Received: 29 March 2022 | Revised: 2 June 2023 | Accepted: 14 June 2023

DOI: 10.1111/mila.12471

WILEY

ORIGINAL ARTICLE

Conceptual engineering, predictive processing, and a new implementation problem

Guido Löhr[1] | Christian Michel[2]

[1]Eindhoven University of Technology, Eindhoven, Netherlands

[2]University of Edinburgh, Edinburgh, UK

Correspondence
Christian Michel, School of Philosophy, Psychology and Language Sciences, University of Edinburgh, Dugald Stewart Building, 3 Charles Street, Edinburgh, EH8 9AD, UK.
Email: chris.michel08@gmail.com

Funding information
Dutch Ministry of Education, Culture, and Science and the Netherlands Organization for Scientific Research, Grant/Award Number: 024.004.031

According to predictive processing, an increasingly influential paradigm in cognitive science, the function of the brain is to minimize the prediction error of its sensory input. Conceptual engineering is the practice of assessing and changing concepts or word meanings. We contribute to both strands of research by proposing the first cognitive account of conceptual engineering, using the predictive processing framework. Our model reveals a new kind of implementation problem as prediction errors are only minimized if enough agents embrace conceptual changes. This problem can be overcome by emphasizing the importance of *social norms* and *conceptual pluralism*.

KEYWORDS

active inference, conceptual engineering, conceptual pluralism, predictive processing, the implementation problem

1 | INTRODUCTION

Simon Blackburn (1999) described the philosophical method as a kind of *conceptual engineering*: “For just as the engineer studies the structure of material things, so the philosopher studies the structure of thought” (p. 1). Of course, as Blackburn acknowledged, studying a structure is not enough. An engineer is not usually hired to simply study the structure of a damaged bridge and to identify the parts that need repairing. She is normally asked to create a plan for repairing these parts and to defend this plan to others who may not consider its implementation

Guido Löhr and Christian Michel contributed equally to the manuscript and are listed alphabetically.

feasible. She must convince them because she cannot repair the large bridge on her own. Still, even convincing them is not enough. Eventually, the bridge must be repaired. This is done in a joint effort involving a large team of specialists that adhere to the previously agreed plan.

Following the engineering metaphor, we can divide philosophical methodology into four distinct phases (Löhr, 2023; see also Isaac et al., 2022, for a different review of the engineering stages). First, we are usually struck by a problem with our conceptual system, say, an inconsistency or conceptual gap. These "conceptual disruptions" (Löhr, 2022) can be prompted, for example, by the introduction of a new technological artifact that generates uncertainty about which concepts to apply or "how to go on" (Wittgenstein, 2010, §151). When reflecting on the disruptions, we then identify the conditions that led to them. We call this phase "conceptual assessment".[1] In the words of Blackburn (1999): "Understanding the [conceptual] structure involves seeing how the parts function and how they interconnect. It means knowing what would happen for better or worse if changes were made" (p. 1).

At some point, we usually try to overcome the disruption by creating and implementing a design proposal, that is, by actually changing our conceptual system. We call this stage "conceptual design". Here, we change our own concepts tentatively and counterfactually. However, when designing our own conceptual system, we risk getting out of touch with the conceptualizations of others. This in turn might generate more conceptual disruptions. Thus, we often try to convince others that *they* should change their linguistic or conceptual system as well. This phase can be called "conceptual activism" (Cantalamessa, 2021). Normally, only if others in our reference network or community adopt our proposed changes can we really consider a conceptual engineering (CE) project completed, at least so we will argue. For this reason, we consider conceptual implementation an important part of CE (see Section 4; Figure 1).

A large part of the CE literature has focused on the question of how conceptual designs can be implemented in society, such that they lead to actual changes in linguistic meaning. In fact, the meta-philosophical debate on the nature of CE has only really taken off since Hermann Cappelen (2018) challenged the possibility of actively engaging in changing our conceptual repertoire at all (see also Deutsch, 2020; Koch, 2021; or Jorem, 2021; see Koch et al., forthcoming, for a review). Cappelen argued for this claim based on the assumption that the factors that determine the meaning of linguistic expressions are either inscrutable or out of our control. Some philosophers have also argued that even if implementing conceptual changes were possible, they may potentially change the topic (called "Strawson's challenge", Cappelen, 2018, p. 105; Pinder, 2020; Brun, 2016; but see Koch, 2023).[2]

In this article, we develop the—to the best of our knowledge—first account of how CE projects can be *cognitively implemented*. We propose a cognitive model of what is going on in our minds when we engage in conceptual work in philosophy, politics, or science—a kind of *psychology of philosophy* if you will (Strevens, 2019, Chapter 4). We engage in such practices even if Cappelen were right and they hardly affect the actual meaning of our words or if engaging in CE is either trivial (Deutsch, 2020) or changes the topic (Cappelen, 2018). We are

[1] We could also call this stage "conceptual analysis" (we thank one of the reviewers for suggesting this), but we want to avoid difficult discussions on the nature of concepts and conceptual analysis. Moreover, conceptual analysis may be distinguished from assessing one's internal conceptual structure. Finally, the term "assessment" is more established in the conceptual ethics literature (see also Löhr, 2023 for this use).

[2] As one of the reviewers pointed out to us: Whether conceptual engineering is pointless depends on our goals. We agree, as argued by Podosky (2022) and Löhr (2021), sometimes we care more about practical goals than whether or not we change the topic. Sometimes changing the topic *is* the goal.

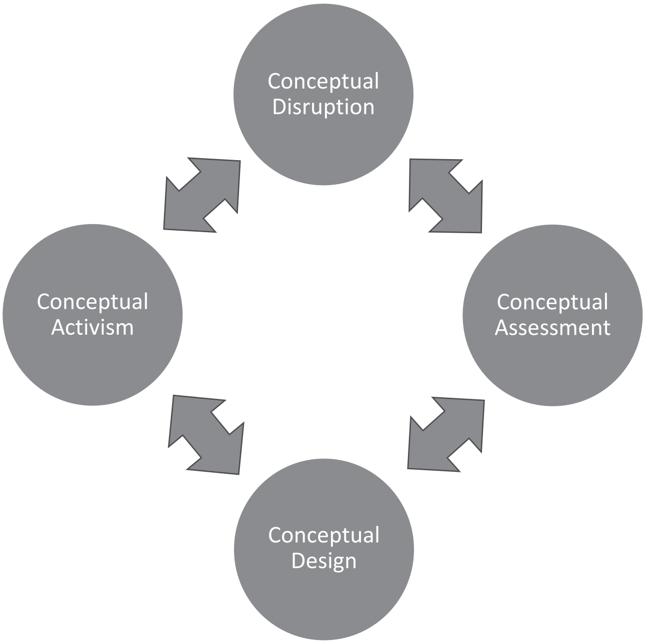

FIGURE 1 Four phases of philosophical methodology construed as conceptual engineering.

especially interested in the cognitive mechanisms underlying the conceptual design and activism stage. We aim to do all of this with a leading and independently plausible theoretical framework in cognitive science called "predictive processing"—PP in short (Clark, 2013, 2016; Friston, 2010; Hohwy, 2013; Parr et al., 2022).[3] The core elements of this framework will be introduced in Section 2.

In Section 3, we apply the predictive processing framework to the different phases of conceptual engineering. We conceive of a *conceptual disruption* as a kind of prediction error, and *conceptual assessment* and *design* as a form of model evaluation and revision. Finally, we conceive of *conceptual activism* as a form of *model enactment* aimed at changing the models of others. We take such a cognitive model of an important philosophical method to be interesting in its own right, but also as offering a proof of concept of the broad applicability of the PP approach. In addition, it contributes to the debate on the nature of CE in three ways. First, our model conceives of CE not as a disembodied intellectual exercise, but as an integrated perception and action cycle. Second, it offers a novel way of distinguishing

[3] A reviewer proposed to endorse a more full-blown "active inference" approach. In fact, we endorse an *active inference formulation of predictive processing* (Clark, 2022). We still decided to keep using the term "predictive processing" (PP) because it is the term members of the conceptual engineering community are more familiar with. Predictive processing is a notion that is also popular in philosophy of the mind and epistemology (see also Parr et al., 2022, p. 198). Technically speaking, PP is a more inclusive framework than active inference. For instance, Parr et al. (2022) state: "[T]he term predictive processing is used in a broader (and less constraint) sense compared to Active Inference" (p. 199). Clark, who initially promoted the notion "predictive processing" and whom we follow, in one of his most recent papers (2022), considers active inference as a specific "formulation" of predictive processing (p. 3) (and, interestingly, there he uses the notion "active inference" more often than "predictive processing"). The exact technical differences do not matter for the current purpose.

conceptual revelations (learning something new about the same concept) from *conceptual revisions* and *replacements* (conceptual change). Third, the model reveals an important cognitive constraint of conceptual engineering—a new kind of (PP) implementation problem. The problem is that we can only minimize our own prediction error by revising our models of the world if enough people change their models as well. Otherwise, we will keep making the wrong social predictions, that is, predictions about the models of other people and their actions. In Section 4, we try to overcome this challenge by emphasizing the importance of expertise, social norms, and conceptual pluralism, none of which has been sufficiently attended to in the CE literature.

2 | THE PREDICTIVE PROCESSING FRAMEWORK IN COGNITIVE SCIENCE

We can liken traditional models of cognition in cognitive science to something that roughly resembles an empiricist view of cognition in epistemology (e.g., Locke, 1998). First, sound and light waves are detected by our sensory system and analyzed to create an uncritical, that is, not yet cognized model of what the world is *really* like—at least its basic or "primary" structures like shapes or motion. This clutter of pure perceptual representation is then used for further processing by our cognitive system. The first and most important function of this cognitive system is a comparison between perception and our model of the world (our "concepts"). Based on the outcome of this comparison, we form a belief and compare it with our desires. Based on this comparison and our other beliefs and desires, we prepare and execute an action, which in turn generates new perceptual input.

Some cognitive scientists call the traditional view of the relation between the senses, cognition, and action "the sandwich model of the mind" (Hurley, 1998; Kirchhoff, 2018; Vetter & Newen, 2014). They compare it to a sandwich because it depicts the mind as consisting of separate layers (the bread being sensorimotor representations and the cheese, pickles and onions being cognition) that communicate with one another in a bottom-up and sequential manner. External input is analyzed in some rudimentary noncognitive manner that is separate from higher-level cognition involving our concepts, beliefs, desires, emotions, and intentions. Moreover, our motor responses in this picture are merely the *output* of this disembodied intellectual process. They are neither part of the cognitive process nor, in any interesting way, involved in perception.

Contrary to the sandwich model of cognition, the predictive processing (PP) model of the mind is more like an onion with multiple layers. Instead of exploiting passively acquired sensory input in a one-way bottom-up feature aggregation process, proponents of the predictive processing approach insist that the mind is cognition all the way down (and up). Cognition and action are two sides of a single coin. The main purpose of the brain is to sustain allostasis—to efficiently prepare the biological organism to anticipate its needs before they arise and to ration life-sustaining resources such as oxygen or insulin (see Barrett et al., 2016; Corcoran et al., 2020). For this purpose, the agent entertains a prediction model of its sensory inputs[4] that stands in a survival-optimizing relation to the external world

[4] Note that "sensory input" needs to be understood—especially in the *active inference* formulation of predictive processing—in the broadest possible sense, including exteroceptive and interoceptive modalities (not just the traditional external senses). Therefore, proprioceptive signals, but also, for example, chemical signals related to the oxygen content in the blood, and so forth, should be seen as "sensory input".

(a so-called hierarchical, probabilistic generative model, Clark, 2013, 2016; Friston, 2010; Hohwy, 2013, 2020).

Prediction errors can be minimized in two ways: by correcting perceptual predictions or by acting on the world such that perceptual predictions are fulfilled. Action can then be understood as a form of prediction in the following way: To grab a piece of pizza, one must predict that the piece of pizza will be in one's hand and then make the prediction come true by moving the hand appropriately. PP can thus be said to unify perception, cognition, and action in the sense that they are considered the consequence of a complex, intertwined process that approximates Bayesian inference. In other words, both perception and action are inferential processes that are the consequence of prediction error minimization of sensory input.[5]

Important for current purposes is that each person's mental model of the world is *hierarchical* in the sense that it is composed of various layers of representation. Each layer generates expectations or predictions (called "priors"), which are compared to the signals from lower levels (see Figure 2). Higher in the hierarchy are relatively stable high-level priors in the form of fundamental "beliefs" that influence all cognition, such as *there is an external world* or *the primary things in the world are objects* (and not, say, color-gradients). Also high in the hierarchy are representations of internalized social norms, which tacitly influence how we think and act. On a middle level, we represent ordinary first-order world knowledge. Below, are sub-symbolic[6] and lower-level perceptual representations. The hierarchical layers are interconnected, and higher levels constrain representations on lower levels in the form of "expectations".

Another important idea associated with the PP approach to the mind is that the model includes a *precision weighting mechanism*, which uses estimates of the precision of the signals to tune the error signals up or down. The mechanism can suppress error signals generated by unreliable or irrelevant input and increase its reliance on prior knowledge. This prevents the models from being unnecessarily updated based, for example, on noisy (unreliable) sensory information. For example, in a foggy environment, we do not normally jump to conclusions and classify an object that looks like a dog with two heads as Cerberus, the monstrous watchdog of the underworld. Instead, we try to calm ourselves down, thinking it might just be a normal dog or an entirely different object we are simply misrepresenting.

Note that even though the PP account is a leading theoretical framework in cognitive science (Hohwy, 2020), it is still very much under construction. Questions that remain controversial are, for example, whether we should be realists about Bayesian inferences in the mind (Clark, 2016; Colombo et al., 2020; Kiefer, 2017) or whether predictive processing is best understood in representational rather than nonrepresentational terms (Kirchhoff & Robertson, 2018). Here, we rely on a rather standard representationalist understanding, endorsed, for example, by Andy Clark (2013, 2016). The key concepts, principles, and mechanisms that this version of PP relies on are gaining more and more theoretical and empirical support (see Clark, 2016 or Hohwy, 2020 for reviews).[7] The implied approximate Bayesian

[5] In active inference, the minimized quantity is "free energy" (e.g., Friston, 2010). Under certain simplifying assumptions, free energy minimization is approximated by (the intuitively more appealing) prediction error minimization (e.g., Friston, 2009).

[6] With "sub-symbolic" we mean representations that are not normally lexicalized or do not correspond to concepts in terms of which we think, like, for instance, edge forms in the visual processing pathway.

[7] For example, the pervasiveness of top-down effects, the hierarchical structure of the brain or the many feedback top-down connections in the brain.

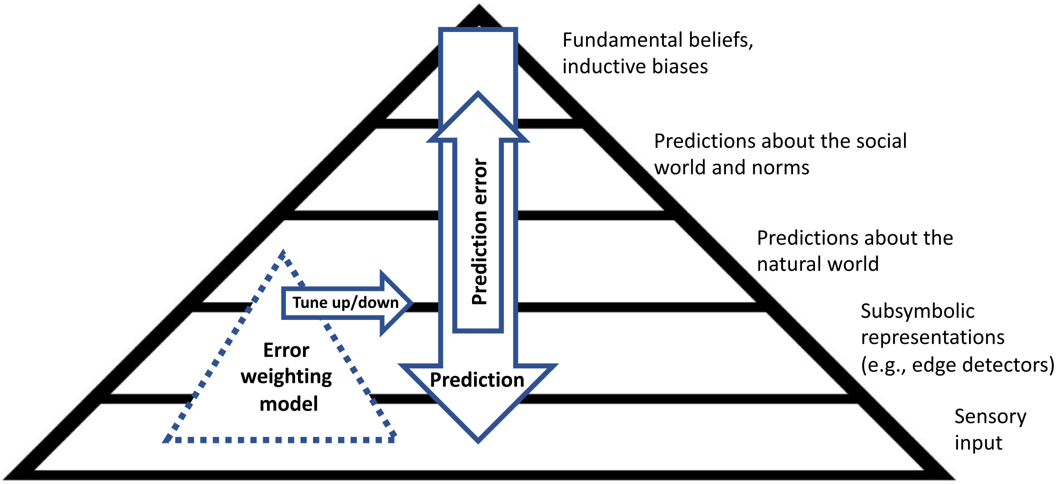

FIGURE 2 A simplified PP prediction hierarchy contains our model of the world, as well as "meta-knowledge" of how to update the model of the world (the error weighting model). Predictions in the many prediction units cascade down and prediction errors are processed upwards. The world model is updated based on the weighted error signals. The error weights correspond to the relevance and reliability of the prediction error signals.

nature of cognition is suggested by many empirical studies (Jiang & Rao, 2021; Millidge et al., 2021; Walsh et al., 2020).[8]

3 | CE AS MODEL ASSESSMENT, REVISION, AND ENACTMENT

PP is a framework of the mind and brain that explains cognitive processes as being aimed at allostasis, that is, the process of adapting the organism to environmental uncertainties and changes. The brain contributes to the overall allostasis by identifying and minimizing prediction errors via changing representational predictive devices that are called "priors". CE is the philosophical method of assessing and modulating representational devices like words and concepts. We can bring both frameworks together by developing a cognitive model of the processes that underly the different phases associated with conceptual engineering: *conceptual disruption*, *conceptual assessment*, *conceptual design*, and *conceptual activism*. Each of these phases will be discussed in the next subsections in terms of the predictive processing framework. A summary of our *PP CE model* can be found in Table 1.

[8] There are also dissenting voices that are more cautious about the prospects of predictive processing more generally, especially with respect to accounting for conceptual thought (e.g., Litwin & Miłkowski, 2020; Williams, 2020). The present work, however, is not a defense of the PP model. We do not rule out that other cognitive architectures may equally well model conceptual engineering. The aim here is to offer one plausible cognitive implementation of a central philosophical method.

TABLE 1 Conceptual engineering terms understood within the PP framework.

Direction of fit	Conceptual engineering	Predictive processing
Model to world	Conceptual disruption	Prediction error
Model to world	Conceptual assessment	Model assessment
Model to world	Conceptual design	Model revision
World to model	Conceptual activism	Model enactment

3.1 | Conceptual disruption as prediction error

From the perspective of predictive processing, a conceptual disruption (an event where we are uncertain about "how to go on", conceptually) can be understood as a prediction error. Prediction errors can occur in the case of an inconsistency in the conceptual system that we are no longer able to ignore. Conceptual disruptions arguably drive much of the research in analytic philosophy. Imagine for example that you are operating under the "prior" that knowledge is justified true belief and you come across Gettier's (1963) paper showing that there are cases where your set of expectations regarding cases of "knowledge" apply but your other priors suggest that these are still not instances of knowledge (a kind of inconsistency). We can think of this conceptual disruption as a prediction error and much of the literature on knowledge as a way to overcome it in a simple and explanatorily powerful manner, that is, as a kind of conceptual design.

A similar kind of conceptual disruption—prediction errors—can be caused by the classifications of other people. Consider the liar paradox, in particular the sentence "This sentence is false". Most people do not know what to make of this statement and will probably shift their attention, that is, give the input a low precision weighting. This shift of attention is driven by higher-level predictions about the lack of existential importance of those prediction errors. As a logician, you do not want to shift your attention. You allow yourself to be confused when someone is telling you "This sentence is false". Your current model of the world is apparently not equipped to make sense of this hypothetical stimulus. It keeps generating error signals that you desperately try to reduce by changing your model. Trying to assign a binary truth value (true or false) to this statement leads to a contradiction that makes it difficult to generate a motor response and expectations about the future. It puts us in a state of conceptual and practical uncertainty.

Another way in which prediction errors can be generated is if our model of the world is not sufficiently adjusted to new events, say, a new technological artifact or scientific finding. This kind of prediction error arguably drives much of the research in the philosophy of technology. For example, imagine seeing a new technological application, say, a highly sophisticated robot, for the first time. Classifying the robot as a mere tool that can be treated in whatever way we like generates the prediction that you may punch and push the robot. This, however, seems to go against certain other priors that anthropomorphize the robot and that predict a more humane interaction similar to that of a person (Coeckelbergh, 2010; Nyholm, 2020). Classifying the robot as a person, however, might commit you to give the robot rights, which contradicts our ordinary predictions about what robots are and how you and other people use them. It would, for example, generate the prediction that it is immoral and even illegal to buy and sell robots, which does not seem to be predicted by your prior that a robot is a product.

3.2 | Conceptual assessment as model assessment

Once we have detected and taken seriously a conceptual disruption—a persistent prediction error—we now advance to the next phase of the CE process: assessing our conceptual resources (*conceptual assessment*). In PP terms, this means that evidence against the model undermines our confidence in it, which motivates the assessment and re-evaluation of the current and alternative models. The main aim of this phase is to identify the conditions in the system that led to the disruption and to find ways to overcome them. Again, one major cause of a disruption is an inconsistency. We, therefore, need to identify the priors we think are in conflict as well as adjacent priors that might be affected if we try to overcome this inconsistency. In the case of the concept of a robot, this might be priors associated with the concepts of personhood and tool.

Besides inconsistencies, we often simply lack the appropriate model to classify a new artifact or event. Such "conceptual gaps" arguably give rise to new simple words (words that are not composed). Imagine a prehistoric community that discovers how to modify a stone such that it can better be used as a hammer. The new artifact might not immediately be classified by the current model because we lack the necessary conceptual structure to make sense of it (Hopster, 2021; Löhr, 2022). We have a model for stones and their different possible properties but perhaps no model or set of priors for thinking about stone as a hammer. To communicate to others about the new artifact, we need to introduce a new label "hammer" that must be integrated into the current conceptual structure, that is, it must find its place among our models for stones, tools, wood, animals, and so forth.

Today, our sophisticated linguistic communities usually come across the opposite problem of a gap—a "conceptual overlap" where two or more models or conceptualizations seem to apply. A popular example to illustrate this problem is the invention of the mechanical ventilator, which assists patients to breathe. This gave rise to patients whose bodies were mostly intact but whose brains displayed no activity (De Boer & Hoek, 2020; Nickel et al., 2022). Do we classify such a patient as dead or alive? Understanding both our conceptual models of death and life is necessary to help us make such a difficult conceptual decision. It involves seeing how parts of each model lead to certain inferences as opposed to others and how they interconnect with other models. "It means knowing what would happen for better or worse if changes were made", to quote Blackburn again.

Finally, the assessment stage also involves identifying constraints on changing our model. Before we can create new priors or conceptual models, we need to study the priors we want to keep stable and which priors we consider expendable or subject to change. For example, in the case of the concept of death, we apparently realized that mental and physical activity is much more important to us than a heartbeat and sustained breathing. Apparently, we value brain activity so much that we classify someone who does not display such activity as *dead* (no longer part of our community). We can imagine a community where things are the other way around, a community that values the body more than the mind and therefore considers a person whose brain is dead "body-alive" (and, therefore, still a member of the community) rather than "brain dead".

In sum, and PP parlance, conceptual assessment implies the counterfactual use of our models of the world that are then evaluated and tested. Such counterfactual testing is crucial for human-level cognition and is possible with the help of the "temporally deep" generative models (see Corcoran et al., 2020) that are posited by the active inference formulation of predictive processing. Furthermore, we can note that the assessment stage is not merely a passive intellectual process. Instead, it is an active kind of problem-solving activity that we often engage in with other people (e.g., in a logic class or as members of the medical profession). In many cases, there is no straightforward recipe for how the model has to be changed to minimize prediction error. Rather, the resolution often necessitates a significant amount of creativity and collective deliberation. Once we have identified the cause of the conceptual disruption, we can go on to the design stage.

3.3 | Conceptual design as model revision

We can overcome conceptual disruptions by means of what we call *conceptual design*. From the PP perspective, we can understand conceptual design as a kind of *structure learning* (see Smith et al., 2020) or *model revision*[9] (revision both of the overall model of the world and parts of the model, that is, models of certain parts of the world). Structure learning means that the model is changed either by adding or deleting nodes or changing structural connections between them. We revise and then select models by temporarily and tentatively changing portions of our overall model of the world. They are then counterfactually tested in the domain in which we noticed inconsistencies. If we can represent the previously problematic domain without any error signal, we can then keep those changes in our own model. The aim is again to minimize the overall prediction error in the long run, such that a perception-action equilibrium is sustained. Note that conceptual assessment and design are construed as an iterative process (see Figure 1).

How can we more clearly distinguish and individuate different forms of conceptual design based on this notion of *model revision*? Specifically, how can we distinguish the notion of *conceptual revelations* (learning something new about our existing concepts) from *intentional conceptual change* (changing our concepts)? This question has occupied several researchers in the CE community (Cappelen, 2018; Haslanger, 2020; see Isaac et al., 2022 for a review). We believe that the PP approach to CE allows us to make interesting distinctions on this topic, which are summarized in Table 2. In particular, it helps us draw a distinction between more or less severe interventions and changes (conceptual reforms and conceptual revolutions): The higher in the network we make changes, the more revolutionary our revision becomes.

We can construe *conceptual revelations* as changes in the *relation* between models or sets of priors rather than changing their internal structure. Revelations arguably happen when a child learns that ice is essentially frozen water, when we learn that birds are avian dinosaurs or that dolphins are mammals as opposed to fish. It does not fundamentally change our models of ice and water but merely relates them in a way that better accounts for the perceptual input, for example, by subsuming the model of ice under the model of water or the model of dolphin under the prior of mammal. Note that conceptual revelations may also be mixed with conceptual revisions. In this process, additional concepts may be introduced that serve as the link. In the *ice is frozen water* example, the concept of "state of aggregation" was important. Ice and (liquid) water are unified by this concept as being different manifestations of the same substance, H_2O.

[9] An anonymous reviewer has suggested using the term "model selection", which includes both the selection of a new or different model and the modification of an existing model. To avoid confusions of this more technical use of "model selection" with a more common-sense notion (as selection of a new/different model), we decided to use "model revision".

TABLE 2 An overview of the different kinds of conceptual design and different forms of impact.

Conceptual design	Model revision (PP)	Examples
Conceptual revelation (learning something new about existing concepts)	Changes in relations between existing priors or models, often by connecting existing models while the models themselves remain relatively stable	Ice is frozen water; water is H_2O; birds are dinosaurs; dolphins are mammals
Conceptual revision (changing the same concept)	Structural change in the same model (adding or deleting nodes, making new connections between models while keeping the model's root node)[a]	A more inclusive concept of family or marriage; Haslanger's revision of the concept of woman; deciding Pluto is not a planet
Conceptual replacement (replacing concepts)	Complete replacement of a model by a new one (such as the introduction of a new root node, while replacing a similar one in its place)	Einstein's notion of time; Machery's (2009) proposal to eliminate the concept of concept
Notions related to the impact on the global PP model		
Conceptual reform	Change to peripheral priors that do not have a larger effect on the overall system	Thinking of a patient without brain activity as brain-dead rather than body-alive
Conceptual revolution	Change to high-level priors that generate substantial change in lower levels	Einstein's theory of general relativity; Copernicus' theory that the Earth rotates around the Sun

Note: These different kinds may overlap.

[a] We consider concepts to be individuated by a root-node, from which other nodes, for example, feature nodes emanate (e.g., Michel, 2020, 2022). This allows us to distinguish between the change to an existing concept and the replacement of a concept.

Conceptual revision means changing a concept rather than merely learning something new about this concept. CE advocates find it notoriously difficult to distinguish conceptual revision from revelation and replacement because it presupposes a theory of concept individuation (Isaac et al., 2022 for a review of this issue). In PP terms, this issue can be tackled in the following way: While a revelation is a change between models or priors, a conceptual revision is an internal change of a prior while leaving its relations to other prior structures more or less intact. For example, we might change the expectations of when to apply the word "woman" not because we learned something new about women but because we hope that this change of the model will lead to a more just world (Haslanger, 2000). Again, we acknowledge that this distinction is hardly ever clear-cut and conceptual revision may also involve revelations.

Finally, *conceptual replacements* are changes where we replace one notion with another, for example, by eliminating one notion of time with a completely different one, while keeping the term "time". In PP terms, this means that we eliminate an existing prior and replace it with a new one, while keeping the same term, in many cases. We might think of Albert Einstein's introduction of a completely new concept of time to physics as one of those replacements that made the old concept of time in physics more or less redundant. It still, however, played some of the explanatory roles of the old notion of time in physics, which is why we can speak here of a replacement rather than merely the introduction of a new model that happens to be given the same name. Similarly, Machery (2009) argued that we should eliminate the notion of concept from psychology and replace it with three other notions (*prototype, exemplar*, and *theory*).

Conceptual revisions, replacements, and revelations may be considered types of "conceptual reforms" if the changes happen mainly in the periphery of the network in a way that does not fundamentally change it. However, we can also make sense of more radical changes, which we might call "conceptual revolutions". Conceptual revolutions occur when we change more fundamental priors, such that the entire system is affected or needs to be updated. For example, we can think of Einstein's concept of time in physics as replacing an older but very fundamental concept. This was, again, not due to a new empirical finding or revelation but an intentional design choice to generate more consistency in the model. It was a case of intentional model revision that led to dramatic changes in the network in both lower and also higher levels.

The result of conceptual revolutions (but also reforms sometimes) may be that the new system is difficult to translate back to the old system, which may lead to a form of *incommensurability* (Baker, 2019; Kuhn, 1962). This will give rise to a problem in communication with others who did not undergo such dramatic changes as we will see below, which makes conceptual revolutions especially interesting from a PP perspective. The difficulty of predicting such changes generates considerable "inferential risk" and may even explain resistance in other speakers to adopting such radical changes, especially if we have to do with fundamental moral concepts like *person* or *right*. Think again of the example of whether sophisticated robots are persons with rights or mere tools. If you decide on the former, this will be highly disruptive throughout the entire conceptual and representational system. This is a real (albeit risky) conceptual revolution because it is difficult to predict the changes in the periphery.

3.4 | Conceptual activism as model enactment

Finally, another option to overcome a prediction error or conceptual disruption according to the active inference formulation of the PP account is via world-changing action. Instead of changing our model of the world, we change the world so that it fits our model—a kind of "model enactment". One way this can work is by simply eliminating objects from the world that generate conceptual disruptions. It has been argued, for example, that certain ultra-realistic robots or new ways of intervening or engineering our genes should be banned because they challenge or disrupt fundamental social concepts like the concept of personhood or the concept of autonomy (Boden et al., 2017; see Nyholm, 2020 for a review). Another, for our purposes more interesting, kind of model enactment involves changing the conceptual models of others. We discuss this latter form of conceptual activism in more detail in Section 4.

Note that both kinds of model enactment may be construed as a form of *conceptual preservation* (Lindauer, 2020)—a conservative form of conceptual engineering. Some authors may prefer reserving the term "conceptual engineering" only for the design phase, but we take this to be a limiting terminological decision, especially for our purpose here. Again, the way we have been using the term *CE* includes several phases or actions with conceptual design as merely one phase among many. As we argue in Section 4, the activism or enactment phase in particular is important for any CE project and can, therefore, not be ignored. The design phase is usually construed as a more or less private matter that happens individually or in small groups (e.g., a logic class or political party). Things become interesting once we try to implement these changes in the broader community.

4 | IMPLEMENTATION: CHANGING THE MODEL OF OTHERS

4.1 | The PP implementation problem

Imagine that you have successfully overcome your personal conceptual disruption. You have just adjusted your system of priors such that no prediction error occurs anymore for the relevant event, object, or word. For example, imagine you just solved the liar paradox after many hours of conceptual design in your armchair. We argue that the PP model of CE now reveals a novel kind of implementation problem. We call it *the PP implementation problem*. This problem arises because we will only engage in conceptual design if it promises the overall minimization of a prediction error in the long run. However, when you successfully change your own model of the world, your motor output will be adjusted to the new priors. These new predictions and behaviors likely generate conceptual disruption or prediction errors *in others*. They will receive perceptual stimuli from you that are not predicted by their models. Moreover, these people will act in ways that are not compatible with your *new* model. This will generate further mismatches in your perception and your motor output will not generate the predicted perceptual input of other people in your community.

Imagine for example that we propose to replace the folk concept of truth with the concepts *ascending truth* and *descending truth*, as proposed by Scharp (2013). At least for certain purposes, both concepts promise to be superior by being able to avoid certain logical inconsistencies. In this case, the conceptual design will consist in making a distinction that is more fine-grained than before. If adopted, this new way of thinking about the world and truth will generate perceptual predictions in other contexts that might generate even more prediction errors than the change was able to reduce. We predict that other people should now do (for instance assert) certain things pertaining to the concept of truth. If this is not the case, we attain a prediction error that forces us to reconsider or reassess our old conceptual system. These new prediction errors might be larger than the prediction error generated by certain logical inconsistencies.

Moreover, while we can usually suppress certain prediction errors if we consider them irrelevant (and hence assign them a low precision), the socially generated prediction error is more difficult (although not impossible) to suppress or ignore.[10] Speakers that create a prediction error in our model usually demand a response that fits more or less their model. Thus, we usually have to “play along” if we want to be understood by others. A major constraint of any PP implementation is then that it depends heavily on the priors of other people: It is difficult to adopt a conceptual change privately. This is what we consider an important constraint on the success of any CE project:

> **The PP implementation constraint:** For model revision to take place, the prediction error likely generated by the designed concept must be predicted to be *significantly lower* than the prediction error of the current model of the world, at least in the long run.

[10] This assumes that some social norms are weighted strongly—and sometimes go against perceptual evidence (thanks to a reviewer to point to this issue). Here is no space to expand on this but we think it is plausible that we have—as the result of evolutionary forces—strong social priors (specifically the bias to conform to social norms) to avoid exclusion from the community or society. Also think of the phenomenon of brainwashing to see how higher-level priors can override evidence.

Note that this is not a normative or metaphysical constraint. It is a cognitive constraint. Even if we decide that a certain model is preferable and should be implemented broadly (perhaps for moral reasons), our minds may not be built in such a way that they allow for following this recommendation.

4.2 | The importance of conceptual pluralism

There are several ways in which we can overcome the implementation problem. One obvious solution is to embrace *conceptual pluralism*. We can adopt a new model in one context, including different social contexts, while switching back to the old model when we are in another context. Conceptual pluralism also offers a way out of the socio-cognitive problem that models between people likely generate more prediction errors if they are out of synch or if we lack sufficient knowledge of their models. In such a case, we can adopt different models and switch between them when interacting with different people. Conceptual pluralism is surprisingly seldom discussed in the CE literature (for exceptions, see Isaac, 2021 or Belleri, 2021, Section 5), but, as we argue, it might be considered a critical element of any CE project.

How does a “switch” between different conceptual frameworks work? The key notion here is again *attention* (e.g., Feldman & Friston, 2010; Hohwy, 2013). Attention is understood in the PP framework as increasing the error-signal sensitivity of the relevant part of the generative model. This is implemented via the precision weighting model. As a reminder, this mechanism tunes down error signals that are estimated to be unreliable or irrelevant and tunes up signals that are estimated to be reliable and relevant. This allows us to effectively “shut down” parts of the model and selectively focus on the parts of it that are relevant to the specific situation. A high-level norm or context prior might serve as such a switch. The concept of a tomato, for example, can be represented in terms of two conceptual frameworks as two little sub-models in the world model. By recognizing the context or situation (e.g., being at home in the kitchen) via the corresponding prior the proper conceptual framework is selected by tuning up its error sensitivity. In this way, the precision weighting mechanism focuses our attention on the “tomato as a vegetable” part of our tomato model rather than the “tomato as a fruit” part.

Conceptual pluralism not only allows us to find a relatively simple way out of the implementation problem. It also inspires us to re-think the success conditions of a CE project. It is often argued that such a project is successful if a new concept is adopted by the community. We argue that CE projects might be successful even if they are only implemented in a small section of the larger community for which the conceptual change was targeted—and even if this part of the community may have adopted the change only in some limited contexts (e.g., in the classroom or at work or when reading Heidegger). Something similar is suggested here for more ordinary contexts. We might for example argue that adopting Haslanger's (2000) notion of *woman* in certain academic contexts sheds light on the world in a certain way that is useful in some contexts while keeping the original notion in other contexts (as Haslanger agrees). Similarly, we might use Scharp's notions of truth in logic and philosophy classes while retaining the ordinary notion in more ordinary contexts.

4.3 | The importance of expertise

Besides conceptual pluralism, another way of overcoming the PP implementation constraint is, of course, to convince others to change their model. We take it that the main challenge for the conceptual activist is then to give the other person or group an incentive (again, cashed out in terms of prediction error minimization) to change their model and to refrain from trying to convince us that we are wrong (that the others themselves engage in a sort of *conservative* conceptual activism as a reaction to our intent to change their model). Other people will change their models only if the change promises an overall reduction in prediction error in the long run.

The first major obstacle to incentivizing a change in one's predictive models is often that the proposed changes do not immediately make sense to us. If the proposed change is significant, it might not even be believable, and the reasons presented for this change might not be intellectually accessible. This might be because they require background knowledge that is lacking or because the details are too complex. For example, imagine someone telling you that Pluto is no longer a planet. Your immediate reaction might be that this itself is a conceptual disruption. You are confused about how a planet can suddenly stop being a planet. Did it explode perhaps? This person explains certain facts to you, but they are somewhat difficult to follow, and you are not sure whether and how you should update your conceptual knowledge about Pluto. We argue that in cases where the reasons for a model change are not immediately obvious to us, we likely still update our world model if the speaker is recognized as more competent (see Table 3).

To use a familiar example in philosophy, imagine that I use the word "arthritis" wrongly in a conversation with a doctor because I think it is something that causes pain in my thigh (Burge, 1979). The doctor corrects me and explains that arthritis is an illness of the joints. Given my belief that this doctor is an expert, I probably accept the correction of the doctor. This means that I update my model by reorganizing a part of it, namely everything related to my priors or model of arthritis. In this case, it would be inappropriate to insist to the doctor that it is *she* who uses the word "arthritis" wrongly and try to change the doctor's model to fit my use of the term. It would be irrational (given my meta-knowledge or hyper-priors) to try to change the world such that now my use of "arthritis" fits the world because I am no expert on arthritis and am also not generally viewed as one.

Translated into PP terms, my precision weighting mechanism determines that the error signal should not be suppressed. I clearly hear what the doctor says (there is no background noise, and the doctor speaks clearly), but given that I predict (have a higher order prior) that the doctor is a highly competent user of the concept of arthritis, I take his conceptualization very seriously. In other words, the prediction error signal between the trustworthy bottom-up evidence provided by the doctor and my top-down priors is tuned up to produce a bottom-up correction of my model. This means that the evidence—judged to be reliable and relevant—overrides prior beliefs in this case. If I had not clearly heard what the doctor said or found the doctor not to be trustworthy, then I might suppress this error signal and move on without adjusting my model.

4.4 | The importance of social norms

However, even if you convince another person of the epistemic advantages of the new conceptual model and of your expertise, none of this ensures that it is preferable for them to in

TABLE 3 Possible competence estimates and consequences (from the "patient's" perspective).

	Engineer	
Patient	**Competent (teacher, expert)**	**Not competent (student, layperson)**
Competent	Ignore/assess further	Ignore/assess further
Not competent	Consider adopting the change	Ignore/assess further

fact implement the changes in the long run. One reason for this resistance might come from the rest of the community, which places further constraints on the conceptual changes we may prefer to implement. If the broader community has not adopted the suggested changes or is actively resisting them, this incentivizes the recipient of an engineering proposal against implementing the recommended changes even if she might accept their superiority. We might have good reasons for choosing certain conceptual changes but if we cannot convince others to adopt these changes, we either become isolated or must accept (suppress) the prediction error caused by our old model of the world (or endorse some form of conceptual pluralism, see Section 4.2). We believe that another way in which the implementation constraint can be met is by changing social norms.

According to Bicchieri (2016), social norms are rules of behavior such that individuals prefer to conform to them under the condition that they expect that (a) most people in their reference network (friends, family, colleagues) conform to them (empirical expectations), and (b) that most people in their reference network believe they ought to conform to them. For example, the rules "do not steal", "greet your friends when you see them", or "call Pluto a dwarf planet" are all social norms in Bicchieri's account. This is because we expect that most people that we care about will follow these rules and that most people believe that we ought to follow them. So, to really implement large-scope conceptual changes, we need to change the relevant expectations of the members of a community (Nimtz, 2021; Thomasson, 2021). You are more likely to adopt a change if you expect that others will not judge you for it, or better still if you expect them to judge you if you do not make these conceptual changes. This currently dominant account of social norms fits well with the PP framework of cognition.

Within the PP framework, we can construe a social norm as a set of complex higher-level priors that represents knowledge about mutual expectations or predictions in the community. The useful role of social norms in the PP model has also been emphasized by Colombo (2014) and Clark (2016, p. 286) as a means to help reduce prediction errors by making behavior mutually more predictable. To really implement a conceptual change (conceptual activism), we must therefore not only convince the individual that a certain change is theoretically preferable or superior. We also have to change the predictions (expectations) of others in our reference network about (a) what other people in their reference network will probably do, and (b) what those people believe one should do, including how one should use a word or apply a concept.

> **Conceptual engineering success condition (according to the PP model):** To implement a conceptual change in the larger society, we must change the predictions of the majority of members in our reference network about how other people in their reference network will probably apply words and how those people believe one should apply the receptive word. In other words, we must change our social norms. More modest conceptual engineering projects can rely on a kind of conceptual pluralism.

We would like to refer here again to Bicchieri (2016) for an account of how we might accomplish the change in social norms, that is, the change in the relevant expectations with respect to the use of a word or concept of others. One way of implementing changes in social norms, according to Bicchieri, is by convincing trendsetters to visibly change their behaviors. This might legitimize the new rule and incentivize others to follow it because they predict that others will follow it as well or at least will not judge the individual who embraces the new rule. For example, an influential intellectual might be convinced to use and explain their use of a new concept of a woman in certain contexts. This then changes the expectations we have with respect to what others expect to be the case and what they expect others believe should be the case.

Another way of incentivizing social change is to engage in some form of *conceptual nudging*, that is, an attempt to change people's expectations indirectly or by means of positive reinforcement, essentially by changing certain environmental constraints. One example of a nudge is for example to decorate one's office with pictures of public leaders or intellectuals who are from a minority in order to reduce bias, for example, during a job interview. The idea is that these pictures change one's predictions about the kind of people we predict or expect to be successful. How exactly we accomplish changing the predictions of the majority cannot be fully discussed here. However, a committed conceptual activist could use the methods introduced in behavioral economics, for example, to try to implement changes in individuals by trying to change their incentives to change their actions. Only if the social norms have been adopted can the individual truly adopt the new conceptual framework and make it her own.

5 | CONCLUSION

According to a popular model of the mind in cognitive science, predictive processing, the mind is a prediction machine whose aim is to minimize prediction errors to keep the individual in the best possible state for surviving and thriving in an uncertain environment. In this article, we applied this view to the topic of conceptual engineering. Prompted by a conceptual disruption (conceived of as a prediction error), conceptual engineers then often assess their generative model of the world to identify where and why the prediction error occurs (conceptual assessment). Within the PP framework, the minimization of the prediction error by revising our model can be construed as a kind of conceptual design. Once a design is identified that best allows for minimizing prediction error in the long run, the conceptual engineer tends to engage in conceptual activism to convince others to adopt the changes. We argued that although implementation is difficult, it is not impossible in the PP model either if we endorse some form of conceptual pluralism or if there are enough incentives for others to engage in model revision, for instance, if we can convince influential individuals of the change in the hope that our social norms will change.

ACKNOWLEDGMENTS

We are grateful to the two anonymous reviewers, Steffen Koch and Jeroen Hopster for the useful comments.

DATA AVAILABILITY STATEMENT

There are no data available.

ORCID

Guido Löhr https://orcid.org/0000-0002-7028-3515
Christian Michel https://orcid.org/0000-0001-9962-5403

REFERENCES

Baker, R. (2019). *The structure of moral revolutions: Studies of changes in the morality of abortion, death, and the bioethics revolution*. MIT Press.

Barrett, L. F., Quigley, K. S., & Hamilton, P. (2016). An active inference theory of allostasis and interoception in depression. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1708), 20160011.

Belleri, D. (2021). On pluralism and conceptual engineering: Introduction and overview. *Inquiry: An Interdisciplinary Journal of Philosophy*, 1–19. https://doi.org/10.1080/0020174X.2021.1983457

Bicchieri, C. (2016). *Norms in the wild: How to diagnose, measure, and change social norms*. Oxford University Press.

Blackburn, S. (1999). *Think: A compelling introduction to philosophy*. Oxford University Press.

Boden, M., Bryson, J., Caldwell, D., Dautenhahn, K., Edwards, L., Kember, S., Newman, P., Parry, V., Pegman, G., Rodden, T., Sorrell, T., Wallis, M., Whitby, B., & Winfield, A. (2017). Principles of robotics: Regulating robots in the real world. *Connection Science*, *29*(2), 124–129.

Brun, G. (2016). Explication as a method of conceptual re-engineering. *Erkenntnis*, *81*(6), 1211–1241.

Burge, T. (1979). Individualism and the mental. *Midwest Studies in Philosophy*, *4*, 73–121.

Cantalamessa, E. A. (2021). Disability studies, conceptual engineering, and conceptual activism. *Inquiry*, *64*(1–2), 46–75.

Cappelen, H. (2018). *Fixing language: An essay on conceptual engineering*. Oxford University Press.

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

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

Clark, A. (2022). Extending the predictive mind. *Australasian Journal of Philosophy*, 1–12. https://doi.org/10.1080/00048402.2022.2122523

Coeckelbergh, M. (2010). Robot rights? Towards a social-relational justification of moral consideration. *Ethics and Information Technology*, *12*(3), 209–221.

Colombo, M. (2014). Explaining social norm compliance. A plea for neural representations. *Phenomenology and the Cognitive Sciences*, *13*(2), 217–238.

Colombo, M., Elkin, L., & Hartmann, S. (2020). Being realist about Bayes, and the predictive processing theory of mind. *The British Journal for the Philosophy of Science*, *72*(1), 185–220.

Corcoran, A. W., Pezzulo, G., & Hohwy, J. (2020). From allostatic agents to counterfactual cognisers: Active inference, biological regulation, and the origins of cognition. *Biology and Philosophy*, *35*(3), 1–45.

De Boer, B., & Hoek, J. (2020). The advance of technoscience and the problem of death determination: A promethean puzzle. *Techné: Research in Philosophy and Technology*, *24*(3), 306–311.

Deutsch, M. (2020). Speaker's reference, stipulation, and a dilemma for conceptual engineers. *Philosophical Studies*, *177*(12), 3935–3957.

Feldman, H., & Friston, K. J. (2010). Attention, uncertainty, and free-energy. *Frontiers in Human Neuroscience*, *4* (215), 1-23. https://doi.org/10.3389/fnhum.2010.00215

Friston, K. (2009). The free-energy principle: A rough guide to the brain? *Trends in Cognitive Sciences*, *13*(7), 293–301.

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

Gettier, E. L. (1963). Is justified true belief knowledge? *Analysis*, *23*(6), 121–123.

Haslanger, S. (2000). Gender and race: (What) are they? (what) do we want them to be? *Noûs*, *34*(1), 31–55.

Haslanger, S. (2020). How not to change the subject. In T. Marques & A. Wikforss (Eds.), *Shifting concepts: The philosophy and psychology of conceptual variability* (pp. 235–259). Oxford University Press.

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

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

Hopster, J. (2021). What are socially disruptive technologies? *Technology in Society*, *67*, 101750.

Hurley, S. L. (1998). *Consciousness in action*. Harvard University Press.

Isaac, M. G. (2021). Post-truth conceptual engineering. *Inquiry: An Interdisciplinary Journal of Philosophy*, 1–16. https://doi.org/10.1080/0020174X.2021.1887758

Isaac, M. G., Koch, S., & Nefdt, R. (2022). Conceptual engineering: A road map to practice. *Philosophy Compass*, *17*(10), e12879.

Jiang, L. P., & Rao, R. P. (2021). Predictive coding theories of cortical function. *arXiv*. https://doi.org/10.48550/arXiv.2112.10048

Jorem, S. (2021). Conceptual engineering and the implementation problem. *Inquiry: An Interdisciplinary Journal of Philosophy*, *64*(1–2), 186–211.

Kiefer, A. (2017). Literal perceptual inference. In T. Metzinger & W. Wiese (Eds.), *Philosophy and predictive processing*. MIND Group.

Kirchhoff, M. (2018). Predictive brains and embodied, enactive cognition: An introduction to the special issue. *Synthese*, *195*(6), 2355–2366.

Kirchhoff, M. D., & Robertson, I. (2018). Enactivism and predictive processing: A non-representational view. *Philosophical Explorations*, *21*(2), 264–281.

Koch, S. (2021). The externalist challenge to conceptual engineering. *Synthese*, *198*(1), 327–348.

Koch, S. (2023). Why conceptual engineers should not worry about topics. *Erkenntnis*, *88*(5), 2123–2143.

Koch, S., Löhr, G., & Pinder, M. (Forthcoming). Recent work in the theory of conceptual engineering. *Analysis*. https://doi.org/10.1093/analys/anad032

Kuhn, T. S. (1962). *The structure of scientific revolutions*. University of Chicago Press.

Lindauer, M. (2020). Conceptual engineering as concept preservation. *Ratio*, *33*(3), 155–162.

Litwin, P., & Miłkowski, M. (2020). Unification by fiat: Arrested development of predictive processing. *Cognitive Science*, *44*(7), e12867.

Locke, J. (1998). *An essay concerning human understanding*. Penguin Classics.

Löhr, G. (2021). Commitment engineering: Conceptual engineering without representations. *Synthese*, *199*(5–6), 13035–13052.

Löhr, G. (2022). Linguistic interventions and the ethics of conceptual disruption. *Ethical Theory and Moral Practice*, *25*, 835–849.

Löhr, G. (2023). If conceptual engineering is a new method in the ethics of AI, what method is it exactly? *AI Ethics*, 1–15. https://doi.org/10.1007/s43681-023-00295-4

Machery, E. (2009). *Doing without concepts*. Oxford University Press.

Michel, C. (2020). Overcoming the modal/amodal dichotomy of concepts. *Phenomenology and the Cognitive Sciences*, *20*, 655–677.

Michel, C. (2022). A hybrid account of concepts within the predictive processing paradigm. *Review of Philosophy and Psychology*, 1–27. https://doi.org/10.1007/s13164-022-00648-8

Millidge, B., Seth, A., & Buckley, C. L. (2021). Predictive coding: A theoretical and experimental review. *arXiv*. https://doi.org/10.48550/arXiv.2107.12979

Nickel, P. J., Kudina, O., & van de Poel, I. (2022). Moral uncertainty in technomoral change: Bridging the explanatory gap. *Perspectives on Science*, *30*(2), 260–283.

Nimtz, C. (2021). Engineering concepts by engineering social norms: Solving the implementation challenge. *Inquiry: An Interdisciplinary Journal of Philosophy*, 1–28. https://doi.org/10.1080/0020174X.2021.1956368

Nyholm, S. (2020). *Humans and robots: Ethics, agency, and anthropomorphism*. Rowman & Littlefield Publishers.

Parr, T., Pezzulo, G., & Friston, K. J. (2022). *Active inference: The free energy principle in mind, brain, and behavior*. MIT Press.

Pinder, M. (2020). On Strawson's critique of explication as a method in philosophy. *Synthese*, *197*(3), 955–981.

Podosky, P. M. C. (2022). Can conceptual engineering actually promote social justice? *Synthese*, *200*(160), 1–22.

Scharp, K. (2013). *Replacing truth*. Oxford University Press.

Smith, R., Schwartenbeck, P., Parr, T., & Friston, K. J. (2020). An active inference approach to modeling structure learning: Concept learning as an example case. *Frontiers in Computational Neuroscience*, *14*(41), 1–24.

Strevens, M. (2019). *Thinking off your feet: How empirical psychology vindicates armchair philosophy*. Harvard University Press.

Thomasson, A. (2021). Conceptual engineering: When do we need it? How can we do it? *Inquiry: An Interdisciplinary Journal of Philosophy*, 1–26. https://doi.org/10.1080/0020174X.2021.2000118

Vetter, P., & Newen, A. (2014). Varieties of cognitive penetration in visual perception. *Consciousness and Cognition*, *27*, 62–75.

Walsh, K. S., McGovern, D. P., Clark, A., & O'Connell, R. G. (2020). Evaluating the neurophysiological evidence for predictive processing as a model of perception. *Annals of the New York Academy of Sciences*, *1464*(1), 242–268.

Williams, D. (2020). Predictive coding and thought. *Synthese*, *197*(4), 1749–1775.

Wittgenstein, L. (2010). *Philosophical investigations*. John Wiley & Sons.

How to cite this article: Löhr, G., & Michel, C. (2023). Conceptual engineering, predictive processing, and a new implementation problem. *Mind & Language*, 1–19. https://doi.org/10.1111/mila.12471