Trends in
Cognitive Sciences

CellPress

25th Anniversary Series: Looking Forward

Forum

Reconstructing the predictive architecture of the mind and brain

Floris P. de Lange [1,2,*,@]
Lea-Maria Schmitt, [1,2] and
Micha Heilbron [1,2]

Predictive processing has become an influential framework in cognitive neuroscience. However, it often lacks specificity and direct empirical support. How can we probe the nature and limits of the predictive brain? We highlight the potential of recent advances in artificial intelligence (AI) for providing a richer and more computationally explicit test of this theory of cortical function.

The past decade has witnessed a paradigm shift in how we view perception and cognition. The idea that our brains use predictive processing to rapidly perceive and comprehend the world has become increasingly influential in cognitive neuroscience. There have been intense research efforts both to build neuro-computational theories of predictive processing that hold promise in providing a general theory of cortical function and to obtain empirical evidence for predictive computations in cortical neurons [1]. For example, it is now well established that the neural response to an identical input is strongly modulated by whether that input is expected or surprising. After learning the regularities in our environment, the neural response to predictable input can be reduced as much as fivefold compared to when the same input occurs unpredictably [2]. This observation is broadly compatible with a predictive neural architecture, in which predictions are continuously updated to minimize neural activity representing prediction errors.

Despite this overall success, several challenges currently limit the appeal of predictive processing as a general theory of cortical function. One issue is that predictive processing is a rather general framework, mostly defined at a computational and algorithmic level and encompassing multiple possible implementations. For example, a common motif in predictive processing is the notion that information processing between neural populations is bidirectional, and that this recurrent message-passing is at the heart of optimal inference in the brain. Indeed, there is abundant anatomical evidence for bidirectional connections. However, the nature and form of the signals that are being passed up and down are still topics of debate. In fact, multiple qualitatively different types of feedback connections have been identified in the cortex [3], suggesting that there may not be one canonical function of feedback signaling. Also, researchers studying predictive processing often assume the existence of two types of neurons, prediction and prediction error neurons, signaling the current best guess and the mismatch between this guess and the actual input, respectively. However, predictive processing models without explicit error representation have also been defined, and even models that only retain a representation of the error. This multitude of possibilities – all captured under the umbrella term of predictive processing – may hinder progress in the field.

Compounding this conceptual unclarity, empirical tests of predictive processing have often been rather indirect. Particularly in the field of cognitive neuroscience, measurements are often made at a much coarser level of granularity than the hypothesized model signals. For example, noninvasive measurements of neural activity using fMRI are a metabolic reflection of synaptic activity of tens of thousands of neurons integrated over hundreds of milliseconds. This makes it difficult to link model predictions about the temporal dynamics of prediction and error neurons to observed neural activity. Observed activity patterns are often compatible with, but not strongly diagnostic for, predictive processing.

How can we move forward? Progress in recent years has often come from more precise measurements: for example, from Neuropixels probes that allow concurrent recording of well-isolated spiking activity from thousands of neurons in multiple areas, or from recent developments in MRI that allow non-invasive recording of activity at submillimeter laminar precision. Another area of potential progress, which we believe holds great promise for improving theories of predictive processing, lies in the development of more precise computational models.

Recent advances in the field of artificial intelligence and machine learning have put artificial neural networks (ANNs) firmly on the map as a promising approach to both model and understand biological brains. ANNs are multilayer neural networks that are loosely inspired by biological neural networks, and come in many flavors: for example, deep or shallow, feedforward or recurrent, discriminative or generative, supervised or unsupervised. These networks can be trained to solve complex tasks like object categorization or natural language processing. There are at least two ways in which ANNs can help push the field of predictive processing forward. First, ANNs can be harnessed as a tool to quantify predictability and surprise in rich naturalistic environments. After training certain classes of ANNs to predict future input, they can provide a rich multilevel quantification of what is expected and surprising in the input we are receiving, at different levels of feature complexity. In other words, the ANNs are used as a tool to estimate the stimulus statistics a predictive brain is expected to track. This allows for much stronger empirical tests of predictive

processing, beyond highly artificial and experimentally controlled situations. This approach of using ANNs to generate stronger empirical tests of predictive processing has been successfully applied to uncover the type and nature of predictive neural computations in vision [4] and language [5,6].

Second, rather than simply using ANNs as a tool to formalize and approximate the predictions the brain might be making, ANNs can serve as mechanistic hypotheses of neural information processing itself. In the field of machine vision, for example, a relatively simple class of feedforward neural networks is the so-called convolutional neural network (CNN). While CNNs achieve impressive performance in categorizing visual input and accurately simulate early feedforward responses within the visual ventral stream [7], they lack several key ingredients that define biological neural networks (i.e., lateral and feedback connectivity). Interestingly, while CNNs mimic primate brain and behavior well for relatively simple visual input, recurrent ANNs (RNNs) are critical for perceptual inference under more challenging conditions, such as occlusion, clutter, and blur [7]. This suggests a potential functional role of feedback for disambiguating input, specifically under challenging conditions. It will be an exciting avenue of future research to determine the exact computational goal(s) of recurrent connectivity, for example by providing ANNs with different feedback training regimes, and directly comparing the geometry of their 'neural' representations as well as their 'behavioral' outputs [8,9].

Whereas older neurocomputational theories were typically limited in scale and biological plausibility, new models and methods allow us to probe the internal representations of different ANN architectures during the processing of naturalistic input. Interestingly, ANNs trained to predict their input are gaining popularity in both language [6] and perception [10]. For example, a recurrent generative network trained to predict future visual input in a self-supervised manner has been shown to reproduce a range of empirical well-known findings in visual cortex and visual behavior, such as the complex time-varying responses of neurons to static input [10]. Also, when training recurrent neural networks to minimize their energy consumption while operating in predictive environments, the networks self-organize into prediction and error units with appropriate inhibitory and excitatory interconnections and learn to inhibit predictable sensory input [11], a key design principle of many predictive processing architectures.

The recent revolution in artificial intelligence and its influence on cognitive neuroscience may suggest that older theories of visual perception will simply be replaced by computational models in the form of ANNs. That, however, would be a mistake [12]. Observing that an ANN can successfully solve a particular cognitive problem does not necessarily provide understanding of how the problem has been solved. Rather, ANNs can serve as computationally explicit formalizations of hypothetical processing mechanisms and provide falsifiable predictions about how internal representations and behavior emerge from architectural and training constraints. In this way, the computational principles identified in ANNs can be used to refine our theories on the predictive nature of neural computations.

In conclusion, it has become abundantly clear in the past decade that brains and minds are predictive, using prior knowledge to constrain incoming information and rapidly make sense of the world. However, the specific design principles underlying this remarkable feat are still less well understood. It is our hope that in the next 25 years cognitive neuroscientists will embrace the new tools furnished by artificial intelligence to unravel how the predictive mind is neurally implemented. We expect that this will catalyze developments in both cognitive neuroscience and artificial intelligence by creating better understanding and development of intelligent systems.

Declaration of interests

The authors declare no competing interests.

[1]Donders Institute for Brain, Cognition and Behaviour, Radboud University, Kapittelweg 29, 6525, EN, Nijmegen, The Netherlands
[2]https://www.predictivebrainlab.com

*Correspondence:
floris.delange@donders.ru.nl (F.P. de Lange).
@Twitter: @flodlan (F.P. de Lange).

https://doi.org/10.1016/j.tics.2022.08.007

References

1. de Lange, F.P. *et al.* (2018) How do expectations shape perception? *Trends Cogn. Sci.* 22, 764–779
2. Schneider, D.M. *et al.* (2018) A cortical filter that learns to suppress the acoustic consequences of movement. *Nature* 561, 391–395
3. Markov, N.T. *et al.* (2013) Cortical high-density counter-stream architectures. *Science* 342, 1238406
4. Uran, C. *et al.* (2022) Predictive coding of natural images by V1 firing rates and rhythmic synchronization. *Neuron* 110, 1240–1257.e8
5. Heilbron, M. *et al.* (2022) A hierarchy of linguistic predictions during natural language comprehension. *Proc. Natl. Acad. Sci. U. S. A.* 119, e2201968119
6. Schmitt, L.-M. *et al.* (2021) Predicting speech from a cortical hierarchy of event-based timescales. *Sci. Adv.* 7, eabi6070
7. Kar, K. *et al.* (2019) Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nat. Neurosci.* 22, 974–983
8. Golan, T. *et al.* (2020) Controversial stimuli: pitting neural networks against each other as models of human cognition. *Proc. Natl. Acad. Sci.* 117, 29330–29337
9. Lindsay, G.W. *et al.* (2022) Bio-inspired neural networks implement different recurrent visual processing strategies than task-trained ones do. *bioRxiv* Published online March 08, 2022. doi.org/10.1101/2022.03.07.483196
10. Lotter, W. *et al.* (2020) A neural network trained for prediction mimics diverse features of biological neurons and perception. *Nat. Mach. Intell.* 2, 210–219
11. Ali, A. *et al.* (2021) Predictive coding is a consequence of energy efficiency in recurrent neural networks. *bioRxiv* Published online November 16, 2021. doi.org/10.1101/2021.02.16.430904
12. Saxe, A. *et al.* (2021) If deep learning is the answer, what is the question? *Nat. Rev. Neurosci.* 22, 55–67