Fletcher, 2015). On a PP understanding, the formation and maintenance of delusion are the result of an aberration in precision weighting, which brings about a circular process “allowing false perceptions and bizarre beliefs to solidify into a coherent and mutually supportive cycle” (Clark 2016b, 80; see Jardri et al. 2017 for preliminary evidence). The formation of delusions would be due to inappropriate estimates of the precision of sensory signals. As Corlett et al. (2010, 347) explain, “during the earliest phases of delusion formation aberrant novelty, salience or prediction error signals drive attention toward redundant or irrelevant environmental cues, the world seems to have changed, it feels strange and sinister, such signals and experiences provide an impetus for new learning which updates the world model inappropriately, manifest as a delusion.”

Over-weighted sensory signals can become over-salient, which generates inappropriately precise prediction errors. If these prediction errors are transmitted up the hierarchy, they cannot readily be “explained away” at higher levels of the generative model. This means the model may appeal to no plausible prediction about the external causes of those over-weighted sensory signals, which may fit the resulting inappropriately precise prediction errors. One solution is to explain away those prediction errors, and make sense of lower level sensory signals, by having the model generate novel (false) predictions that fit the over-weighted prediction errors. These false predictions would become progressively inflexible, and insensitive to novel sensory evidence, thereby giving rise to a delusional belief.

For example, if you do not feel the usual affective experience when you face your beloved—though you can judge that the face in front of you is identical to that of your beloved—you might have an unexpected sensory experience, which might be associated with a prediction error in your brain. This prediction error would signal that your normal affective response to somebody looking
like your beloved is unexpectedly lacking. If this prediction error is given too much weight, then your unexpected sensory experience will be flagged as especially significant. If the probabilistic model implemented by your brain is updated on the basis of this overly precise prediction error, the model will issue new predictions to the effect that this person in front of you, who looks like your beloved, feels in fact like a stranger to you. The result of perceptual and active inferences aiming at minimizing your prediction errors is a false inference such as ‘That person is not really my beloved!’ This false inference would provide you with an apparently good explanation of your abnormal sensory experience; and if this false inference acquires too much precision relative to future prediction errors, it will lead to the formation of a delusion.

The maintenance of a delusion is explained in terms of the suppression of sensory signals that would weaken the rigidity of the deluded belief. In terms of PP, “top-down priors are weighted more in inference; in addition, the dampened down error signal is a less efficient supervisor in the ongoing revision of prior belief, making it more likely those priors stray from reality” (Hohwy, 2013, 159). The maintenance of the delusion is thus explained by an increase of the precision assigned to top-down predictions and a decrease of precision assigned to prediction errors.

Two points are worth making in this context. First, PP advocates apply the same explanatory strategy to different species of delusion, as well as to other positive symptoms of schizophrenia (Corlett et al., 2009). Second, to explain delusions, PP advocates straddle the distinction between personal and sub-personal levels of explanation, using a variety of concepts from folk psychology, psychiatry, neuroscience, statistics and machine learning. Before we unravel the relationship between these concepts, let’s clarify the first point.

Different delusions have remarkably different features at the personal level of explanation: they differ in content, degree of conviction, phenomenology, in the extent to which they are associated with negative affect, and in their motivational power (Appelbaum, Robbins, & Roth, 1999; Bortolotti, 2009; Davies & Coltheart, 2000; Gallagher, 2009; Garety et al., 2005). Some delusions have bizarre content—like the delusion that aliens took control of one’s own actions. Other delusions have more plausible content—like the delusion that one is being followed by malevolent people. Some delusions are polythematic, while others are monothematic—like Capgras delusion, where one believes that a close relative or friend has been replaced by an identical-looking impostor. While some schizophrenic patients with delusional disorder present the “jumping to conclusion” bias in some probabilistic reasoning tasks (Huq, Garety, & Hemsley, 1988), and tend to reject other people’s opinions that counter their own delusions (McCabe, 2004), they do revise their non-delusional beliefs taking account of other people’s opinions and available evidence (Kaliuzhna et al, 2012).

In the face of this diversity, the basic type of explanation offered within PP remains the same. Consider, for example, delusion of control, which consists in “an abnormality in the feeling of being in
control of one’s action” (Frith 2005, 763). Within PP, delusions of control are explained in terms of a deficit in sensory attenuation (Limanowski, 2017). Usually, “sensory attenuation is a necessary consequence of reducing the precision of sensory evidence during movement to allow the expression of proprioceptive predictions that incite movement” (Brown et al., 2013, 413). Sensory attenuation would guide proprioceptive active inference by assigning low levels of precision to sensory prediction errors at the time of movement execution. In delusions of control, sensory prediction error signals would not be attenuated, but would be estimated as being highly precise. As sensory prediction errors are highly precise, they would bring about a drastic update of the neurally implemented probabilistic models, which would result in “falsely inferring an opposing exogenous force” (Brown et al., 2013, 423). Thus, the personal-level delusional belief that actions are controlled by another agent would be formed as a causal consequence of a deficit in sensory attenuation, which PP advocates explain in terms of inappropriately precise sensory prediction errors.

The second point we made above is that PP explanations seem to straddle personal and sub-personal levels of explanation. This is apparent in the case of delusions of control. For example, in an attempt to make contact with personal-level explanations, Brown et al. (2013) clarify the falsehood of inferences is not determined from within the sub-personal predictive system, but only from a scientific third person point of view: “from the point of view of the subject–its inferences are Bayes optimal. It is only our attribution of the inference as false that gives it an illusory or delusional aspect” (Brown et al., 2013, 423). Likewise, Clark (2016b) suggests delusions “appear bizarre and unfounded to the external observer, yet from within now constitute the best, because the only, explanation [of sensory data] available” (206-207).

These types of remarks highlight the normative qualification ‘inappropriate’ in inappropriately precise prediction error makes sense only against the background of computational psychiatrists’ expectations about what counts as appropriate information processing (Colombo forthcoming, on miscomputation in computational psychiatry). These remarks also motivate the question: How does the sub-personal, PP explanation of delusion relate to personal-level explanations appealing to considerations of epistemic rationality? In particular, how does the notion of false inference in PP explanations of delusion relate to personal-level notions of rational inference? In Hohwy’s words, “the question is how we should understand this idea [of delusion as false inference] without succumbing to crass neuroanthropomorphism” (Hohwy, 2013, 23).

3 Inference. Personal and sub-personal

An inference can generally be understood as a rule-governed transition from one body of information to another. While this type of transition consists in an activity (i.e., the activity of
inferring), the term ‘inference’ can also refer to the outcome of inferring, to the body of information the transition issues. We can specify what inference is, in two ways at least.

First way: an inference is a form of “personal-level, conscious and voluntary” reasoning, which requires a personal-level ability to follow certain rules (Boghossian 2014, 3; see also Frege 1979; Harman 1986; Neta 2013).

Consider this example. You go to the market, where you know you can find fresh strawberries. You draw a random strawberry from a bag. The strawberry you have drawn looks good. You conclude all strawberries in that bag are good. Reaching this conclusion involves a transition from the thought that this strawberry from this bag is good to the thought that all strawberries in the bag are good. The transition is conscious, as you are aware of the observations you make and the conclusion you reach on the basis of your observation. The transition is voluntary, as you willingly take the observations you made to provide good evidence for your conclusion; but you could have refrained from drawing that conclusion, had your observation been different. The transition also requires you be able to follow rules you accept, for example the rule ‘If this randomly drawn strawberry is good, then all strawberries from the same bag will probably be good.’ Applying this rule to your initial thought, you infer a new thought that can be evaluated for its truth and evidential support. Likewise, you and other people can evaluate the rule you applied in terms of its truth-conduciveness, coherence, or other epistemic properties.

Here is the second way to cash out what inference is: inference is a rule-governed transition from one content-bearing state to another, which may be carried out sub-personally in biological or artificial systems. This transition need not involve any conscious state, voluntary action or ingenuity on the part of the system, though it requires the system be able to token semantically evaluable states, including, perhaps, inference rules themselves.

Consider this other example. AlphaGo is the first computer program to defeat a professional human Go player (Silver et al., 2016). This system combines a computational architecture and algorithms associated with deep neural networks and tree search. AlphaGo’s deep neural network architecture allows it to construct increasingly abstract representations of the current state of a game of Go, which are then used to search a subset of the possible moves ahead in the game, and to evaluate the chance of victory for each one of these moves. AlphaGo uses reinforcement learning algorithms to bootstrap itself towards better and better performance, incrementally learning from its mistakes. Thus, given a board configuration, AlphaGo makes a search ahead, evaluates many possible moves, and selects one move, so as to improve its chance of winning the game. In carrying out these operations, AlphaGo transitions from one content-bearing state to another, in a rule-governed manner. But AlphaGo’s following rules with respect to the content of its states is a matter of automatized, non-conscious processes, where AlphaGo cannot obviously vet the rules it follows.
Nonetheless, there is a sense in which AlphaGo can make inferences. It is the same sense assumed by some approaches in current perceptual psychology, which hypothesize that perceptual sub-personal systems carry out unconscious statistical inferences (Rescorla, 2015).

In a statistical context, inference consists in drawing conclusions from sample data about the characteristics of a target process that is assumed to have generated the data.\(^1\) Statistical inference is something that both people and sub-personal systems like the algorithms in AlphaGo can perform. The difference is that statisticians can figure out why they are moving from certain thoughts to others. They are aware, or can become aware, of the different steps involved in the inferences they perform, and can evaluate both the evidential support of their conclusions, and the aptness of the statistical rules they follow. By contrast, sub-personal systems need not assign any interpretation or meaning to the states involved in their inferences. Their rule-following need not involve any comprehension of why they are moving from some state to another.

In summary, there are at least two ways of understanding inference. One understanding lies at the personal level of explanation, the other can lie at the sub-personal level. In the light of this distinction, to say that an inference is false is to say that a rule-governed transition—either from some personal-level thoughts to others, or from some non-conscious content-bearing states to others—delivers some inaccurate representation. It is to say, for example, that your conclusion is false that all strawberries in that bag are good; or that AlphaGo’s estimate is false that it is currently most likely to win the game.\(^2\)

With this distinction in place, we can now return to our question: what’s the relationship between explanations of delusion at the personal and sub-personal level that refer to false inference? Let’s consider four possible answers.

### 4 Relating personal and sub-personal levels of explanation

The first answer is that personal and sub-personal explanations are mutually autonomous. This answer denies the explanatory posits of PP bear any relationship to the posits of personal-level explanations. One reason for autonomy is that personal-level explanations should illuminate the phenomenology or rationality of mental phenomena. But explanations that refer to sub-personal transactions unavailable to conscious scrutiny cannot make delusional belief intelligible and cannot illuminate the structure of the phenomenal experiences of individuals suffering from delusions (Bennett & Hacker, 2003; Gerrans, 2014, Sec. 1.1-1.3). Likewise, personal-level explanations, which

---

\(^1\) Unlike in statistics, inference is typically distinguished from learning in a Machine Learning context. Learning is typically associated with the task of estimating the statistics of some hidden variable or the parameters of a model, while inference is associated with the task of predicting the value of a target variable.

\(^2\) Intriguingly, AlphaGo suffered from problems which David Silver, the lead researcher on AlphaGo, called ‘delusions’, that is: persistent, long-term mis-evaluations of specific board positions.
posit folk-psychological concepts displaying rational connections, would be irrelevant to explanations that are aimed at uncovering neurocomputational mechanisms, such as PP explanations.\(^3\) The folk-psychological notions of *belief* and *justified inference*, or phenomenological analyses of people’s subjective experiences would contribute nothing to the projects of identifying common kinds of neurocomputational processes and of providing functional explanations of mental phenomena in terms of those processes. If explanations of delusion at the personal and sub-personal level are mutually autonomous, then they do not constrain one another, and learning about one explanation provides us with no additional knowledge about the explanation at the other level.

A second answer is that explanations at the sub-personal and personal level stand in an *enabling* relationship (Hurley, 2008; McDowell, 1994). According to this view, “subpersonal informational and causal theories explain how personal-level phenomena become possible—are enabled—but need not share structure with personal-level descriptions of processes as rational or conscious” (Hurley, 2008, 3). A set of conditions enable a phenomenon just in case they are causally necessary to the occurrence of that phenomenon. The enabling relationship allows PP explanations to relate causally to personal explanations of delusional belief; they would provide us with information about how certain neurocomputational mechanisms work that enable mental states of a delusional type. In so doing, PP sub-personal explanations allow us to answer counterfactual questions about how mental states of a delusional type are causally affected by changes in sub-personal factors, for example by changes in precision estimates of sensory signals. Answering this type of counterfactual questions can help us refine our causal understanding of the factors that give rise to the formation and maintenance of delusional beliefs, which is crucial for finding effective treatments for psychiatric disorders, but does not illuminate what delusion consists in.

A third answer concerning the relationship between personal and sub-personal explanations is *reduction*. This is classically understood as a logical relationship, where explanations or concepts from one domain can be logically deduced from the explanations or concepts from another domain (Nagel, 1961). Thus, a personal-level explanation of delusion reduces to the PP explanation just in case the personal-level explanation is derivable from the PP explanation with the help of bridge principles connecting the concepts of the two explanations. In particular, the relationship between the concept *false inference* in PP—viz., an informational transition involving aberrant precision weighting—and the personal-level concept *false inference*, which refers to an incorrect conscious inference about reality, would be specified by synthetic, extensionally equivalent statements. While

---

\(^3\) Explanation of people’s behaviour couched in the folk-psychological vocabularies, which are used by ordinary people to identify and track mental states and to explain and predict behaviour, is the paradigm case of personal-level explanation. But personal-level explanations include also explanations of people’s behaviour and mental states couched in scientific vocabularies. As should be clear from Section 2.1, some scientific accounts of mental disorders, such as the DSM-5, offer personal-level explanations couched in both folk-psychological and scientific vocabulary.
in previous work we have argued any specific PP model does not have the status of a reducing explanation (Colombo & Wright, 2017; Fabry, 2017), cashing out the relationship between PP and personal-level explanations of delusion in terms of reduction would mean current folk-psychological understanding of belief and inference should at the very least be revised under the pressure of the empirical and theoretical successes of PP explanations.

The fourth answer is co-constitution. Where two explanations or concepts of a target phenomenon stand in a co-constitutive relationship with respect to that phenomenon, they each provide us with information that makes intelligible what the phenomenon consists in, possibly affecting the nature of the phenomenon itself. To unpack this idea, it is helpful to consider how McDowell (1994) construes explanation that is constitutively relevant to a phenomenon, and then to extend McDowell’s idea to define the way we understand co-constitution here.

Engaging with Lettvin et al.’s (1959) neurocomputational explanation of frogs’ sensory systems, McDowell acknowledges that a full causal understanding of frogs and their behaviour requires both froggy and sub-froggy explanations, since the latter would causally enable the former. But, he argues, “[t]he sub-personal account of a sensory system which treats it as an information-processing device that transmits its informational results to something else inside an animal cannot adequately characterize what its sensory systems are for the animal” (Ibid., 197). If we want to adequately characterize what its sensory systems are to the frog, then we must place the whole animal in the context of its competently inhabiting its environment. The idea is that an explanation is constitutively relevant to a phenomenon just in case, in virtue of the concepts it employs, the explanation makes intelligible what the phenomenon consists in.

If two explanations of a given phenomenon stand in a co-constitutive relationship with respect to that phenomenon, then, in virtue of their concepts, they both contribute to make intelligible what the phenomenon consists in, possibly affecting the very nature of that phenomenon. This does not mean the two explanations constitute one another, but that each is partially constitutive of a target phenomenon. Co-constitution entails that, over time, explanations and concepts at one level are susceptible to revision (and sometimes elimination), given discoveries and conceptual change at other levels of explanation (Churchland, 1993; Colombo, 2013). It also entails the nature of a target phenomenon can change as a function of conceptual changes in some of its explanations. Given the four options we have sketched, what’s the best for understanding the relationship between personal-level explanation and sub-personal PP explanation of delusion?

5 Co-constitution and PP
We assume—in broad agreement with Anscombe’s (1963), Davidson’s (1963), and others’ contributions on “acting under a description”—that the understanding of the nature of mental
phenomena and behaviour displayed by people depends on the linguistically embodied concepts people employ to describe those phenomena. For example, we may say that, when confronted with counterevidence, your resistance to revising your belief that your partner is an impostor is epistemically irrational, where your inferences about the identity of your partner are systematically and tenaciously false. Here we use expressions like ‘epistemically irrational’ and ‘systematically false inference’ to characterise the nature of a certain type of mental state.

The descriptions available to members of an epistemic community to make sense of mental phenomena include both personal and sub-personal concepts, which psychiatrists, neuroscientists, biologists, computer scientists, philosophers, the media, and so forth, inject into public discourse. As novel descriptions become available, members of the community may re-conceptualise the same type of behaviour. For example, we may say your resistance to revising your belief that your partner is an impostor is a symptom of ‘Bayes optimal perceptual inference,’ where your cognitive system relies more on neurally encoded predictions relative to prediction errors.

If people in a community use these sub-personal descriptions to make sense of what delusion is, then, as a matter of fact, personal and PP sub-personal explanations are not autonomous. PP explanations have in fact been offered to illuminate the phenomenology of delusion, as they would lend support to the idea that there is no sharp distinction in the phenomenological structure of delusion, hallucination, and belief (e.g., Mishara & Sterzer, 2015). Likewise, personal-level concepts displaying rational connections with one another would be relevant to make sense of how sub-personal processes can be ‘Bayes optimal’ or ‘inappropriately precise.’ If these observations are correct, then autonomy does not adequately characterise the relationship between personal explanation and PP sub-personal explanation.

Reduction is not adequate either, for several reasons (Colombo & Wright, 2017; Fabry, 2017). Here is one: the connectability of the concepts of a reducing explanation to the concepts of a reduced explanation is a necessary condition for reduction. This means that personal-level explanations of delusion must be wholly translatable into PP vocabulary. If the reducing PP explanation must explain the set of phenomena that the personal-level explanation to be reduced covers, then the reducing PP explanation must be more inclusive than the reduced personal-level explanation. The PP explanation must contain the concepts contained in the personal-level explanation.

It’s important to clarify three points. First, our focus is broader than Anscombe’s and Davidson’s, who were concerned with explanation of intentional actions in terms of the agent’s reasons for acting. Second, in explaining a given phenomenon, one can use different descriptions. The availability of multiple descriptions does not entail there are multiple phenomena. Third, we are concerned with explanation under a description, where explanations are answers to why or how questions, and an explanation of a phenomenon under different descriptions can support different counterfactuals.
However, the explanation of delusion offered by PP is not more inclusive than personal-level explanations of delusion. As we noted above, the vocabulary employed by psychiatrists and laypeople to make sense of delusion at the personal-level includes normative concepts highlighting epistemically irrational aspects of delusion. These aspects are not covered by PP explanations. In fact, where personal-level explanations posit delusions as *epistemically irrational*, PP explanation posits delusions as *Bayes optimal*. This indicates personal-level explanations of delusion are not wholly translatable into PP vocabulary. For this reason, the connectability requirement is not satisfied, and personal-level explanations of delusion cannot be reduced to PP explanations.

If PP is true, then the sub-personal account of delusion it delivers explains how delusions are *causally enabled*—it provides us with knowledge of the causal factors that are necessary for the appearance and maintenance of delusion, of the causal mechanisms underlying delusion. After all, according to PP, the orchestrated functioning of perceptual and active inference would be the underlying mechanism of much, if not all, mental phenomena. The symptoms of mental disorders, including delusion, would thus causally depend on malfunctioning in the delicate balancing of precision estimates, prediction, and prediction error in a specific environment. So, at least, PP explanations and personal explanations can stand in an enabling relationship.

They can also stand in a *co-constitutive relationship*, since they both contribute to illuminate what it is to be deluded. In particular, if the descriptions under which people behave include both personal and sub-personal concepts, then both types of descriptions will contribute to our understanding of what kind of behaviour or mental phenomenon a certain person displays. PP is committed to the idea that brains, engage in statistical analyses aimed at inferring the causes of incoming sensory signals (Colombo, 2017, Sec. 5). If the descriptions used to understand people’s mental phenomena are among the sensory signals brains use to update their predictions, then descriptions like ‘irrational belief’ or ‘inappropriately precise prediction error’ may have an impact both on people’s self-understanding, and on mutual, communal understanding of what it is to be deluded.

The co-constitutive relationship between descriptions at the personal and sub-personal level has implications also for understanding historical and cross-cultural variation in the concept of delusion. As the descriptions available to characterise certain kinds of people, behaviours, or mental phenomena change, the communal understanding of the nature of particular kinds of people, behaviour, or species of belief, may also be subject to change. The availability of descriptions like ‘superstitious belief,’ ‘degenerate belief,’ ‘biased reasoning,’ ‘inappropriately precise prediction error’ will differentially affect statistical models of reality implemented in human brains at different historical times and in different epistemic communities. In this way, the concept *delusion* is displayed
as a “moving target”: changing descriptions of delusion in a social environment may alter the nature of delusion itself in that environment.

To clarify the idea that sub-personal PP explanations and personal explanations stand in a co-constitutive relationship governed by predictive processes, Ian Hacking’s (1995a) notion of looping effects of human kinds is helpful. According to Hacking, the continuous classification and systematization of knowledge about humans, which come most prominently from folk conceptions, scientific achievements, and bureaucratic and legal taxonomies, can create “new ways to be a person” (Hacking, 1995b, 239). The idea is that social reality is structured, and in part constituted (at least in some cases), by the descriptions people apply to the behaviours and mental phenomena they observe. Such descriptions can bring to light new human kinds. Child abuse, homosexuality, and multiple personality disorder are examples Hacking discusses of human kinds brought to light by certain ways of describing certain people and their behaviours. When a description, a label, becomes widespread in a community, people may come to recognize themselves and others in those new concepts and labels. And this recognition may impact the self-understanding of those people, and the social responses to their behaviour and beliefs, which in turn impacts usage of that description.

Scientific concepts and explanations—for example, those offered by PP—might impact people’s conceptions of mind and behaviour. These changes may lead to looping effects, where classifications, descriptions, and explanations of human kinds lead to “changes in individuals of that kind, which means that the kind itself becomes different... Next, because the kind changes, there is new knowledge to be had about the kind. But that new knowledge in turn becomes part of what is to be known about members of the kind, who change again” (Hacking, 1995a, 369-70). These looping effects may determine new social roles, which in turn set new normative expectations about appropriate or justified behaviour (or belief) in a community. For example, labels like ‘deluded belief,’ ‘pervert,’ and ‘possessed,’ carry normative significance, both in terms of setting relevant norms for epistemic or moral valuation, and in terms of institutional and communal responses to the behaviour of people described in those ways.

These looping effects reveal that, at least sometimes, some linguistically embodied concepts influence the nature of mental disorders, as well as how they are framed, experienced, and treated. If looping effects determine, at least in part, normative expectations about certain kinds of people, then not every person who makes false inferences, or whose brain assigns too much precision to incoming sensory signals, may aptly be called ‘delusional.’ The aptness of the labelling depends on a complex web of communal and institutional expectations about belief formation and belief updating. These expectations include descriptive expectations about typical belief formation and updating in a community, and normative expectations about conditions under which a belief violates shared, communal epistemic norms. That’s why Clark (2016b) suggests delusions “appear bizarre and
unfounded to the external observer” (206). Delusions violate expectations of certain external observers, although PP advocates explain them as “optimal (but false) inferences.” If PP governs looping dynamics involving human kinds, then the aptness of a description in a community will ultimately depend on the power of that description to allow members of the community to interact smoothly, and to stabilize the milieu of the community.

Now, to better appreciate how PP-governed looping effects might influence how delusion is framed, experienced, and treated, we should take into account that looping effects can come about via a number of different avenues in the course of individuals’ and communities’ cognitive trajectories, which are embedded in a rich, socio-culturally structured environment. At an individual scale, and over a relatively short period early in development, perceptual and active inferences would pick up non-linguistic behavioural regularities that track the descriptive and normative expectations of members of their communities about (ir)rational belief. While previous, cumulative structuring of the epistemic environment of the baby channels the baby’s data gathering process, it also determines the estimated precision of sensory signals relative to model predictions in the baby’s cognitive system.

At longer time scales, this continuous cycle of perceptual and active inference is scaffolded and channelled by linguistic structures, labels, and narratives, which make available new domains of thinking, contributing to the spread, policing, and negotiation of epistemic norms across times and communities (Dennett, 2000; Sterelny, 2012). Language allows PP processes to use compressed, re-coded representations of complex, abstract patterns in the statistics of the environment (Colombo, 2013b). In Clark’s (2005, 265) words, “language is the key cognitive tool by means of which we are able to objectify, reflect upon, and hence knowingly engage with, our own capacities of thought, reason and self-control”. In this sense, language would contribute to the looping effects brought about by PP processes taking place in structured social environments. Narratives would be a particularly effective means for channelling socially distributed processes of prediction error minimization aimed at social learning and regulation. This is because narratives would effectively contribute to determine the precision with which people categorize, describe, and understand themselves and others morally and epistemically as a function of available linguistic labels and previous sensory experiences and social feedback (Fabry, 2018; Hutto & Kirchhoff, 2015; Menary, 2008).

To make this point concrete, the expanding, linguistically embodied, concept of Bayes optimal active inference in delusion might lead people diagnosed with delusion disorder to experience themselves as less irrational than they would have been at earlier times. To the extent PP displays belief as enslaved by action, and to the extent PP concepts enter looping effects at longer time-scales, delusion may be reconceived as a disorder of both practical and epistemic rationality.
While existing research shows that individuals prone to delusions generally request less information to make a decision in some reasoning tasks, they also display more indecisiveness when facing real-life dilemmas and greater effort to monitor the consequences of one’s own actions (McKay, Langdon, & Coltheart, 2006; Sterzer et al., 2018).

If people get to understand the nature of mental disorders, and of delusion in particular, as computational and fundamentally bound up with action control, then looping effects might lead to changes in the public attitudes towards patients diagnosed with delusional disorder. While the public may blame patients less, they may also perceive them as recklessly erratic, since the computational routines running on their brains might be seen as biased, opaque and unpredictable. Interestingly, Schomerus et al. (2012) provide evidence that between 1990 and 2006 the proportion of people who conceive of schizophrenia as a genetic and brain disease has increased in a sample of European and North American countries. While this conception normalizes mental disorders as diseases like any other, it makes patients more fatalistic about their capacity to control the course of their malady.

Distributed over multiple, institutionally organized, embodied predictive systems at longer time scales, the same type of process of prediction, prediction error, and model updating would allow successive generations to test, re-shape and transmit existing normatively-laden folk-psychological and scientific labels and narratives via cultural learning (Henrich, 2016; Heyes, 2016). At this spatially larger and temporally longer scale, institutions like universities, churches, families, newspapers, social media, and so forth, would afford highly precise predictions about patterns of behaviour corresponding to certain human kinds. Pressures from below, from individuals displaying or failing to display those patterns, may trigger prediction errors leading to updating previous socially distributed and institutionally regulated predictions.

If the speculative suggestions we have offered are on the right track, then PP, during cultural learning and socio-culturally sculpted cognitive history, would tap linguistic labels and narratives, re-shaping people’s understanding of the personal-level concepts, within which they conduct and evaluate their life. As Roepstorff’s (2013, 225) surmises, linguistic and narrative practices “may help establish priors or even hyperpriors, sets of expectations that shape perception and guide action.” In this sense, human kinds would act as (hyper-)priors that are the “result of shared expectations that are communicated in interactions, mediated by representations, solidified through materiality, and extended into an action space” (Ibid). Likewise, following Hirsh et al. (2013), Clark hypothesises that “narratives function as high-level elements in the models that structure our own self-predictions, and thus inform our own future actions and choices” (Clark, 2016b, 286). Narratives featuring human kinds would “tend to feed the structures and expectations of society back in, so that they become reflected in the models that an individual uses to make sense of her own acts and choices. Personal
narratives may thus be considered as another species of communal uncertainty-reducing device” (Ibid).

Narratives featuring concepts like *rationality*, *belief*, and *inference* would partially constitute what it is to have irrational beliefs, which typically “elude expectations about normal belief formation” (Murphy, 2013, 116). This understanding could have produced folk-psychological and clinical conceptions of *delusion* that refer to both what it is to have a rational belief and what kinds of normal beliefs fit the worldview shared by a certain community (Berrios, 1996). In the case of PP accounts, the personal-level concepts of *delusion* and *irrational belief* have been used to talk about and understand sub-personal false inferences underlying delusion. The human kind of a *schizophrenic patient* has informed the scientific understanding of delusion in the context of scientific research within the PP framework. In turn, the scientific understanding of delusion in terms of false inference may loop back into the human kind of a *schizophrenic patient* in virtue of linguistic labelling and narration in folk-psychological and clinical contexts. The interesting twist here is that researchers working on PP provide an account of how embodied, socially situated brains may bring about the very co-constitutive processes underlying these looping effects.

6 Conclusion

In this paper, we have articulated a proposal for how the sub-personal explanations delivered by PP could have any traction on personal-level concepts within which we conduct and explain our lives. In particular, we have considered the case of delusion and argued that, if PP were true, then personal and sub-personal concepts should display a dynamic co-constitutive relationship with respect to the target phenomena. This relationship is governed by predictive processes at various spatial and temporal scales.

This claim raises several questions. One question is whether the co-constitutive dynamics that would fall out of PP can be evaluated in the light of empirical evidence. One may examine, for example, how a PP account of the large-scale (supra-personal) historical dynamics of psychiatric concepts can fit historical reconstructions like Berrios’s (1996) of how the language of present-day psychiatry has taken shape. Another relevant source of evidence is from computer simulations of the emergence of norms governing the usage of certain explanations in psychiatry, where an individual’s belief revision mechanism is a species of a PP mechanism (e.g., Muldoon, Lisciandra, & Hartmann,

---

5 In many cultures, socially distributed personal narratives include religious beliefs and convictions based on widely shared traditions. In these narratives, religious beliefs may not count as ‘irrational’, because they are the norm, rather than the exception (cf., Murphy, 2013). This lends additional support to the idea that the ascription of personal-level notions such as ‘false belief’ or ‘irrationality’ depends, at least sometimes, on culturally evolved norms that are shaped and re-shaped through narrative practices.
2014), and, more generally, from Bayesian models of cultural evolution and collective sense-making (see, e.g., Suchow, Bourgin, & Griffiths 2017).

A second question is whether PP implies some form of epistemic relativism, whereby there are no absolute norms of rationality. While we have suggested epistemic norms governing evaluation of belief and inference might be, at least partly, constitutively dependent on linguistic labelling, narrative practices, and interactions within epistemic communities, people’s epistemic systems may also include structural, relatively invariant, hyper-priors—perhaps, in the form of topological or hardwired structural constraints in the flow and integration of information—that ground abstract, general epistemic facts, allowing for mutual evaluation across communities. Such structural hyper-priors may be the default cognitive resources people use across contexts to explain behaviour, and would contrast with contextual hyper-priors that can change more rapidly based on instruction or sensory evidence.

A third set of questions concern our assumption that the basic commitments of PP are true. Here we have assumed an empirical flavour of PP, which corresponds to the view that the basic commitments of PP are true of how brains work. This is a claim about the reality of PP posits, which can be rejected in light of empirical evidence. If PP turns out to be false, PP cannot offer an account of how embodied brains bring about looping effects. Yet, our argument about the co-constitutive relationship between the personal, everyday concepts with which people understand and conduct their lives, and sub-personal PP concepts will still stand. It will stand because a different flavour of PP is aesthetic, and this flavour corresponds to the view that explaining mental phenomena in terms of PP organizes scientific research to understand how prediction error minimization underlies the organisation of the world we inhabit (Prinz 2019). While PP would be untestable in this flavour, it would inject an array of concepts within public discourse. This may re-shape people’s image of delusion, and mind more generally.

References


