Book Reviews

Clark, A. 2016. *Surfing Uncertainty. Prediction, Action and the Embodied Mind*. Oxford / New York: Oxford University Press. 424 pp. ISBN 978-0-19-021701-3

It was the formal model of the Turing Machine that inspired 20th century cognitive science and philosophy to a large extent. The vision of the brain that depicts it as large network machine, related to recent successes of deep learning neural networks and other elaborated forms of machine learning, could take a similar role in the 21th century. At least, this is a strong impression which a reader of Andy Clark's latest book may get. In it, Clark, who is also known for his defence of *the extended mind hypothesis* (Clark & Chalmers, 1998), presents us a broad picture of a unifying theory of how the brain (and the mind) works. Central to Clark's book is the thesis of predictive processing (PP) that the brain can be described as a multi-layer-hierarchical generative neural network that constantly predicts the incoming streams of sensory signals and learns from the resulting prediction errors. Building his arguments on critically and carefully assessed evidence from a large stock of empirical studies, Clark suggests that this single model can, in an integrative way, explain perception, learning, awareness and action and a many other sorts of related mental phenomena, like, for example, imagining, emotions, social cognition, illusions and mental disorders as autism and schizophrenia. Importantly, he also claims that the proposed model of brain functioning combines quite well with the core ideas of embodiment and enactivism.

The book is divided in three parts, each having a main topic, which is supplemented by previews and prearrangements of topics covered more extensively in a later chapter of the book. This allows Clark to slowly unfold the complex, interwoven topics of his book, making their golden threads easily accessible, even to outsiders or newcomers to the field of cognitive science. Part I of the book is largely devoted to introducing the reader to the basic prediction processing model of perception and attention, which in part II is extended to include action (motor command and control) and then, shown to facilitate explanations of diverse mental phenomena. In the final part III, Clark goes on to reflect about the predictive processing approach to the brain from a more general perspective discussing 'productive laziness', the relation between embodiment/enactivism and predictive processing, doubts about the adequacy of the model (raised by the Dark Room Puzzle and novelty-seeking), and the role of representations.

The book is well written, as the author uses clear, precise and easy accessible language, mostly devoid of technical terms. Many notions, not part of the regular philosophical vocabulary, are explained when they are introduced. The style is entertaining and creative, for section headings the author has often selected clever, catchy phrases, innuendos, sometimes even a clever pun. At some points, Clark may seem to repeat his main ideas, somewhat like an ongoing mantra. But perhaps it helps the reader to not lose sight of the big picture that is hidden in the mosaic of fascinating details Clark writes about.

The general picture Clark suggests may appear unlikely true at first: Isn't mental life way too complex to be explained by such a simple model that identifies sub-personal prediction-minimization as the primary function of the brain? How likely is it that the evolution of the brain, which is shaped by historical contingencies, culminates in a coherent system exhibiting a single primary function, instead of a patchwork of various functions? But during the course of the book,

the reader may be surprised about how much light a predictive brain might shed on the mind.

In chapter 1 the author introduces the reader to the predictive processing model of perception, contrasting it with traditional approaches of cognitive science that "depict perception as a cumulative process of 'bottom-up' feature detection" (13). Referring to Hohwy et al. (2008), a first illustration of PP's explanatory power is presented in the case of binocular rivalry (33-37).

According to PP, a top-down and lateral flow of neural signals aims at predicting the current sensory signals, while attempting to minimize the resulting prediction error. Those predictions errors that still occur are propagated both upwards and laterally through the network (25), which Clark summarizes by the slogan: *signalling the news*. In reaction to the feed forward processing of the prediction error the network is updated. This "functional asymmetry" (31) in information processing, which is central to PP, initiates, in Clark's words, "an energetic dance between multiple top-down and bottom-up signals" (14). When the system succeeds to predict the incoming sensory stream, by *generating the signal for itself*, a perceptual experience is formed (14). These ideas of prediction-driven learning in a hierarchical multi-layer network are first briefly outlined and accompanied by a short overview on historical advances in machine learning, beginning with the 'Helmholtz machine', and finally explained in most detail on page 31, where Clark describes the implementation of a predictive processing network by Rao and Ballard (1999).

The mathematical details of the PP model are largely left out. This has the clear advantage that the book remains accessible for readers without sufficient background in mathematics. However, at some points in the book, the missing mathematics seems to leave some questions open (e.g. in chapter 7). What should be addressed by models of perception in general, of course, is the problem of how an organism can perceive its environment, the world, although it has only direct access to the stimulation patterns of its sensory receptors. Clark mentions this problem in passing, in section 1.2., where he reacts to a chicken-egg worry about how prediction is possible without knowledge:

> How does all that knowledge – the knowledge that powers the predictions that underlie perception [...] – arise in the first place? Surely we have to perceptually experience the world *before* we can acquire the knowledge to make predictions about it? (Clark, 2016, p. 14)

As many other optimists about neural networks who are impressed by their recent empirical successes, he claims that predictive-driven learning provides a "powerful way to make progress under such initially unpromising conditions" (17) since when you are "able to detect only the ongoing changes in your own sensory registers [...] [o]ne thing you can do [...] is busily to try to predict the next state of those very registers". Yet, what about the doubts about connectionist models as have been expressed by proponents of the classical approach (knowledge bases) to learning and cognition (e.g. Fodor & Pylyshyn 1988, Chomsky 1980)? As Clark seems to believe, the main problems about connectionist models had been the dependency on pre-categorized data and the distribution of error in the network which have been solved (according to Clark). Awareness of the need of connectionist explanations of the compositional structure of thought are largely missing in Clark's picture, as well as critical reflections on the sheer amount of data neural networks need to be trained on in order to learn (and how this relates to Chomsky's poverty of stimulus argument).

An immediate consequence of the PP account of perception is that the same neural network is taken to explain both perception and (abstract) learning.

In chapter 2 the mechanism of precision estimation and prediction error weighing is introduced: Additionally to predicting the incoming sensory signals, the PP network estimates the precision (inverse variance) of its predictions and balances the bottom-up influences of prediction errors on the network according to the estimated precision of its respective predictions. This allows the network "flexibly to extract signal from noise" (56). Clark, relating to work by Karl Friston, then describes attention "simply as means by which certain error unit responses are given increased weight, hence becoming more apt to drive response, learning and (as we shall later see) action" (57). Results on gaze allocation are interpreted as evidence in favour of the model. Once, we accept the error weighting process and the attention model, it is small a step towards action, because, as Clark suggests, generative models include best sampling expectations and action can be partially seen as a tool for precision-expectation-based sensory sampling (65). Although precision weighting is by far the most important aspect and explanatory ploy of the PP model, a clear and detailed depiction how it actually works is missing.

In chapter 3 the constructive nature of perception is addressed. Since predictions are assumed to form our perceptions, it is suggested that visual illusions are a consequence of exceptional circumstances which do not fit into the well-adapted perceptual expectations of the neural network (85). In a similar fashion perception of omissions (e.g. of notes in a music piece) is explained (89). Again, the worry how predictions allow an agent to perceive the world at all, here stated in form of a lucky imagining or hallucinating argument, is countered by maintaining that perception is different from imagination and hallucination in that it is counterfactually robust and allows attention-based modulation of sensory prediction errors (92/93). One might object perhaps, that his defence is somewhat too quick, because we would, for example, also like to know what exactly constitutes the robustness and veridity of perception (in contrast to hallucination), if both processes depend on the same neural apparatus. The general picture of imagination, here described, is that it is co-emergent with perception: If a *generative* network (a network that generates signals that are predictions of sensory signals) is the foundation for perception then the same network perhaps can generate signals independently of sensory stimulation (93). This account of imagination as being co-emergent with perception is underlined by citing evidence for an overlap between activation patterns that "encode the scenes when merely imagined and when they are perceived" (97). Although Clark in this chapter has also to say something about memory and a relatively new proposal to describe memory in terms of a hierarchical predictive network ("PIMMS and the Past"), it is definitely the most speculative and incomplete part of the book. Memory, especially biographical memory, also raises some questions about the correctness of the predictive processing account of the brain. For biographical memory, seems not to be involved in sub-personal prediction-making. So why is there episodic biographical memory?

Chapter 4 shows how an account of action can be build from the basic PP model. Following Friston and others, it is suggested that in contrast to perception motor control is *subjunctive*. Instead of predicting actual proprioceptive trajectories, what happens in motor control, is the prediction of non-actual trajectories that would result in performing the desired action (121). Central is the claim that instead of adjusting predictions to reduce prediction errors, in action the reduction is achieved by making the prediction real (121). As Clark presents the case, by referring to works by Friston and others, this picture is supported by empirical evidence. What distinguishes the model from traditional approaches to motor control is (i) that no inverse model of motor control and (ii) no cost functions are postulated (125). However, it has in Clark's opinion

also the implication that desires, rewards and pleasure are not the causes of action but only a consequence. Rather, our behaviour is caused by "sub-personal webs of probabilistic expectation" (129). If Clark is right about that point, the PP account of action is deeply at odds with our regular conception of agency for we normally distinguish between what we expect (in a descriptive sense) and what we desire or want to be the case (in a normative sense). But perhaps, the PP proposal that predictions of non-actual future proprioceptive signals can be the cause of bodily action, can be best understood by identifying these predictions, at least in some cases, with desires and intentions.

In chapter 5 Clark deals with the question how an agent comes to understand his own actions and, importantly, the actions of others. The relevant kind of understanding considered here, is 'experiential understanding' as it was for example described by Maurice Merleau-Ponty. In Clark's words: "Some kind of deep, primary, or 'embodied' understanding that enables us to appreciate the meaning of an observed action" (152). One problem that is related to experiential understanding is that the meaning of an action is context-dependent, as Clark illustrates by a "Dr Jekyll or Mr Hyde" example of a man holding a knife to a human chest. Clark's claim is then that the process that generates experiential understanding cannot solely rely "on the feedforward ('bottom-up') flow of sensory information" (153). Instead he considers the top-down predictions and the flexible precision-weighting, which are part of the PP model, to handle this problem. Going from there, he argues for a deflationary view on mirror neurons that assumes that mirror neurons are produced by associative sequence learning. Of course, there remains a problem of action attribution, but this is accounted by precision-weighting on proprioceptive prediction errors (158).

In chapter 6 Clark discusses the mind-world relation in the light of the PP model of perception. The main question posed in the chapter is whether the PP model of perception implies that perception is indirect. He cites views expressed by Frith (2007) and Hohwy (2007) who hold that we actually perceive the brain's model of the world or the brain's best hypothesis. Clark agrees with them in so far that perception is "in some sense an inferential process" but he thinks that their views are mistaken in two ways. He first rejects the idea that the inference-based routes that produce perception introduce a "representational veil between agent and world" (170). Instead he claims that only by probabilistic apparatus of prediction-driven learning the agents is able "to see through the veil of surface statistics" (170). The details are missing here though. Their second mistake, Clark holds, is "a failure to take sufficient account of the role of action" (170), stressing that prediction-driven learning presents us no action-neutral image of the world but one full of possibilities for action (171). Aside from that topic we find some discussions on the question of innateness of knowledge, the close relationship between perception and action, decision-making, and externalism.

Chapter 7 explores conscious experience through the lens of PP. Clark is largely concerned with showing how the PP model was used to give explanations of mental health conditions like schizophrenia. Schizophrenia is associated with two characteristic positive symptoms, hallucinations and delusions. Clark reports the suggestion made by Fletcher and Frith (2009) that both symptoms can be explained by "falsely generated and highly weighted waves of prediction error" (206). As Clark tells us, in the PP model, high weighted waves of prediction error are equivalent to low weighted precision of predictions. All what counts is the balance between top and bottom levels (212). This is assumed to initiate a self-entrenching process (80/81) in which

hallucinations cause delusions which reinforce the original hallucinations. The highly weighted proprioceptive error signals are taken to form the impression that one's own actions are performed by someone else (219). The more detailed explanation given on page 219 though seems to imply that precision weighting is not merely jointly determined by top-down predictions and the incoming sensory signals, but actually in some cases also performed independently of these factors, to compensate a bad tuning of the prediction machine, as Clark puts it, "under such conditions, the only way to restore movement is to artificially inflate the precision of the higher level states". But this, so it seems, is not permitted by the general constraints of the model. At least, this passage is unclear in important ways, and some mathematical hints could have helped to clarify the exact mechanism. Besides schizophrenia, Clark discusses PP accounts of the feeling of conscious presence, the lack of which is linked to depersonalization disorder (227), and emotions like fear in the dark (235).

In Chapter 8 Clark examines the relationship between PP and embodiment by relating the PP account of the brain to the idea of predictive laziness (Simon, 1956) and the principle of ecological balance (Pfeifer & Bongard, 2006). The idea is that many cognitive and agency tasks are solved by heuristics that actively use the body and the environment as a resource. We find a treatment of the Darkened Room Puzzle (Friston, Thornton & Clark, 2012), which asks why a creature that is driven towards a reduction of prediction error is not inclined to increase deprivation and why the minimization of prediction errors is not inconsistent with a striving for novelty. The answer Clark suggests is twofold. What he first observes is that there are creature-defining expectations (basic needs) that cannot be adjusted in a way that leads the organism to seclude itself to a dark room and wait for death (264). The "positive attractions of novelty" are, Clark admits, more difficult to explain, but he thinks that part of the answer is found in "culturally-mediated lifetime learning" (266), an innate tendency "to seek out 'just-novelty-enough' situations" (266) and the creation of designer environments (ch. 9) that "actively favour [...] novelty-seeking and exploration" (266).

Chapter 9 mainly tries to answer what is so special about humans from the perspective of PP. Clark suggests two ideas. One is that the human neural network adapted in ways that allow an "even more complex and context-flexible hierarchical learning than is found in other animals" (276). Complementary to this is the second idea that humans create socially and culturally formed environments which constantly "provide new and ever-more-challenging patterns that will drive learning" (277). We find this latter idea already in Dewey's description of the school as a social institution (Dewey, 1916), but the novelty or relevance of Clark's description perhaps consists in applying it to learning as a mental process *per se*.

Clark in his aim to cover the grand vision of the predictive processing brain in all its aspects is sometimes too busy to concentrate on describing and discussing the important details that should be elaborated when the ambition is to present a workable theory. Details, that bear philosophical challenges as well. For example, one substantial philosophical claim Clark states in his introduction is that "to match the given picture [...] [by predicting] *just* is to understand a lot about [...] [a domain] and [...] [domain-relevant] causes" (5) that seems to be foundational for the whole PP approach to perceptual experience (as it conveys some form of immediate understanding). This perspective on prediction is quite optimistic, and its justification is not discussed. There are other researchers on machine learning who are more critical on the relation between prediction and understanding (e.g. Wheeler, in a conference talk in 2017). And the yet unresolved interpretation problem in quantum mechanics also indicates that there are important

differences between making successful predictions and understanding. For the project of the book, these are perhaps minor issues though.

Overall, *Surfing Uncertainty* is an impressive book that can be read by researchers and graduate students likewise. The author achieves his goal to present a broad picture of "a well-supported vision of the brain" (28) as a predictive machine and combines it with the specific proposal of predictive processing. He can show that predictive processing is conceptually elegant, computationally well-grounded and has a good chance of being neutrally implemented (as his citing of several research studies in computer- and neuroscience indicates). While the book cannot answer all the questions it poses, it may inspire future discussions and help to deepen our understanding of the mind in unexpected ways.

Literature

Bernecker, S. 2014. How to understand the extended mind. Philosophical Issues, 24, 1-23.

Chomsky, N. .1980. On Cognitive Structures and their Development. A reply to Piaget. In: Piattelli-Palmarini, M. (ed.). Language and Learning. The debate between Jean Piaget and Noam Chomsky. Harvard University Press.

Clark, A. / Chalmers, D. 1998. The extended mind. Analysis, 1, 7-19.

Clark. A. 2013. Expecting the World. Perception, Prediction, and the Origins of Human Knowledge. Journal of Philosophy, 9, 469-496.

Clark, A. 2015. What „Extended Me" knows. Synthese, 11, 3757-3775.

Clark, A. 2016. Surfing Uncertainty. Prediction, Action and the Embodied Mind. Oxford: Oxford UP.

Clark, A. 2017. A Nice Surprise? Predictive Processing and the Active Pursuit of Novelty. Phenomenology and the Cognitive Sciences, 1-14.

Clark, A. 2017. How to Knit Your Own Markov Blanket. Resisting the Second Law with Metamorphic Minds. In: Metzinger, T. / Wiese, W. (eds.). Philosophy and Predictive Processing, 3. Frankfurt am Main: MIND Group.

Dewey, John. 1916. Democracy and education. New York: The Free Press.

Fletcher, P. / Frith, C. 2009. Perceiving is believing. A Bayesian approach to explaining the positive symptoms of schizophrenia. Nature Reviews Neuroscience, 10, 48-58.

Fodor, J. / Pylyshyn, Z. 1988. Connectionism and cognitive architecture. A critical analysis. Cognition, 28(1-2), 3-71.

Friston, K. 2010. The free-energy-principle: a unified brain theory? Nature Reviews Neuroscience, 11, 127-128.

Friston, K. / Thornton, C. / Clark, A. 2012. Free-energy minimization and the dark-room problem. Frontiers in Psychology, 3, 1-7.

Frith, C. 2007. Making up the mind. How the brain creates our mental world. Oxford: Blackwell.

Hohwy, J. 2007. Functional integration and the mind. Synthese, 159(3), 315-328.

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

Lupyan, G. / Clark, A. 2015. Words and the World. Predictive Coding and the Language-Perception-Cognition Interface. Current Directions in Psychological Science, 4, 279-284.

Pfeifer, R. / Bongard, J. 2006, How the body shapes the way we think. A new way of intelligence. Cambridge, MA: MIT Press.

Rao, R. / Ballard, D. 1999. Predictive coding in the visual cortex. A functional interpretation of some extra-classical receptive field effects. Nature Neuroscience, 2(1), 79.

Simon, H. A. 1956. Rational choice and the structure of the environment. Psychological Review, 63(2), 129-138.

Searle, John. 1980. Minds, Brains and Programs, Behavioural and Brain Sciences, 3, 417-457.

Wheeler, G. 2017. Remarks on Machine Learning and Instrumental Predictions. Reasoning and Argumentation in Science. Conference at the Munich Center for Advanced Studies, 31.5.-2.6.2017.

Martin Nitsch, Heinrich-Heine-Universität Düsseldorf