JNphi

Too Rational: How Predictive Coding's Success Risks Harming the Mentally Disordered and ill

Lee Elkin, Karolina Wiśniowska

Abstract

The so-called predictive coding or predictive processing theory of mind has attracted significant attention in the brain and behavioral sciences over the past couple of decades. We aim to discuss an important ethical implication of the theory's success. As predictive coding has become influential in the study of mental disorders and illness, particularly on autism spectrum disorder (ASD) and schizophrenia, we highlight a significant risk of further harming an already stigmatized population. Specifically, because predictive coding is undergirded by Bayesian inference, and Bayesian inference is often thought to imply 'rationality', the cognitive framework engenders a risk of strengthening existing negative attitudes towards individuals having mental disorders and illnesses by associating such individuals with also having 'irrational brains.' In defending the salience of the risk, we base our argument on historical examples of socially harmful effects propagated by psychological constructs of intelligence and suggest that predictive coding may be headed down a similar path. We conclude that scientific researchers promoting predictive coding should proactively fulfill a moral duty to mitigate harm to those that are the subject of their studies. Thus, we implore those promoting predicting coding to take care in the language used and examine the theoretical and practical implications of such language as the program continues forward.

Key Words: predictive coding, Bayesian brain, ASD, schizophrenia, ethics

Introduction

The so-called predictive coding or predictive processing theory of mind has attracted significant attention in the brain and behavioral sciences over the past couple decades, drawing on empirical and theoretical findings in cognitive linguistics (Tenebaum *et al.*, 2006), neuroscience (Friston, 2009; 2010), philosophy (Clark 2013; Hohwy 2013), and psychology (Oaksford & Chater, 2007). Its application has enjoyed much empirical success in scientifically explaining a vast range of cognitive phenomena, including (but not limited to) perception (Knill & Pouget, 2004; Howhy *et al.*, 2008) and higher levels of cognition such as beliefs and desires (Yon *et al.*, 2020). Alongside, predictive coding has become influential in the study of developmental disorders such as autism spectrum disorder (ASD) (Pellicano & Burr, 2012; Van de Cruys *et al.,* 2014) and mental illness such as schizophrenia (Horga *et al.,* 2014; Powers *et al.*, 2017).

Corresponding author: Lee Elkin
Address: Erasmus School of Philosophy/EIPE, Erasmus University Rotterdam, Burg. Oudlaan 50 (Bayle 5-69), 3062 PA Rotterdam, The Netherlands and Karolina Wiśniowska, Institute of Philosophy/INCET, Jagiellonian University, ul. Grodzka 52 31-044 Kraków, Poland
e-mail ✉ elkin@esphil.eur.nl
Received: 06.03.2022; **Accepted:** 08.04.2022

Despite its rise in popularity, there is no shortage of critics (see e.g., Baetu, *et al.,* 2011; Glymour, 2011; Colombo & Wright, 2017; Colombo *et al.*, 2021). Most critics have pointed to some theoretical deficiency or implausible assumption(s) of the framework, but one dimension of criticism, or at least a point of concern, that seems obvious to us, though apparently not to others, pertains to the ethical consequences of the cognitive theory's success. On the relation between predictive coding and ethics, Andy Clark (2017) discussed in a blog post possible explanatory implications concerning biased perception in policing. Aside from this blog post, though, ethical considerations, in general, appear not to even be a second thought among proponents and critics of predictive coding alike.

Here, we want to highlight a glaring concern with the theory that does not further an ethical debate via scientific explanations of, say, biased perception in policing. Rather, our aim is to illuminate how predictive coding's growing empirical success may in the near future be the cause of a whole new ethical problem through unintended consequences, similar to other psychological constructs like measures of intelligence, in amplifying negative stereotypes and stigma towards mental disorders and illnesses, especially given the recent rise to prominence of predictive coding in scientific research on ASD and schizophrenia. Because predictive coding is undergirded by Bayesian inference, and Bayesian inference is often thought to imply 'rationality', the cognitive framework engenders a risk of strengthening existing negative attitudes towards individuals having mental disorders and illnesses (Corrigan *et al.*, 2004) by associating such individuals with also having 'irrational brains', thus causing further harm to already stigmatized groups.

What is predictive coding?

Predictive coding is a theory of mind built on the philosophical thesis that the brain aims at minimizing prediction errors through a hierarchical computational framework (Clark, 2013), or described in information-theoretic and physics terminology, minimizing free energy (Friston, 2009). The rationale of the thesis is that the brain must optimally infer the correct "hidden" state of the world under uncertainty due to sensory information signals obtained through perception sometimes being ambiguous or noisy. In achieving the goal of inferring the correct state in the face of perceptual uncertainty, it is claimed that the brain is organized into a hierarchical system in which informational inputs are checked against prior predictions in a bottom-up manner. If there are mismatches between sensory inputs and prior expectations, i.e., prediction errors, the brain corrects itself from the top down, taking into account the mismatches from which it forms new or updated predictions with respect to the relevant space of possible states.

In more detail, the inferential procedure is casted as an active employment of a generative model by the brain in which prior expectations determined by a prior probability distribution over uncertain states are combined with the likelihood of the sensory data, i.e., the probability of the data *d* given some state*s*, thereby yielding a posterior probability that is adopted in the new or updated expectation. The inferential procedure generally falls in line with *Bayesian inference*, which is said to be approximated by the brain. On the most promising accounts of predictive coding, the cognitive model includes weight or *precision* on the top-down predictions and bottom-up evidence, generalizing the Bayesian inferential model, which allows for more or less weight to be placed on either side (Yon *et al.*, 2020). The precision construct in such generalized computational cognitive models has become the key to the theory's recent success, especially in explaining certain mental disorders and illnesses.

For instance, slow development associated with autism spectrum disorder (ASD) is plausibly explained under the predictive precision-weighting model through high precision placed on prediction errors. When the brain consistently places too much weight on prediction errors, it prevents abstract representations due to its fixation on each token sensory input, causing the brain to attend to accidental or misleading environmental cues that should instead be down weighted or ignored entirely (Van de Cruys *et al.,* 2014). Constant overweighting of the bottom-up part of the process can lead to slower processing

of sensory information and stunt cognitive development. By contrast, hallucinatory episodes associated with psychosis and schizophrenia are plausibly explained under the predictive-weighting model through high precision placed on prior expectations. Within the framework of predictive coding, disproportionately higher precision constantly placed on priors implies that incoming sensory data is down-weighted or ignored by the brain, which can in turn give rise to perceptions of objects that are not present but are expected by the brain to be present based on prior expectations. An overweighting of the top-down part of the process can therefore give rise to hallucinations. Experimental evidence in support of strongly weighted prior expectations has recently been found (Powers *et al.*, 2017).

While these accounts of cognitive phenomena associated with autism and schizophrenia appear to be conceptually plausible, the predictive coding approach in neuroscience and psychiatry is not without its challenges. There are, of course, rival theories that explain symptoms of ASD (see e.g., Baron-Cohen *et al.,* 2000). Even more concerning for the theory are confirmed but contradictory predictions with respect to hallucinations/psychosis, considering that empirical evidence has been found for both overweighting prior expectations as well as overweighting incoming sensory information (Sterzer *et al.,* 2018 for a review). This conflict indicates an underdetermination problem for the predictive coding theory of mind.

Despite the aforementioned challenges, however, the framework has had significant influence on many domains in cognitive science. As Rescorla (2020) suggests, there are, of course, alternative frameworks to Bayesian perceptual psychology in the field of cognitive science, but none in the current paradigm come close in explanatory power.

Discussion

Although constructs of cognitive ability are necessary in advancing a scientific understanding of the brain and behavior, some constructs are historically known for yielding harmful effects on social groups, whether intended or unintended. Among the most extreme, scientific notions of intelligence in the early twentieth century became instrumental in catapulting some eugenics movements. For example, past compulsory sterilization programs, as well as marriage laws and immigration restrictions, in the United States targeted 'feebleminded' people, or people having low intelligence (Kröner, 1999). The U.S. Supreme Court's ruling in Buck v. Bell 1927 serves as a prime example of the harmful effects of intelligence constructs, which deemed compulsory sterilization of feeble minds constitutional. Justice Oliver Wendell Holmes wrote that the compulsory sterilization order for the plaintiff, Carrie Buck, permitted under the Virginia Sterilization Act was not a violation of the constitution and remarked that "[T]hree generations of imbeciles are enough" (Buck v. Bell, 1927). In this case, and others, the definition of 'imbecility' was (often loosely) based on the Binet-Simon intelligence scale (Lombardo, 1985), which was believed by many eugenicists to be a useful psychological tool.

Even in case of subsequently revised measures that have been administered with no ill-intention, problems with standardization remain, especially since IQ measures, which are still used today, were developed around Western, white, middle-class populations with certain cultural values (Mensh & Mensh, 1991). Non-member groups thus would have been (and still are) at a significant disadvantage in testing for intelligence, as they likely would (and will continue to) have lower scores as a result of the design. These historical observations clearly indicate that the institutionalization of measures of intelligence coupled with tendentious testing and framing of results has had a largely negative effect on society, and its reach has extended to education (Fancher, 1985; Mensh & Mensh, 1991), industry (Scarr, 1978; Fancher, 1985), and social domains generally (Chase, 1977; Mensh & Mensh, 1991; Andersen, 1994). While the scientific community has come to recognize these problems and evidence that intelligence is not a predictor of income (Zagorsky, 2007), nor are race or sex predictors of intelligence (Hunt, 2010), the notion has had a lasting effect given that it still resides in character appraisal.

Considering the past harms from constructs of intelligence, we contend that predictive coding might end up on a similar path as it concerns autistic and schizophrenic individuals (and others more broadly as the framework continues to be applied) that the cognitive science community should carefully consider since there is no feasible take back, especially in public opinion, once the cat is let out of the bag. We base this prediction on how Bayesian inference is construed in other domains. In the computational and statistical sciences, for example, 'Bayesian inference' is often equated with 'rational inference' given its philosophical foundations (Ramsey, 1926; de Finetti, 1974). Thus, the 'Bayesian brain' is interchangeable with the (probabilistically) 'rational brain'. But as described in the previous section, autistic and schizophrenic individuals tend to stray from being optimally Bayesian by either giving too much weight to the likelihood (ASD and schizophrenia) and not enough to the prior or *vice versa* (schizophrenia). Deviation from being optimally Bayesian implies a tendency towards 'irrationality' given that being optimally Bayesian supposedly implies 'rationality'. We thus find that a likely conclusion will emerge from cognitive scientific research, though not necessarily with intent or malice, that individuals diagnosed with ASD and schizophrenia, and others more broadly as the framework continues to be applied, have 'irrational brains.'

We find the potential realization of this consequence extremely worrisome because those having mental disorders or illnesses can, in fact, be rational, on an ordinary conception, across various contexts, while healthy individuals can, in fact, be irrational on different occasions (Kahneman, 2011 for a review of cognitive biases). But the misplaced negative ascription is not our main worry. Our main worry concerns the harm that will be inflicted on an already stigmatized and stereotyped population resulting from the program. Among the harms that such individuals already experience is a lack of self-improvement and management of their mental health due to a reluctance to seek help because of the stigma attached to mental illness in public opinion. People with mental illness are often thought of as a danger to the public and themselves, leading to social distancing (Parcesepe & Cabassa, 2013). Mental health stigmatization has, in turn, caused patients to underreport mental health conditions in comparison to reporting non-mental health conditions (Bharadwaj *et al.*, 2017). Without medical help, however, such patients go untreated.

A failure to improve and manage one's conditions due to the judgmental grip of society strangling the will of mentally disordered and ill individuals is not only harmful but unjust given a denial of fair and equal opportunity to living a mentally health life. Because mental health stigma is real, not hypothetical, individuals affected by mental disorders and illness have indeed been harmed by negative public sentiment. We thus caution those promoting the predictive coding program that further harm might be inflicted. The reason is that the descriptor 'irrational' not only coheres with 'danger' in ordinary language but might even amplify it, causing increased fear of those with mental abnormalities. This outcome would not only be unfortunate, but socially devastating for such groups. In line with the Belmont Report's beneficence principle concerning research ethics and Mill's Harm Principle (1859), we suggest that scientific researchers promoting predictive coding proactively fulfill a moral duty to mitigate harm to those that are the subject of their studies. Thus, we implore those promoting predicting coding to take care in the language used and examine the theoretical and practical implications of such language as the program continues forward.

Acknowledgments

We would like to thank Tomasz Żuradzki for his valuable feedback on the paper. We are grateful for receiving financial support from the European Research Council and Dutch Research Council that supported the research for this paper.

Conflict of interest statement

None declared.

Funding

Lee Elkin and Karolina Wiśniowska were supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant No. 805498). Lee Elkin was also supported by the Dutch Research Council (NWO) through the VIDI project ENCODE, project no. VI.Vidi.191.105.

References

Andersen, ML. The many and varied social constructions of intelligence. In TR Sarbin and JI Kitsuse (eds.), Constructing the Social. Sage Publications, 1994: 119-38.

Baetu I, Barberia I, Murphy RA and Baker AG. (2011). Maybe this old dinosaur isn't extinct: What does Bayesian modeling add to associationism?. Behavioral and Brain Sciences 2011; 34: 190-191.

Baron-Cohen S, Ring HA, Bullmore ET, Wheelwright S, Ashwin C and Williams SCR. The amygdala theory of autism. Neuroscience & Biobehavioral Reviews 2000; 24: 355-364.

Bharadwaj P, Pai MM and Suziedelyte A. Mental health stigma. Economics Letters 2017; 159: 57-60.

Chase A. The Legacy of Malthus: The Social Costs of the New Scientific Racism. Knopf, 1977.

Clark A. Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behavioral and Brain Sciences 2013; 36: 181-204.

Clark A. Neuroethics, the predictive brain, and hallucinating neural networks. The Neuroethics Blog, 2017. <http://www.theneuroethicsblog.com/2017/12/neuroethics-predictive-brain-and.html>

Colombo M, Elkin L and Hartmann S. Being realist about Bayes, and the predictive processing theory of mind. The British Journal for the Philosophy of Science 2021; 72: 185-220.

Colombo M and Wright C. Explanatory pluralism: An unrewarding prediction error for free energy theorists. Brain and Cognition 2017; 112: 3-12.

Corrigan PW, Markowitz FE and Watson AC. Structural levels of mental illness stigma and discrimination. Schizophrenia Bulletin 2004; 30: 481-491.

De Finetti B. Theory of Probability: A Critical Introductory Treatment. Wiley, 1974.

Fancher RE. The Intelligence Men: Makers of the IQ Controversy. W.W. Norton and Company, 1985.

Friston K. The free-energy principle: a rough guide to the brain? Trends in Cognitive Sciences 2009; 13: 293-301.

Friston K. The free-energy principle: a unified brain theory? Nature Reviews Neuroscience 2010; 11: 127-138.

Glymour C. Osiander's psychology. Behavioral and Brain Sciences 2011; 34: 199-200.

Hohwy J. The Predictive Mind. Oxford University Press, 2013.

Hohwy J, Roepstorff A and Friston, K. Predictive coding explains binocular rivalry: An epistemological review. Cognition 2008; 108: 687-701.

Holmes OW and Supreme Court of The United States. U.S. Reports: Buck v. Bell, 274 U.S. 200, 1927. [Periodical] Retrieved from the Library of Congress. https://www.loc.gov/item/usrep274200/

Horga G, Schatz KC, Abi-Dargham A and Peterson BS. Deficits in predictive coding underlie hallucinations in schizophrenia. Journal of Neuroscience, 2014; 34: 8072-82.

Hunt E. Human Intelligence. Cambridge University Press, 2010.

Kahneman D. Thinking, Fast and Slow. Macmillan, 2011.

Knill DC and Pouget A. The Bayesian brain: the role of uncertainty in neural coding and computation. Trends in Neurosciences 2004; 27: 712-719.

Kröner HP. From eugenics to genetic screening. In The Ethics of Genetic Screening. Springer, 1999: 131-145.

Lombardo PA. Three generations, no imbeciles: New light on Buck v. Bell. New York University Law Review 1985; 60: 30-62.

Mensh E and Mensh H. The IQ Mythology: Class, Race, Gender, and Inequality. SIU Press, 1991.

Mill JS. On Liberty. John W. Parker and Son, 1859.

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. The Belmont report: Ethical principles and guidelines for the protection of human subjects of research. U.S. Department of Health and Human Services, 1978.

Oaksford M and Chater N. Bayesian Rationality: The Probabilistic Approach to Human Reasoning. Oxford University Press, 2007.

Parcesepe AM and Cabassa LJ. Public stigma of mental illness in the United States: a systematic literature review. Administration and Policy in Mental Health and Mental Health Services Research 2013; 40: 384–399.

Pellicano E and Burr D. When the world becomes 'too real': a Bayesian explanation of autistic perception. Trends in Cognitive Sciences 2012; 16: 504-510.

Powers AR, Mathys C and Corlett PR. Pavlovian conditioning–induced hallucinations result from overweighting of perceptual priors. Science 2017; 357: 596-600.

Ramsey FP. Truth and probability. In The Foundations of Mathematics and Other Logical Essays. Routledge, 1926.

Rescorla M. A realist perspective on Bayesian cognitive science. In A Nes and T Chan (eds.), Inference and Consciousness. Routledge, 2020.

Scarr S. From evolution to Larry P., or what shall we do about IQ tests? Intelligence 1978; 2: 325-342.

Sterzer P, Adams RA, Fletcher P, Frith C, Lawrie SM, Muckli L and Corlett PR. The predictive coding account of psychosis. Biological Psychiatry 2018; 84: 634-643.

Tenenbaum JB, Griffiths TL and Kemp C. Theory-based Bayesian models of inductive learning and reasoning. Trends in Cognitive Sciences 2006; 10: 309-318.

Van de Cruys S, Evers K, Van der Hallen R, Van Eylen L, Boets B, de-Wit L and Wagemans J.Precise minds in uncertain worlds: predictive coding in autism. Psychological Review 2014; 121: 649-675.

Yon D, Heyes C and Press C. Beliefs and desires in the predictive brain. Nature Communications 2020; 11: 1-4.

Zagorsky JL. Do you have to be smart to be rich? The impact of IQ on wealth, income and financial distress. Intelligence 2007; 35: 489-501.