Philosophical Psychology

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/cphp20

Predicting ordinary objects into the world

Arthur C. Schwaninger

PHILOSOPHICAL PSYCHOLOGY
https://doi.org/10.1080/09515089.2022.2123310

Routledge
Taylor & Francis Group

ARTICLE

OPEN ACCESS Check for updates

Predicting ordinary objects into the world

Arthur C. Schwaninger

Department of Philosophy, Theoretical Philosophy, University of Zurich, Zurich, Switzerland

ABSTRACT
Ordinary objects are experienced to endure over space and time, to not be collocated with each other, to be composed of proper parts, and to survive the loss of some of their parts. These qualities are on the one hand difficult to reconcile for theorists of perception and on the other hand pose a variety of problems when considered in isolation. Relying on the theoretical framework of predictive processing, this paper argues that we can use the category of a robust predictive process to conceptualize qualities such as persistence and compositionality in a unified manner. Traditional problems concerning the structural properties of ordinary objects, such as the question of when two objects compose, can then be reformulated using this new category.

ARTICLE HISTORY
Received 27 October 2021
Accepted 3 September 2022

KEYWORDS
Predictive processing; metaphysics; perception; ordinary objects; phenomenology; kantianism

1 Introduction

Humans experience ordinary objects (OOs), such as chairs and tables, as things that endure over space and time, that are not collocated with each other, that are composed of proper parts, and that survive the loss of some of their parts. Multiple articles in the theory of perception try to make sense of how our visual representations generate the 'visual ontology' (Skrzypulec, 2016, p. 261) of our experiences (e.g., Green, 2019, 2021; O'Callaghan, 2016; Skrzypulec, 2016, 2018). By turning to findings in the cognitive sciences, they explain why it is that we see certain parts of our visual field to compose to single objects with the properties we experience them to have. These approaches are reminiscent of Gestalt psychology that tries to detect rules that the mind employs to generate perceptual units (Jäkel et al., 2016).

Based on recent developments in the neurosciences, this paper makes two contributions to this field of research by turning toward predictive processing (PP)—a novel theory of brain functioning. First, it explains how the category of *robust predictive processes* (RPPs) allows us to unify classical concepts employed to characterize the objects in our visual ontology such as compositionality and persistence under a single umbrella term. Second,

CONTACT Arthur C. Schwaninger arthur.schwaninger@uzh.ch Department of Philosophy,Theoretical Philosophy, University of Zurich, Philosophisches Seminar, Zürichbergstrasse 43, Zurich CH-8044, Switzerland

using this new concept the paper explains why it is that the sensory data are interpreted by the brain to have the structure the way we experience it.

The paper first discusses the problem of underdetermination and why there is reason to believe that certain structural qualities of OOs are the result of our cognitive processes (Sec. 2). The paper then introduces our modern understanding of cognition and how it differs from traditional approaches that have been employed in philosophical discussions (Sec. 3). Section 4 presents the core idea of this paper and elaborates how the category of RPPs allows us to unify different aspects of OOs in our visual ontology. Finally, I explain the advantages of employing RPPs in our theorizing (Sec. 5).

2 Imposing structure onto the visual image

When information about the external world is received through our eyes, it is further projected to a region in the thalamus called the lateral geniculate nucleus (LGN) and from there continues its path to the visual cortex where the information gets further processed. While traditional neuroscience has mostly investigated the computations in the visual cortex (see, e.g., Hubel & Wiesel, 1962), it is today a well-known fact that only roughly 20% of the information entering the LGN comes from the senses; the remaining 80% comes from other parts of the brain that convolute the sensory information before it continues to the visual cortex (Churchland & Sejnowski, 1988; Varela et al., 1991, p. 94–95).

But why would the brain add new information to the visual input? The reason is that it has no other feasible choice. It is a universal fact that the stimulations onto the senses of any cognizing system are underdetermined by their causes. The resulting problem is sometimes referred as the *inverse problem* of perception and suggests that any cognitive system is required to find a way to invert the mapping from external causes to sensory effects – a difficult task, given that different causes can result in having the same effects. This means that the information about the shapes of objects and their relations to one another as well as their parts 'must be derived from the sum of retinal information together with various assumptions about the structure of the world' (Shapiro, 2011, p. 29)—assumptions about objecthood that are not necessarily veridical or amount to primary properties but might merely be improving the fitness of the cognizing system. The information entailed in the sensory data allows in principle for many possibilities as to how OOs could be experienced. Given a cognizer that receives sensory data, there is a *multiplicity of possibilities* of what the experienced qualities of OOs could be and how the visual ontology is structured. Using feedback loops that add information from different parts of the brain to the information coming from the senses, the lacking structural information in the data is complemented by assumptions made by the brain about the world. What is investigated in this paper are these additional structural aspects that our brains impose on the sensory data for the purpose of seeing objects in a way that is beneficial to our needs but, due to underdetermination, are not necessarily veridical. Thus, *given* our visual sensory input this paper suggests that the structural aspects of our visual ontology are a product of how the brain computes sensory information. For instance, whether the universe consists of particles in motion that persist over space and time or whether we occupy a universe of cellular automata that are motionless (see, e.g., Wolfram, 2002) is a question for physicists or metaphysicians to decide. Our sensory data simply cannot provide us with an answer to this question. However, it is clearly the case that we experience OOs to persist over space and time, and this paper offers an explanation of how the brain adds this structural element to the given visual sensory data.[1]

It is important to note that much of the cognitive science literature does not clearly state that certain aspects of our visual ontology are not features of the world independent from us. The inverse problem is typically explained in a way that implies that the brain has the ability to reconstruct the 'properties of external objects' (Wiese & Metzinger, 2017, p. 4), although in fact much of our visual ontology is not at all a reconstruction of how the world is. This does not mean that no aspect of our visual ontology is veridically reconstructed as sometimes suggested (see, e.g., Hoffman, 2019), but it does mean that at least certain structural aspects of OOs are added by our perceptual processes without there being any corresponding mind-independent structure in the world.

3 Two views on cognition

According to traditional views in the neurosciences, the perceptual part of the brain is mostly concerned with extracting information from proximal stimulation to obtain accurate representations of the world. This view was not confined to neuroscientists but was also generally accepted amongst philosophers working on perception and the nature of OOs. Consider, for instance, Quine (1995), according to whom we have an innate feature-detection machinery that allows us to identify OOs:

> . . . [T]here is a harbinger of [the positing of objects] already in our *innate* propensity, and that of our animals, to confer salience on those components of a neural intake that transmit corporeal patches of the visual field. *(p. 254, emphasis added)*

These innate feature detectors have, according to the classical view, evolved to accurately detect what there is in the world. Perception is thus seen to largely depend on universally shared mechanisms that detect specific features in the perceptive field. This view was, for instance, reinforced by Hubel and Wiesel in their work on the visual cortex for which they both received the Nobel Prize in 1981. According to Quine, their work shows that the

> . . . overall traits of the scene are selected and communicated as wholes. One bank of brain cells responds *exclusively* to scenes in which there are some conspicuous diagonals from upper right to lower left. Other banks of cells *specialize in other broad traits*. Various traits thus abstracted get superimposed to kick off the appropriate response. (Quine, 1993, p. 114, emphasis added)

A consequence of thinking that the brain possesses an innate feature-detection machinery that responds in a universally shared determinate manner to the features of the stimulus is that perception becomes characterized as a one-directional causal process. It assumes a dominating one-directional causal chain between a physical object, its neural representation, and the utterance made by the speaker. Quine made the following elaboration:

> Causal continuity is the fact of the matter: the causal chain from Mama or the rabbit to utterance of the observation sentence 'Mama' or 'Lo, a rabbit.' Psychologists fix upon one or another point on this causal chain and call it the stimulus, to which the response is conditioned. *(Quine, 1993, p. 113–114)*

Since Quine's writings, the belief that the receptive fields of feature detectors are specified innately has been discarded by many neuroscientists and theoreticians working on machine learning (Hinton, 2007; Karni et al., 1994; Sharma et al., 2000). Perception is not a passive process by which the features of an object are detected by innate specialized modules of encoding. The new perspective is most prominently included in *predictive processing (PP)*, a novel framework from neuroscience research that offers nothing less than a unified theory of cognition. PP is relevant for us since it can be understood as an 'intermediate-level model.' That is to say, it is unspecific about precise neurophysiological details but aims to explain the most general computational methods that are in operation across different brain structures and length scales (Spratling, 2013). It is especially relevant for our purposes as it bridges some of the gaps between understanding the general purpose of individual neural activity and first-person human experiences:

> The PP schema is especially attractive because it deeply illuminates the nesting of the neural economy within the much larger nexus of embodied, world-involving action . . . . PP suggests new ways of making sense of the form and structure of human experience . . . . (Clark, 2016, p. 2–3)[2]

Replacing our traditional understanding of cognition, PP suggests that the brain is hierarchically structured and that higher cortical regions actively generate percepts, trying to *predict* the sensory data in a *robust* way on the basis of prior probabilities and likelihood estimates. This is realized by a hierarchical network with a two-directional information flow where the so-called *recognition network* runs 'bottom-up' from the incoming signals to the higher cortical regions while the *generative network* runs 'top-down' in the opposite direction. This structure allows lower cortical regions to send information about false predictions upstream and allows higher cortical layers to communicate the needed feature-detection adjustments to the lower levels. The features to which neurons respond are not innately determined. Rather, these features are constantly reevaluated and adapted, depending on the current needs for making successful predictions (Clark, 2016; Friston, 2005; Hohwy, 2013; Kersten et al., 2004).[3] Contrary to the classical view, such as the one Quine was relying on above, our visual experiences are shaped by the brain's needs to make robust predictions on the activity of lower regions in the visual cortex. Whenever a prediction is inaccurate, the brain's inner model is updated – it 'learns' and adapts in the form of synaptic and possibly structural plasticity. According to PP, the problem of underdetermination is not solved by an innate hardwired feature-detection machinery but rather by a kind of approximation of Bayesian inference. Prior beliefs are encoded by generative networks and are used to generate percepts that predict the neural activations in lower cortical regions – ultimately the activity of those neurons that is directly caused by the sensory data.

The advantage of a cognitive system that does not rely on an innate feature-detection machinery is that it is fully adaptive to a continuously changing environment in using its resources most effectively. This is because the complexity of the world is far greater than the complexity of the brain such that the brain is forced to compress sensory data to make probabilistic rather than determinate models of the world. Due to its limited resources, the brain cannot keep track of everything that is going on in the world but constructs an internal model of the world that is most *likely* to predict future events. For this, the model must be highly adaptive and cannot rely on innate feature detectors. Constructing relevant feature detectors rather than having them innately engrained is one of the essential techniques employed in PP to build an internal model of the world.

If it were, theoretically, not the case that the world surrounding us had a much greater complexity than our brains—, thus, if it were not the case that there would be constant unforeseen changes in the environment – then the PP strategy of cognizing the world would be far from optimal. For instance, a probabilistic approach to cognition would make little sense in a universe with only a handful of possible states and a simple rule determining the upcoming state if the cognizer had more than enough resources to process these states.[4] Thus, there is an implicit ontological assumption about the world surrounding us toward which PP is adjusted: that the world surrounding the cognizer is much more complex than the cognizing system itself and is constantly changing, making innate feature detectors pointless. In such a highly complex world, the cognizing system by definition has limited resources to grasp the full complexity of the world surrounding it. As a consequence, much of what is going on cannot be comprehended by the cognizing system and is perceived as *noise*. Living in a complex world therefore necessitates cognition to deal with noise, if it ought to be successful.

This is where the significance of *robustness* comes into play. In a complex world, the cognizer is surrounded by noise and must therefore have an internal model of the world that is largely unaffected by this in order to maximize the average prediction error. For a predictive cognitive process to be qualified as making *good* predictions, the process must not only have high accuracy and precision when predictions are made on some generic sensory data. It must also make correct predictions when environmental changes occur or when the sensory data are convoluted with noise due to the cognizer's biological constitution. Such predictive processes are then said to be *robust*.[5] For instance, predictions of neural activity should not be completely different whenever the shapes in the perceptual data are rotated or when luminance changes. In the case of visual perception, the perceptual capacity of being robust results in having *perceptual constancies*. *Shape constancy* is for instance the capacity of seeing the same shape under a variety of different perspectival conditions. *Location constancy* is the capacity to detect the same object at varying distances.[6] Perceptual constancies are to be understood as a consequence of minimizing prediction errors in a robust way. In the PP framework, this robustness is ensured by the hierarchical structure of the network in which higher cortical regions predict the activity of lower regions.

4 Predicting the essential qualities of ordinary objects into the world

In this section, I show how certain structural properties of our visual ontology cannot be derived from the visual sensory data alone but are to some extent the result of our predictive cognitive processes. While there exist numerous ways how our visual data could be interpreted, it will be shown that the discussed structural qualities of OOs can be derived from the category of a *robust predictive process* (RPP) given our ordinary sensory input. The category of RPPs is therefore argued to unify certain concepts within our visual ontology.

The postulated principle governing the visual ontology of a system *A* which has a cognitive system that operates by the principles of PP is the following:

> **Principle of Predicted Structure:** *A*'s visual ontology is structured in a way that robustly maximizes the probability of a minimal average prediction error for future sensory input.[7]

In the following, different principles are presented that are derived from the *Principle of Predicted Structure* and concern a specific aspect of our visual ontology.

4.1. Object individuation by categorization

Whether or not OOs exist in the world is a subject of dispute amongst metaphysicians (see, e.g., French, 2019). Some say that there exist no tables but only particles arranged tablewise, generating the same experience as if tables existed (Merricks, 2001; Unger, 1979; van Inwagen, 1990). While the existence of OOs might be controversial, it is generally accepted that there is nothing within the visual data themselves that determines whether OOs exist.[8] Although this paper takes no position on metaphysical matters, it is without any doubt that we ordinarily have the impression that the material world consists of individual objects possessing certain properties. It is thus a question of why and how this sensation of seeing individuals of a certain kind comes about, given our ordinary sensory data.

Early accounts of perception tend to treat object detection and categorization as two separate issues. Seeing an object *x* of kind *F* is thereby a two-step process of first detecting objects in the visual data and then categorizing them. An early proposal concerning the first step of segmenting objects from the background was, for instance, given by the Gestalt theorist Wertheimer (1923). According to him, there exists a cognitive mechanism that groups together 'elements' of the visual field resulting in perception of discrete, individual objects. The cognitive principles guiding this grouping behavior consist of various factors such as similarity and relative proximity.

As Palmer and Rock (1994) correctly pointed out, Wertheimer falsely presupposed 'elements' in the visual field when formulating grouping principles: "[T]here are no independent 'elements' or 'units' to be grouped; there is simply an unstructured image" (Palmer & Rock, 1994, p. 39). For this reason, Palmer and Rock formulated an alternative grouping principle called the *uniform connectedness principle* according to which a connected region of uniform visual properties is grouped as one perceptual unit. A computational realization of the uniform connectedness principle is said to be feasible by first locating the boundary of an object as described by Hubel and Wiesel (1968) and then distinguishing the bounded domain from the background, resulting in the individuation of the object that can later be categorized.

Palmer and Rock admitted that the uniform connectedness principle has difficulties in explaining how disconnected regions in the visual field can belong to one and the same object if the object is partially covered by another object. The principle also has difficulties explaining how similar surfaces can still be distinguished as in the case of animals with camouflage.

PP resolves these problems by challenging the assumption that individuation and categorization occur separately. Instead, PP relies on the idea that the features by which objects are individuated and distinguished from the background (as determined by the recognition network) are defined by their categoric membership (as determined by the generative network). This seems paradoxical: One would expect that in order to learn a category it is necessary that one has already detected objects having similar features. Thus, how can categories be taken as a criterion for object detection? Hinton (2007) discovered an algorithm that makes it possible to simultaneously detect and categorize an object by treating individuation and categorization as a single process of tiny, incremental representational improvements in a continuous space. This algorithm was inspirational for PP and overcame the 'chicken-and-egg problem: Given the generative weights we can learn the recognition weights and given the recognition weights we can learn the generative weights' (p. 5).

In the case of the brain's cortex, 'predictor neurons code information about object category; error neurons signal mismatches in predicted and observed object category' (Koster-Hale & Saxe, 2013, p. 838). Higher cortical regions encode models of the environment that generate hypotheses in the form of perceptions that try to predict the structural characteristics of the world such as categoric membership. For instance, given a certain stimulus, the cognitive system computes the category *F* that it finds most likely to be present in the stimulus. Conceptually similar to Hinton's algorithm, when the brain selects the category *F* for the purpose of minimizing prediction error, it also determines the object and its boundary in the stimulus.

This can be generalized to any cognizing system with finite resources that has the objective of minimizing prediction error. Such a system must have some representation that encodes information about categoric membership because sensory data are only predictable when they are considered as an instance of a category that generalizes certain features in the data for compression purposes.[9] A system can roughly be said to *impose* the mental category *F* it has learned from earlier experiences onto images. Imposing a category onto an image successfully (i.e., with little prediction error) allows the system to individuate objects in the visual field, for instance, by computing for all parts in the image the probabilities for belonging to an instance of a certain mental category.[10] Not imposing or being able to impose any category leads to the maximum prediction error since no visual object is individuated for the purpose of tracking it and predicting its behavior. For a cognizer *A* to minimize prediction errors, it is essential to individuate objects in the visual image provided over the visual system:

> **Principle of Individuation:** A region in the visual image is individuated and tracked as an object *x* if this robustly maximizes the probability of minimizing *A*'s average prediction error.

The *visual image* refers to the information that is received at the retina and structured in the form of a 2-D grid corresponding to the physical arrangement of the photoreceptors. From the retina, this information is passed forward to the LGN and the visual cortex.

Figure 1 illustrates how the theoretical insights of PP about perception exemplify themselves at the level of our conscious experiences. Figure 1a shows an image that is at first glance unrecognizable, but eventually the Dalmatian dog can be seen. Once the dog is detected, it is typically impossible to experience the image the same way as before – as an assemblage of unorganized black spots. According to PP, the category of the dog is imposed onto the image because this is the assumption with the highest probability of resulting in small prediction errors when tracking the spots in the image over time. Without already possessing the right category, it is impossible to detect the dog in the image, and methods like those proposed by Wertheimer or Palmer and Rock would clearly fail.

(a) (b)

Figure 1. Image (a) depicts a Dalmatian dog and was taken from Gregory (1970). Image (b) is reproduced from Clark (2015).

How prior perceptions affect the outcome of the category imposed onto the sensory data can be illustrated by looking at Figure 1b. When our eyes move from top to bottom, we first see the letter *A* followed by the letter *B* since the latter is the most probable category to be displayed after the former, given the sensory input. For analogous reasons, when we move our eyes from left to right, we see the shape in the center as the number *13*. Whatever category is imposed, corresponding regions of the visual field are then treated as composing to a single object.

To determine the category of an object in the visual field one might naively believe that the brain computes "'comparisons' of a presented [visual] patch with some sort of standard or prototype in memory" (Raffman, 1994, p. 48) in order to evaluate the category *F* of an object *x*. Thus, if the perceived object *x* is similar to an image of type *F* or a set of well-defined, static features associated with *F*, then *x* is also of type *F*. This intuition reflects itself in early computational models of object recognition where a large number of categorized images are stored in memory and compared to the sensory input image (see, e.g., Abu-Mostafa & Psaltis, 1987; Hopfield, 1982; Huberman & Hogg, 1984; Kohonen, 1978; Willshaw et al., 1969). Many theorists of perception have rejected the idea of the brain storing a large number of picture-like representations in order to determine an object's category (see, e.g., Biederman, 1987; Ullman, 1989). As I discuss above, the PP framework opposes this naive view by suggesting that minimizing prediction error is what determines categorization and with this the individuation of visual objects. Not only does PP reject the idea of the brain comparing features in the sensory input to some pictorial representation in the brain, it also suggests that there is no comparison to statically, predetermined features. Rather, the features against which the sensory input is evaluated are always evolving and changing.

This does not mean that according to PP the cognitive system does not make any sort of comparison. When the cognizing system generates its predictions of the input, it relies on the generative network that models the joint distribution over both the input and the approximated 'hidden causes,' which allows it to generate instances of predicted input data. The cognitive system learns and adapts according to the similarity between this joint distribution and the distribution over causes given the input data expressed by the recognition model (Kiefer & Hohwy, 2020).[11] In contrast to the naive view on cognition that imagines some comparison between external features and fixed, possibly innate features stored in the brain, the generative network encodes for changing structures. The brain makes a comparison between the sensory data and the self-generated predicting image that is context dependent and not static but under constant change in accordance with the always-evolving features defining the respective category. When the self-generated image is a bad prediction, the neural network model adapts in order to achieve better prediction results in the future.

4.2. Object persistence

It is clear from introspection that, within our visual ontology, objects persist over time and space, and we expect them not to disappear from one moment to another without any reason.[12] This means that the corresponding parts in the visual field do not become a new object from one moment to another but persist to be the same object. In Figure 2, an example of the visual image over a short amount of time is presented. Due to the underdetermination of perception there is no empirical fact of the matter that necessitates the connection between the ball in one image with the ball in the next image, since our senses receive sequences of sensory data that do not determine whether the objects at one moment in time are the same as those in the next. Yet, we clearly have the impression that the ball remains the same ball – the same object.

The *Principle of Individuation* explains how an object x comes about in our visual ontology. Assuming that we have at least one such object, then, according to the postulated *Principle of Predicted Structure*, the following principle must hold for any predicting cognitive system A, since persistence is a structural property:

> **Principle of Object Persistence:** An object x that is located at the spatial-temporal position (r_1, t_1) continues to be perceived by A as the same object at (r_2, t_2), if this robustly maximizes the probability of minimizing A's average prediction error.

The important question is whether the *Principle of Object Persistence* correctly captures our intuitions concerning object persistence, namely that OOs generally persist through space and time, given the visual images we receive over our retina. Thus, with regard to Figure 2, the question is whether the *Principle of Object Persistence* provides reasons as to why we see the ball at one point in time to be the same ball at the next point in time.

Giving an affirmative answer to this question is almost trivial. When one considers the opposite scenario of a cognizer that experiences OOs as constantly coming into and going out of existence, the cognizer would be unable to make any useful predictions about upcoming sensory inputs. In the case of the ball in Figure 2, after times t_1 and t_2, the best prediction that minimizes prediction error at t_3 is *not* that the ball at t_1 and t_2 suddenly

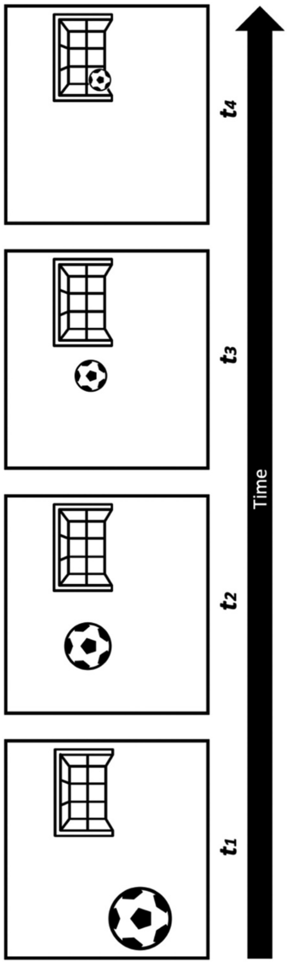

Figure 2. Objects are experienced as persisting through time and space even though the sensory data do not necessitate the connections between the objects at different points in time.

disappears and a new ball appears at t_3. Instead, a trajectory of what the cognitive system identifies as a single object is computed and predictions are made on where the ball lands at t_3. Therefore, according to the interpretation of PP presented here, OOs persist within our visual ontology whenever the visual data allow the brain to make predictions.

4.3. Parthood and compositionality

The different aspects concerning parthood and compositionality extend the case of object persistence. The central question that we are concerned with here regarding our visual ontology is how we account for the fact that some parts of our visual image compose to one unified object while some parts do not compose. For instance, how can it be that my arm composes an object with the rest of my body, but my nose and the Eiffel Tower do not compose?[13]

Within the metaphysics literature, van Inwagen (1990) was the first to ask under what conditions two objects, x and y, jointly compose an object, z, and called this the 'special composition question' (SCQ). Considering several initially promising answers to the SCQ, van Inwagen showed that none results in plausible implications. For instance, one plausible answer to the SCQ that van Inwagen considered was *contact*. So, one might suggest that any contact between x and y will result in the new composite object z. Drawing evidence from ordinary experiences, he concluded that if contact were to be a criterion of composition, then this would lead to very undesirable consequences, such as that whenever two people shake hands, they briefly constitute a single object. He argued that there seems to be a tension between the prospect of a universal answer to the SCQ and the adoption of an ontology that is in line with our ordinary beliefs about, and experiences with, OOs.

Clearly, being a metaphysician, van Inwagen was seeking something different than what we are interested in here. He sought an account for compositionality for a 'mind-independent, discourse-independent world' (Horgan & van Inwagen, 1993, p. 695). Our interest here is the visual ontology of humans and not what metaphysicians have in mind when talking about compositionality concerning a mind-independent world. Nevertheless, famous puzzles concerning compositionality in metaphysics are also puzzles for any account of compositionality within the visual ontology for given sensory data. This might be the case because these puzzles are only mistakenly conceived as metaphysical puzzles rather than puzzles concerning our visual ontology; however, this paper does not take sides on this subject.

One such famous puzzle is Unger's (1980) 'problem of the many.' According to this problem, a multiplicity of possible subsets of an object's basic composing elements seem to constitute a new object, leading to a population explosion. Take, for example, a cloud consisting of *N* droplets. Then it seems that any subset of *N*-1 droplets also forms a cloud which is different to the one consisting of *N* droplets. A complete account of the visual ontology must also provide an answer to this problem, even if visual ontology is not the concern of metaphysicians.

Sider (2001) introduced the technical term of *maximality* that is useful for addressing this problem. The property of maximality is a second-order property that can be attributed to a property, *F*, such as *being a cloud*. F is maximal if and only if large proper parts of an *F* are not themselves *F*s; for instance, a part of a cloud is not itself a cloud. What I intend to illustrate now is that the *Principle of Predicted Structure* implies maximality when the cognizer is exposed to our ordinary sensory data. It is for this reason that our brain avoids Unger's problem and that we have the experience of only seeing a small number of clouds rather than millions of clouds when looking into the sky.

Notice that the *Principle of Predicted Structure* implies the following corollary about the structure of our visual ontology:

> **Principle of Compositionality:** The visual objects *x* and *y* compose to *z* if *z* is expected to be on average easier to predict than *x* and *y* independently.

Thus, categories are selected and imposed onto the data such that regions in the visual image turn into individuated objects we perceive as single standing objects. The important thing to notice about this principle is that it entails Sider's maximality property for the objects of our visual ontology when we assume that sensory data are noisy and come from a complex, changing environment as presupposed by PP (see end of Sec. 3).

To see this, consider Figure 3. A black square is depicted which is moving around in front of a cat, thereby only allowing the cognizer to garner a small number of visual cues about there being a cat in the data. Such a lack of information and noise in the data is commonplace in ordinary, everyday perception. Every time we walk around or move our eyes, there is a 'dramatic change of sensation' (Frith, 2007, p. 201) and yet the world appears to be stable.

Let us first assume a cognizer with a cognition process that assigns categories merely by making a comparison of the visual image with some static features or prototypical image in memory as in the naive approach to cognition discussed above. In this case, at time t_2, the object in the image could just as well be interpreted to be a dog because at that moment in time

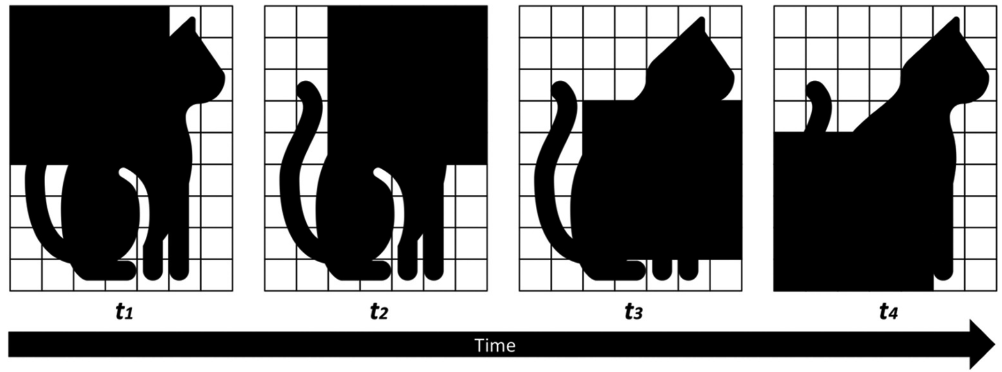

Figure 3. The cognizing system must select the category that is expected to be the most robust against noise over time in order to minimize prediction error.

the relevant information that makes a distinction between dogs and cats (the head) is unavailable, and the body of the animal might just as well be a dog's body. (For simplicity, I am assuming that the sole relevant visual characteristic distinguishing the animals is given by the shape of the head.) This type of cognizing system will have a visual experience of seeing a cat, then a dog, and then a cat again. Or in general, the system will have an unstable visual experience of a world in which categories as well as the boundaries of objects alter from one moment to the next.

In contrast, a cognizing system that has experiences by minimizing prediction error will experience a visual ontology of stable rather than abruptly altering objects and object kinds. It will impose the category onto the data that is the most robust toward possible future noise in the sensory data because this makes the data predictable. The cognizer's category would, for instance, not change at time t_2 to 'dog' because this would not be a robust prediction over time. Any categoric change leads to a high prediction error since a category is needed in the first place to make a prediction about an upcoming sensory event.

This so far extends to what we have already seen when we were looking at object persistence, but now the important point is that our findings generalize to the parts of the object. To see this, let us again first assume a cognizer with a cognition process that assigns categories by comparing the visual image with some prototype in memory. This cognizer is confronted with the sequence of images in Figure 3. As we have seen, the object in the image could, at time t_2, just as well be interpreted to be a dog. But furthermore, some small parts – for example, the tip of the cat's claw – might not be well in line with the prototype of a cat and would thus be treated as separate objects. Or the whole region of the cat minus some small part *A* could look equally alike to the prototype as the cat minus some small part *B*. Therefore, the cognizer might interpret there to be two cats at hand or randomly select one possible instance over the other.

Furthermore, imagine that the cognizing system were, for instance, not to select the category of the cat as a whole at time t_1 but, instead, were to impose the categories of cat-head, cat-body, and cat-tail separately. In this case there would be a great amount of prediction error at time t_2, when the square overlaps the head, as the activated category of cat-head would have to be turned off completely (including any further encoding for the spatial connection between the cat's head and all its other body parts). At t_3, the representation of the head would have to be reactivated, generating further prediction error due to the sudden reappearance of the object. If, on the other hand, the cognizing system were to impose the category of a whole cat at t_1 and maintain this categorical assumption for as long as it is a reasonable option, then no great prediction error would be expected at t_2 or t_3, making this the category of choice. A selected category that is not robust against such noise is not reliable to minimize prediction error, because in daily perception we often only receive a glimpse of an object.

Contrary to a cognizer that selects categories by naive comparisons, a cognizer operating by making robust predictions will always see the maximal number of jointly predictable parts in the visual field to compose as one object because this increases the chances of making good predictions from prior experiences. Such a cognizer will look into the sky and see one cloud rather than each part of a cloud forming an additional cloud as considered in the problem of the many. In contrast, a naive categorization by similarity analysis leads to a multitude of possible categories where different parts of the visual image might match the category well and the boundary of the object would not be well determined.

To be sure, a cognizer's perceptive system that categorizes to make robust predictions will on the one hand see objects as composing wholes (e.g., one cloud rather than millions of clouds) but will nonetheless not see objects composing arbitrarily. My nose and the Eiffel Tower do not compose because the mental category needed to individuate such an object does not lead to robust predictions. There is no connection between my nose and the Eiffel Tower that would minimize prediction error, were they to compose to a single object.

4.4 Collocation

From the *Principle of Individuation*, the following must hold: One and the same region in the visual field is simultaneously individuated as two different objects x and y if this robustly maximizes the probability of minimizing A's average prediction error. Clearly, we humans never experience objects to be precisely collocated. This becomes most apparent when we think of the famous duck – rabbit illusion (Jastrow, 1899) in which we are unable to see both a duck and a rabbit at the same time at the same location.

The fact that visual objects are never seen to precisely coincide does not follow trivially from the above-mentioned principle for our sensory input. As we have seen in Section 3, for a PP approach to operate successfully, the ontological assumption must be made that the environment is complex and in flux. It is sometimes suggested that additional ontological factors determine the fact that we do not see OOs to coincide. For instance, in the context of explaining *binocular rivalry*,[14] which arguably also applies to the duck – rabbit illusion, Hohwy et al. (2008) suggested that our generative top-down model has learned that 'only one object can exist in the same place at the same time' (p. 691). According to them, our experiences of the world shape our brain's prior expectations to the extent that, by Bayesian inference, the brain computes a much lower probability for the image to be both a duck *and* a rabbit rather than a duck *or* a rabbit.[15]

The problem with Hohwy et al.'s analysis is that they only maximized probabilities for a given moment in time that is case specific rather than general and thereby failed to account for general robustness over time. In accordance with PP, the cognizing system *A* has no innate categories of perception but is, rather, constantly in the process of constructing these categories to account for the many small changes of the environment. As we have seen in the case of compositionality and persistence, if *A* were to alter its hypothesis every time some other hypothesis seems more probable, it would have an extremely high prediction error in a changing and noisy environment. It would look at a cloud and suddenly see a huge horse appear, and then in the next moment a huge unicorn due to constant formation changes of the cloud. This is what any system based on PP tries to avoid, and, given that categories are not innate or predetermined, any indecisiveness where the cognizing system begins to perceive a multiplicity of possible objects for the same visual input generates complete instability of the system.

Minimizing prediction error in the long run comes at the cost of not being considerate about alternative possibilities in the moment. For this reason, the brain must occasionally disregard hypotheses even if they have a very high probability of being the optimal choice for the specific point in time. The first consequence of this is that the following principle about collocation must hold:

> **Principle of Object Collocation:** One and the same region in the visual image is never simultaneously individuated as two different objects.

The second consequence is that having a discomforting visual experience like the duck – rabbit illusion is 'something any organism with a predictive engine will to an extent experience' (Kaaronen, 2018, p. 11).

It might seem that our not seeing OOs coincide with each other results from the physical fact that the subatomic particles constituting OOs also do not coincide. Therefore, one might try to argue that OOs do not coincide in our experience because of how the world really is rather than being a consequence of how our cognition works. However, this again misses the point that the empirical data are underdetermined and there is a multiplicity of possibilities to interpret the world. For instance, all empirical data in physics are also compatible with the assumption that each particle in the world precisely coincides with a massless 'dummy' particle that does not causally interact with anything in the world. The data might also be compatible with each particle in the world having half its mass but being always collocated with another particle of equal mass. In any case, our cognition decides on one of many possibilities, and our experience of OOs as necessarily non-collocated can be derived from how cognition works rather than from the empirical data.

5 Reformulating questions about ordinary objects and conclusion

The analysis shows that any system that receives our ordinary visual input and experiences the world through RPPs will perceive the world to be structured in the following way: it consists of individuals that tend to persist over space and time, that are not collocated with each other, and for which maximality is guaranteed despite compositionality not happening arbitrarily. Problems such as the problem of the many are, within our visual ontology, an immediate consequence of a simplistic view on cognition that relies on the intuitive assumption that categories must be applied to objects based on the similarity in the arrangement of their parts. But this is not how the brain subconsciously categorizes objects.

We notice that the notion of an RPP allows us to unite several aspects of what it means to be an OO under one umbrella term. This provides a more unified understanding of the object concept which is to be preferred over other accounts of our visual ontology that depend on numerous, independent principles (cf. Green, 2019, 2021; Skrzypulec, 2016). Questions concerning our visual ontology that are ordinarily formulated in terms of standard metaphysical concepts can now be formulated in terms of RPPs by taking a subjectivist position and moving from metaphysics to epistemology or perception. For instance, instead of asking, 'Do x and y compose?' we can ask, 'Are x and y jointly easier to predict on average than independently?' Or instead of raising the question, 'Does x persist through space and time?' we might ask, 'Does the individuation of x also lead to good average predictions if x is located at some other location in space and time?' Or 'Does x survive the loss of its part y?' can be replaced with 'Does x without y still lead to small average prediction errors?' While the traditional questions have no straightforward answers and seem to be not directly related or at least hard to connect, the new formulation relies on the unifying category of RPPs and offers a straightforward answer. This is not to say that no implicit ontological assumptions were presupposed in our analysis, since we presupposed the visual input we already ordinarily have. However, due to underdetermination, certain cognitive import is required to give structure to our visual ontology. The concept of RPPs unifies the missing structural aspects and also allows us to explain in a unified manner how our intuitions about our visual ontology come about, given our ordinary sensory input.

Notes

1. Some of the qualities we associate with OOs are not only underdetermined by perception but even *inconsistent* with physical reality (e.g., the inseparability of individuals due to quantum entanglement) (Schwaninger, 2019).
2. Notice that this is not an uncontroversial issue. See, e.g., the exchange between Hohwy and Seth (2020) and Schlicht and Dolega (2021).
3. Note that the notion of *prediction* is a technical notion and refers to mechanisms we are not consciously aware of.
4. For simplicity, the cognizer is implicitly assumed to not be part of the world.
5. In the machine learning literature, the lack of robustness is characterized by and related to the concept of "overfitting." A learning algorithm that tries to take too many possible instances of a hypothesis class into account does not generalize the data well but is instead overfitting the data, resulting in a lack of robustness toward new data.
6. See Burge (2014) for a discussion on how perceptual constancies relate to the mind.
7. In a noisy world, robustness is in the generic case already implicitly entailed in minimizing the average prediction error. However, for the purpose of emphasis and generality, it was made explicit.
8. Hofweber (2019) is an exception amongst metaphysicians in these regards as he does claim that existence questions can be answered empirically.
9. Despite us having the ordinary impression of there being a qualitative difference between *concepts* as general mental categories and *percepts*, this difference is only gradual in PP: "Percepts are … basically shorter-term expectations and concepts longer-term expectations" (Hohwy, 2013, p. 72).
10. As Rescher correctly points out, concepts or general mental categories are required for individuation and identification: "The fact is that ideation is an indispensable preliminary and requisite for individuation …. To be an identifiable item, that is to say, is to answer to the idea of a certain sort: identification is not possible without sortification and sortification is not possible without recourse the corresponding ideas" (Rescher, 2016, p. 10). Although some scholars avoid the use of the term "category" in favor of "representation" or "hypothesis," I maintain "category" as it reminds us of the bridging character of PP between the personal and sub-personal level, and it is also in line with some of the neo-Kantian literature that closely relates to Gestalt psychology.
11. This similarity between distributions can be quantified in terms of the Kullback-Leibler divergence (Cover & Thomas, 2006, ch. 2.3).
12. Studies show that infants already expect objects to move along continuous trajectories and are surprised if they do not do so and disappear (Spelke et al., 1995).
13. Note that the following properties under investigation are similar to those of a system that operates by optimally compressing sensory data (Petersen, 2019). Merely taking data compression into account does not, however, succeed in recreating all the structural principles of our visual ontology. This is because prior experiences are not considered when only taking the notion of compression rather than prediction into account. The account presented here consequently suggests that a sensory input might deliver different objects at one point in time than in another. For instance,

objects that were not composing at time t_1 might be composing at another time t_2, depending on the experiences the cognizer has made previously. This is exemplified by Figure 1b, where the sensory input remains constant but our experience varies, depending on what we had previously experienced.

14. This is the phenomenon that, when the left and the right eye are presented with a different image, the subject alternately sees only one image at a time rather than two.
15. Notice that I do not equate "seeing a duck and a rabbit at the same time" with "seeing a duck-rabbit" since this might be misinterpreted to refer to a single object, a duck-rabbit, rather than two coinciding objects.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the Forschungskredit of the University of Zurich under GrantFK-19-065.

Notes on contributor

Arthur C. Schwaninger is currently a researcher at the University of Zurich. He studied Philosophy of Science at the University of Cambridge, Computational Science at ETH Zurich, and worked in the field of Computational Neuroscience at the Human Brain Project. His research interests include Theoretical Neuroscience, Deep Learning, Kantianism, and the metaphysics of ordinary objects.

ORCID

Arthur C. Schwaninger http://orcid.org/0000-0002-0466-2437

References

Abu-Mostafa, Y. S., & Psaltis, D. (1987). Optical neural computers. *Scientific American*, *256* (3), 88–95. https://doi.org/10.1038/scientificamerican0387-88

Biederman, I. (1987). Recognition-By-Components: A theory of human image understanding. *Psychological Review*, *94*(2), 115–147. https://doi.org/10.1037/0033-295X.94.2.115

Burge, T. (2014). Perception: Where mind begins. *Philosophy*, *89*(3), 385–403. https://doi.org/10.1017/S003181911400014X

Churchland, P. S., & Sejnowski, T. J. (1988). Perspectives on cognitive neuroscience. *Science*, *242*(4879), 741–745. https://doi.org/10.1126/science.3055294

Clark, A. (2015). Radical Predictive Processing: Radical Predictive Processing. *The Southern Journal of Philosophy*, *53*, 3–27. https://doi.org/10.1111/sjp.12120

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

Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory* (2nd ed.). John Wiley & Sons.

French, S. (2019). There are no such things as ordinary objects. In J. Cumpa & B. Brewer (Eds.), *The nature of ordinary objects* (pp. 260–273). Cambridge University Press.

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1456), 815–836. https://doi.org/10.1098/rstb.2005.1622

Frith, C. (2007). *Making up the mind: How the brain creates our mental world*. Blackwell Publishing.

Green, E. J. (2019). A theory of perceptual objects. *Philosophy and Phenomenological Research*, *99*(3), 663–693. https://doi.org/10.1111/phpr.12521

Green, E. J. (2021). Binding and differentiation in multisensory object perception. *Synthese*, *198*(5), 4457–4491. https://doi.org/10.1007/s11229-019-02351-1

Gregory, R. L. (1970). *The intelligent eye*. Weidenfeld and Nicolson.

Hinton, G. (2007). To recognize shapes, first learn to generate images. *Progress in Brain Research*, *165*, 535–547.

Hoffman, D. D. (2019). *The case against reality: How evolution hid the truth from our eyes*. W.W. Norton & Co.

Hofweber, T. (2019). Empirical evidence and the metaphysics of ordinary objects. In J. Cumpa & B. Brewer (Eds.), *The nature of ordinary objects* (pp. 27–47). Cambridge University Press.

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

Hohwy, J., Roepstorff, A., & Friston, K. (2008). Predictive coding explains binocular rivalry: An epistemological review. *Cognition*, *108*(3), 687–701. https://doi.org/10.1016/j.cognition.2008.05.010

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

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, *79*(8), 2554–2558. https://doi.org/10.1073/pnas.79.8.2554

Horgan, T., & van Inwagen, P. (1993). On what there isn't. *Philosophy and Phenomenological Research*, *53*(3), 693–700. https://doi.org/10.2307/2108094

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, *160*(1), 106–154. https://doi.org/10.1113/jphysiol.1962.sp006837

Hubel, D. H., & Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology*, *195*(1), 215–243. https://doi.org/10.1113/jphysiol.1968.sp008455

Huberman, B., & Hogg, T. (1984). Adaptation and self-repair in parallel computing structures. *Physical Review Letters*, *52*(12), 1048–1051. https://doi.org/10.1103/PhysRevLett.52.1048

Jäkel, F., Singh, M., Wichmann, F., & Herzog, M. H. (2016). An overview of quantitative approaches in Gestalt perception. *Vision Research*, *126*, 3–8. https://doi.org/10.1016/j.visres.2016.06.004

Jastrow, J. (1899). The mind's eye. *Popular Science Monthly*, *54*, 299–312.

Kaaronen, R. O. (2018). A theory of predictive dissonance: Predictive processing presents a new take on cognitive dissonance. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.02218

Karni, A., Tanne, D., Rubenstein, B. S., Askenasy, J. J., & Sagi, D. (1994). Dependence on REM sleep of overnight improvement of a perceptual skill. *Science*, *265*(5172), 679–682. https://doi.org/10.1126/science.8036518

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as bayesian inference. *Annual Review of Psychology*, *55*(1), 271–304. https://doi.org/10.1146/annurev.psych.55.090902.142005

Kiefer, A., & Hohwy, J. (2020). Representation in the prediction error minimization framework. In S. Robins, J. Symons, & P. Calvo (Eds.), *The Routledge companion to philosophy of psychology* (2nd ed., pp. 384–409). Routledge, Taylor & Francis Group.

Kohonen, T. (1978). *Associative memory: A system theoretic approach*. Springer-Verlag.

Koster-Hale, J., & Saxe, R. (2013). Theory of mind: A neural prediction problem. *Neuron*, *79*(5), 836–848. https://doi.org/10.1016/j.neuron.2013.08.020

Merricks, T. (2001). *Objects and persons*. Oxford University Press.

O'Callaghan, C. (2016). Objects for multisensory perception. *Philosophical Studies*, *173*(5), 1269–1289. https://doi.org/10.1007/s11098-015-0545-7

Palmer, S., & Rock, I. (1994). Rethinking perceptual organization: The role of uniform connectedness. *Psychonomic Bulletin & Review*, *1*(1), 29–55. https://doi.org/10.3758/BF03200760

Petersen, S. (2019). Composition as pattern. *Philosophical Studies*, *176*(5), 1119–1139. https://doi.org/10.1007/s11098-018-1050-6

Quine, W. V. O. (1993). In praise of observation sentences. *The Journal of Philosophy*, *90*(3), 107–116. https://doi.org/10.2307/2940954

Quine, W. V. O. (1995). Naturalism; Or, living within one's means. *Dialectica*, *49*(2–4), 251–263. https://doi.org/10.1111/j.1746-8361.1995.tb00164.x

Raffman, D. (1994). Vagueness without paradox. *The Philosophical Review*, *103*(1), 35. https://doi.org/10.2307/2185872

Rescher, N. (2016). On kinds of things and cognitive idealism. *Idealistic Studies*, *46*(1), 1–17. https://doi.org/10.5840/idstudies201751554

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

Schwaninger, A. C. (2019). What simulations teach us about ordinary objects. *Open Philosophy*, *2*(1), 614–628. https://doi.org/10.1515/opphil-2019-0045

Shapiro, L. A. (2011). *Embodied Cognition*. Routledge.

Sharma, J., Angelucci, A., & Sur, M. (2000). Induction of visual orientation modules in auditory cortex. *Nature*, *404*(6780), 841–847. https://doi.org/10.1038/35009043

Sider, T. (2001). Maximality and intrinsic properties. *Philosophy and Phenomenological Research*, *63*(2), 357–364. https://doi.org/10.1111/j.1933-1592.2001.tb00109.x

Skrzypulec, B. (2016). Ontology of early visual content. *Philosophical Psychology*, *29*(2), 261–276. https://doi.org/10.1080/09515089.2015.1070259

Skrzypulec, B. (2018). Thisness and visual objects. *Review of Philosophy and Psychology*, *9*(1), 17–32. https://doi.org/10.1007/s13164-017-0337-4

Spelke, E. S., Vishton, P., & von Hofsten, C. (1995). Object perception, object-directed action, and physical knowledge in infancy. In M. S. Gazzaniga (Ed.), *The cognitive neurosciences* (pp. 165–179). MIT Press.

Spratling, M. W. (2013). Image segmentation using a sparse coding model of cortical area V1. *IEEE Transactions on Image Processing*, *22*(4), 1631–1643. https://doi.org/10.1109/TIP.2012.2235850

Ullman, S. (1989). Aligning pictorial descriptions: An approach to object recognition. *Cognition*, *32*(3), 193–254. https://doi.org/10.1016/0010-0277(89)90036-X

Unger, P. (1979). There are no ordinary things. *Synthese*, *41*(2), 117–154. https://doi.org/10.1007/BF00869568

Unger, P. (1980). The problem of the many. *Midwest Studies in Philosophy*, *5*(1), 411–468. https://doi.org/10.1111/j.1475-4975.1980.tb00416.x

van Inwagen, P. (1990). *Material Beings*. Cornell.

Varela, F., Thompson, E., & Rosch, E. (1991). *The embodied mind: cognitive science and human experience*. MIT Press.

Wertheimer, M. (1923). Untersuchungen zur Lehre von der Gestalt: II. *Psychologische Forschung*, *4*(1), 301–350. https://doi.org/10.1007/BF00410640

Wiese, W., & Metzinger, T. (2017). Vanilla PP for philosophers: A primer on predictive processing. In T. Metzinger & W. Wiese (Eds.), *Philosophy and predictive processing* (pp. 1–18). MIND Group.

Willshaw, D. J., Buneman, O. P., & Longuet-Higgins, H. C. (1969). Non-Holographic associative memory. *Nature*, *222*(5197), 960–962. https://doi.org/10.1038/222960a0

Wolfram, S. (2002). *A new kind of science*. Wolfram Media.