Acta Bioethica 2024; 30(2): 263-272. DOI: http://dx.doi.org/10.4067/S1726-569X2024000200263

COMPUTING ADDICTION: EPISTEMIC INJUSTICE CHALLENGES IN THE CULTURE OF COMPUTATIONAL PSYCHIATRY

Min Wang[1], Zhoukang Wu[2], Liangjiecheng Huang[3], Xiaochu Zhang[4], Xiaosong He[5]

Abstract: Computational psychiatry (CP), based on artificial intelligence technology, plays an increasingly important role in scientific research and medical diagnosis. Epistemic concerns in the ethics of artificial intelligence have also been at the center of debate in CP, but the different epistemic forms of injustice caused by the internal cultures of CP remain unexplained. We distinguish between data-driven and theory-driven cultures and their research purposes via practical examples of CP models deployed in addiction. A data-driven culture may advance medical understanding of biological categories of mental illness, whereas a theory-driven culture provides better explanatory mechanisms between symptoms and biology. We discuss testimonial injustice caused by the silencing of patient voices in a data-driven culture, and hermeneutic injustice caused by the non-sharing of hermeneutic resources in theory-driven culture based on Miranda Fricker's account of epistemic injustice. We analyze the factors underlying nuances in epistemic forms between the two, such as naturalistic-dominated medical understanding and the system's epistemic privileging. The above epistemic risks all indicate the intricacies of mental disorders and require that success be assessed in terms of actual benefit to patients. Finally, we emphasize the importance of the patient's phenomenology and call for greater inclusion of patients in psychiatric decision-making processes.

Keywords: epistemic injustice, computational psychiatry, addiction, testimonial and hermeneutical injustice, epistemic privileging

Adicción a la informática: Los retos de la injusticia epistémica en la cultura de la psiquiatría computacional

Resumen: La psiquiatría computacional (PC), basada en la tecnología de la inteligencia artificial, desempeña un papel cada vez más importante en la investigación científica y el diagnóstico médico. Las preocupaciones epistémicas en la ética de la inteligencia artificial también han estado en el centro del debate en la PC, pero las diferentes formas epistémicas de injusticia causadas por las culturas internas de la PC siguen sin explicarse. Distinguimos entre las culturas basadas en los datos y las impulsadas por la teoría y sus propósitos de investigación mediante ejemplos prácticos de modelos de CP mostrados en la adicción. Una cultura impulsada por los datos puede hacer avanzar la comprensión médica de las categorías biológicas de las enfermedades mentales, mientras que una cultura impulsada por la teoría proporciona mejores mecanismos explicativos entre los síntomas y la biología. Basándonos en el relato de Miranda Fricker sobre la injusticia epistémica, discutimos la injusticia testimonial causada por el silenciamiento de los pacientes en una cultura centrada en los datos y la injusticia hermenéutica causada por el hecho de no compartir estos recursos en una cultura centrada en la teoría, Analizamos los factores que subyacen a los matices en las formas epistémicas entre ambas, como la comprensión médica dominada por el naturalismo y el privilegio epistémico del sistema. Todos los riesgos mencionados indican la complejidad de los trastornos mentales y exigen que el éxito se evalúe en términos de beneficio real para los pacientes. Por último, hacemos hincapié en la importancia de la fenomenología del paciente y pedimos una mayor inclusión de los pacientes en los procesos de toma de decisiones psiquiátricas.

Palabras clave: injusticia epistémica, psiquiatría computacional, adicción, injusticia testimonial y hermenéutica, privilegios epistémicos

Dependência de computação: desafios da injustiça epistêmica na cultura da psiquiatria computacional

Resumo: Psiquiatria computacional (PC), baseada na tecnologia da inteligência artificial, tem um papel crescentemente importante na pesquisa científica e no diagnóstico médico. Preocupações epistêmicas na ética da inteligência artificial tem também estado no centro das discussões da PC, mas as diferentes formas epistêmicas de injustiça causadas pelas culturas internas da PC permanecem inexplicáveis. Distinguimos entre culturas orientadas por dados e orientada por teorias e seus propósitos de pesquisa via exemplos práticos de modelos de PC empregados em dependências. Uma cultura orientada por dados pode avançar a compreensão médica de categorias biológicas de doença mental, enquanto uma cultura orientada por teorias fornece melhores mecanismos explicativos entre sintomas e biologia. Discutimos injustiça testemunhal causada pelo silenciamento de vozes de pacientes em uma cultura orientada por dados e injustiça hermenêutica causada pelo não compartilhamento de recursos hermenêuticos em cultura orientada por teorias, baseados no relato de Miranda Fricker sobre injustiça epistêmica. Analisamos os fatores subjacentes às nuances nas formas epistêmicas entre os dois, tais como a compreensão médica dominada pelo naturalismo e o privilégios epistêmicos do sistema. Todos os riscos epistêmicos citados anteriormente indicam as complexidades dos transtornos mentais e requerem que sucesso seja avaliado em termos de benefício real aos pacientes. Finalmente, enfatizamos a importância da fenomenologia do paciente e clamamos por maior inclusão de pacientes em processos de tomada de decisão psiquiátrica.

Palavras chave: injustiça epistêmica, psiquiatria computacional, dependências, injustiça testemunhal e hermenêutica, privilégios epistêmicos

[1] Department of Psychology, School of Humanities and Social Sciences, University of Science and Technology of China, Hefei, Anhui, China.

[2] Department of Psychology, School of Humanities and Social Sciences, University of Science and Technology of China, Hefei, Anhui, China.

[3] Department of Psychology, School of Humanities and Social Sciences, University of Science and Technology of China, Hefei, Anhui, China.

[4] Hefei National Laboratory for Physical Sciences at the Microscale and School of Life Sciences, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei, Anhui, China.

[5] Department of Psychology, University of Science and Technology of China, 96 Jinzhai. hexs@ustc.edu.cn

Introduction

Naturalistic approaches to treating mental disorders have been a long-standing pursuit of psychiatry. In response to the difficulties of classifying mental disorders and the challenge of establishing psychiatric biomarkers, computational psychiatry (CP) is progressively gaining ground as a way forward in key fields such as mental health and behavioral science*(1)*. This emerging paradigm promises to enhance objectivity and consistency in diagnosing and treating psychiatric disorders while achieving notable precision and reliability. Concurrently, this field brings to the fore significant ethical considerations, including the foundational principles of biomedical ethics*(2)*, data ownership and protection, the risk of reductionism and the neglect of conscious experience*(3,4)*. However, an equally important epistemic ethical dimension in CP remains to be explained: the propensity of computational methods, particularly within the realm of addiction, to engender epistemic injustice, a scenario in which individuals are wronged in their capacity as knowers and communicators of their experiences.

Epistemic injustice, a term that encapsulates the unfairness inflicted upon individuals in their role as contributors to knowledge, has garnered substantial academic scrutiny since its introduction by Miranda Fricker*(5)*. A significant segment of this scholarly exploration is dedicated to discerning the origins of epistemic injustice within the realm of scientific endeavor. Within the broad scope of CP, different methodologies, characterized as data-driven and theory-driven cultures, coexist and sometimes converge*(6)*. There are unique ethical challenges for both, particularly with the creation, interpretation, and application of knowledge in clinical settings ——the nuances of epistemic injustice as it manifests within these two distinct cultures of CP. Using machine learning diagnostics and computational cognitive models in addiction research as examples, we highlight the potential silencing of certain moral claims and the lack of a linguistic framework for certain experiences.

CP encompasses a vast array of methodologies, each with unique potential and challenges in the realm of mental health. Our primary objective is to determine whether a blanket dismissal of CP in addiction research would be ethically remiss, given its prospective advantages. However, it is imperative to critically assess whether a specific CP approach genuinely augments clinical outcomes or merely represents an intricate computational endeavor. With addiction as our focus, we first describe the basic differences between data-driven and theory-driven cultures in CP. Subsequently, we examine specific case studies that exemplify these cultures. To support our central argument, we analyze these cases through the lens of the epistemic injustice framework and emphasize the ethical implications of testimonial and hermeneutical injustices in the application of CP in addiction research.

Computational psychiatry in two cultures in addiction studies

The CP is an interdisciplinary field centered on theories such as reinforcement learning, dynamic systems, neural networks, Bayesian decision-making, and sequential sampling to understand, diagnose, and treat mental disorders*(7,8)*. In contrast to psychological theories, the basic theories of computational psychiatry originate from the fields of mathematics, computer science and cognitive neuroscience. These fields present systematic methodologies to bridge various levels of analysis and offer insights into neurocomputational functions*(9)*. Thus, the CP algorithm completes, to some extent, the transition from the rational analysis of a specific problem to the algorithmic complexity of the solution and its potential biological implementation.

Breiman's conceptual framework differentiates "algorithmic modeling" and "data modeling"*(10)*. The former, rooted in a predictive paradigm, seeks to forecast the outcomes produced by the data-generating process based on given inputs, without considering the black box of the process. In contrast, "data modeling" endeavors to elucidate the inner workings of the data-generating process by analyzing the relationship between its inputs and outputs. As it migrates to psychiatry, two parallel trajectories are recognized: machine-learning approaches, which prioritize the prediction of psychiatric outcomes, and explanatory modeling, which is used to unravel the computational and biological underpinnings of psychiatric

disorders*(11)*. Although these trajectories are often labeled data-driven or theory-driven respectively, it is critical to understand that they are not mutually exclusive. A typical example is the application of deep neural networks. Depending on the context, such a network can serve an explanatory role by mirroring the biophysical dynamics of psychiatric dysfunction, or a predictive role by classifying and forecasting diagnoses. Arguably, both cultures draw on common statistical tools and methodologies, while diverging in their ultimate goal: prediction versus explanation.

Biological psychiatry defines the brain as the organ that generates, maintains and supports psychological functions. The data-driven culture of CP responds precisely to this vision of biological psychiatry, namely the identification of biomarkers for psychiatric diagnosis. Data-driven paradigms typically perform discriminative classification problems in two ways: supervised learning and unsupervised learning*(12).* The former involves training a model on a labeled dataset for which the results are known. This approach enables the model to learn the relationships between input features and outcomes (e.g., specific biological characteristics and the prevalence or cure rate of a certain disease) and to subsequently make predictions on new, unseen data. In contrast, the latter method attempts to uncover hidden patterns or structures in data. Since it does not rely on labeled results, unsupervised learning is widely used in situations where the results are not clearly defined or are in the exploratory stage of research. *(13)* As the cornerstone of both types of learning, computer science algorithms (e.g., support vector machines or deep neural networks) are often used to construct linear or nonlinear relationships between inputs and outputs and to avoid model overfitting through parameter regularization and cross-validation*(14).* All these measures contribute to increasing the ecological validity of machine learning and successfully applying it to addiction research and practice.

Since people with substance use disorder have a clear history of drug intake and physical symptoms of dependence, it is common to use supervised learning to identify changes in their brain structure and function*(15)*. This method utilizes a priori boundaries between people with addiction and people without addiction to label the dataset and train the model. Based on spontaneous fluctuations in the brain, multiple research groups have shown that machine learning can use resting-state MRI data to distinguish patients with cocaine/alcohol/nicotine use disorders from healthy controls and predict withdrawal following treatment with related drugs*(16-20)*. For unsupervised learning, this technique has been used to identify new subtypes of psychiatric disorders based on symptomatology and genetic information. For instance, work by Sun et al.*(21)*. revealed distinct biological subtypes of patients with opioid use disorder by applying a clustering algorithm to heritability. Additionally, as a potential application of machine learning technology, the United States has deployed automated predictive drug monitoring programs (PDMPs) to identify patients with potential for opioid addiction or abuse*(22).* By tracking prescription drug purchase and usage status, PDMPs can predict which patients are at higher risk for prescription drug abuse or overdose and alert authorities or health care providers of potential drug diversion or abuse. Notably, the machine learning parameters used to predict addiction outcomes do not indicate underlying psychological or neural processes, so they cannot be explained mechanistically.

On the other hand, the theory-driven culture of CP is a response to the explanatory gap faced by biological psychiatry and neuroscience. Concepts from biological psychiatry have guided the development of several generations of antipsychotics, antidepressants, and anxiolytics and have benefited a wide range of clinical groups. However, a large explanatory gap exists between the almost irrational effectiveness of psychotropic drugs and the mechanistic understanding of their effects on psychological function*(23)*. For instance, neurotransmitters are understood as chemicals that shuttle information from one neuron to another. These substances, whose drug-induced changes relieve psychiatric symptoms, provide a kind of conceptual leap that seem to bypass the interpretation of the multiple layers of representation that mediate receptor function and behavioral changes. In other words, there is currently a lack of appropriate intermediate levels

of description between the pharmacological level and the patient's cognitive level, and psychiatry needs to build a bridge between the molecular and the phenomenological. Advances in human neuroscience have the potential to bridge parts of the explanatory gap that persists in biological psychiatry. A significant area of progress is the field of decision-making, which is fundamental to the majority of psychiatric conditions*(24)*. A focus on abnormal decision-making provides a unique opportunity to couple cognitive and neural processes in individuals with mental disorders. The computational revolution in cognitive neuroscience underpins this opportunity and lays the groundwork for theory-driven culture.

Theory-driven paradigms are explanatory statistical models, expressed as equations, that are posited as underlying mappings from neural computations to cognitive states. This paradigm often fits the free parameters related to cognitive mechanisms through reinforcement learning algorithms. For instance, the update of action value in the Q-learning model is assumed to track the reward prediction error in the individual decision-making process, and this mechanism has been shown to be mediated by striatal signals that receive dopaminergic projections*(25)*. In this way, patients with mental disorders can observe gaps more easily than healthy populations when interpreting model parameters and can provide cognitive inferences about underlying neurocomputational dysfunction. Animal models of addiction have revealed that addictive behavior is a process of transitioning from operant conditioning to classical conditioning, but how this conceptualization can be generalized from mice to human studies has been a mystery(26). The reinforcement learning theory of addiction provides a unique opportunity to understand the transition from goal-directed to compulsive behavior in addictive behaviors. Specifically, addictive behavior involves a shift in the balance between two key neurological systems. The goal-directed system is flexible and sensitive to changes in outcome values, while the habitual system is connected to cues and is less sensitive to outcomes. Reinforcement learning models explain how drugs act as rewarding reinforcers in the early stages of addiction, allowing the goal-directed system to dominate. With repeated use of drugs, addicted individuals become less sensitive to the negative reinforcing effects of drugs, and the habit system replaces the goal-oriented system, ultimately promoting compulsive behavior*(27)*. In this process, the parameters of reinforcement learning act as a concatenation of cognitive inferences and neural patterns, revealing changes in dominance between the ventral and dorsal striatum that accompany behavioral changes in people with substance disorders. These algorithms, which are used to fit the parameters of cognitive models, shed light on the neurocomputational mechanisms underlying addictive behaviors.

For an explanatory model to be effectively utilized, several prerequisites must be satisfied*(28)*. First, the model should accurately correspond to the true data-generating process, that is, the model's structure and parameters should reflect the underlying biological, psychological, or social processes that produce the observed outcomes. Second, the parameters in the model should have clear, interpretable effects on the model's predictions. They should be able to be identified independently of each other to avoid confounding effects. A common practice to verify these conditions is data simulation, which involves generating data from the model and checking whether the model fitting routine can recover known parameters from that the data. Finally, the model must be statistically tested to determine the reliability of its parameters and predictions, including assessing the statistical significance of the parameters and the confidence intervals of the predictions. The above verification is a necessary but not sufficient condition for reliable explanatory modeling. One potential pitfall in the field is overinterpretation of the results. The best-fitting model is simply the best model tested and its parameters are estimated with associated uncertainties. Importantly, explanatory models are mathematical abstractions of the complex phenomenon of mental disorders, and empirical data are not completely consistent with any candidate model.

Epistemic injustice in computational psychiatry

The conceptual apparatus of epistemic injustice serves as a pivotal tool for the critical examination,

comprehension, and prospective amelioration of prejudiced practices entrenched in the processes of knowledge generation, application, and dissemination, particularly within the domain of computational psychiatry*(29,30)*. Epistemic harms, conceived as ethical transgressions, emerge within these processes and may culminate in epistemic injustices. Such injustices are instantiated when the contributions and insights of individuals from historically underrepresented or marginalized groups are systematically diminished or when their roles as knowledge contributors are unjustly impugned*(5)*. These injustices are further compounded when individuals are iniquitously stripped of the necessary hermeneutical tools to decode and make sense of the world, or when they are precluded from effectively engaging with knowledge that has been formulated in the absence of their experiential input or contextual understanding.

The conceptual architecture for analyzing epistemic injustice was first systematically formulated by the philosopher Miranda Fricker, who distinguishes between the two different forms of epistemic injustice mentioned above: testimonial and hermeneutical*(5)*. Testimonial injustice occurs when a listener prejudicially diminishes the trustworthiness of a speaker's testimony, which is evidenced by actions that silence, depreciate, or distort the speaker's contributions, effectively assigning a deficit of credibility*(5)*. Conversely, hermeneutical injustice arises when individuals or groups encounter difficulties in understanding and communicating their experiences, due to an insufficiency of validated and accessible collective interpretive tools. In other words, the experiences of individuals or groups remain opaque to themselves or to the broader community, due to a lacuna in the collective hermeneutic resources necessary to articulate these experiences*(5)*. Taking the process of conceptualizing "sexual harassment" as an example, before the term existed, women who experienced harassment also tended to experience hermeneutical injustice because of the lack of a public concept to specify this violation. As a subset of medicine, CP is often taken for granted as consistent with both forms of epistemic injustice and even exacerbates them. In this discourse, we contend that computational psychiatry is not immune to these structural and enduring patterns of epistemic exclusion; rather, we argue for nuanced differences in the forms of epistemic injustice that the two cultures manifest in their epistemic practices.

Testimonial injustice in a data-driven culture

The pursuit of a naturalistic understanding of mental disorders in a data-driven culture can be viewed as an epistemic problem. Naturalistic understanding can be defined as the interpretation of medical experiences and variations through the lens of biological norms, focusing on the boundaries between standard functionality and dysfunction*(31)*. The data-driven paradigm is generally considered to follow this standard. This approach positions mental disorders as phenomena that can be systematically classified and understood through machine learning techniques, emphasizing data-driven objective analysis. In addiction research, information from tens of thousands of patients' brains, receptors, and genes is used as a model input to classify and predict diseases. Although researchers aim to reduce mental disorders to a purely biological or neurochemical imbalance, this approach often falls into the trap of dimensionality in the data. Feature selection for machine learning usually involves value judgments. From the perspective of epistemic injustice, the current naturalistic approach to biometric dominance, especially within the supervised learning paradigm of addiction, has been widely criticized. Psychiatric classification endeavors to systematically organize mental disorders, which are characterized by a multifaceted interplay of biological, psychological, and social elements. Unfortunately, the intricate causal relationships and interactions within these disorders are often not fully understood. As a result, most etiological explanations and biomarkers of psychopathology have been hypothetical in nature*(32)*. However, in the context dominated by the biomedical model, alternative understandings other than naturalism are often excluded, marginalized or disparaged. Current naturalistic understandings of CP run the risk of promoting epistemically unjust attitudes, actions, and assumptions, which lead to the belief that examining biological characteristics of disease is the right thing to do scientifically or clinically.

The direct factor in testimonial injustice caused by data-driven culture is naturalistic-dominated medical understanding. Another indirect factor may be the researcher's or physician's assessment of the credibility of patients with mental disorders, which is evident in the practice of addiction medicine. The testimonies of patients with addiction can be viewed in biased way as less credible due to their epistemic status. For instance, in PDMPs, studies have shown that risk scores given by algorithms can disenfranchise patients from care, even when they know they are not addicted to opioids and have never abused drugs*(33)*. This unfair treatment occurs due to undue loss of credibility, which often stems from negative views of their epistemic contributions because of their other shortcomings*(34)*. Even from a third-person perspective, people with mental disorders are sometimes considered to be responsible for these disadvantages, thus exacerbating marginalization. On the other hand, erroneous assessments of epistemic subjects' abilities potentially influence researchers' and doctors' preference for "black box" objective evidence and tend to discount the subjective evidence of conscious experience provided by patients*(30)*. As a result, patients are excluded from the decision-making process as objects of epistemic inquiry rather than beneficiaries of epistemic searches for diagnosis and treatment. This constitutes a form of preemptive testimonial injustice in which the hearer prejudicially assumes that testimony is irrelevant or unreliable even when such testimony is never solicited*(32)*.

Sartre conceptualized illness as a process of moving away from immediate physical experience toward a more reflective and epistemic understanding, culminating in a medical diagnosis*(35)*. This "existence precedes essence" perspective deconstructs illness into four distinct levels. The first three levels are mainly the way patients themselves constitute the "illness", while the fourth level is the physiological conceptualization of "disease" as we know it in medicine. Data-driven classification decisions require value-laden judgment in balancing different risks, where the patient's first-person experience is relevant. Social justice issues arise when the patient's perspective is marginalized because it also has epistemic implications, and incorporating different perspectives is a means of detecting value implications and debating them*(36)*. The naturalistic understanding pursued by the current data-driven culture may be harmful at the level of epistemic practice. Because it fails to represent all relevant perspectives, can result in an epistemic wrong. It is necessary to include the patient's first-person knowledge, which provides an implicit value-laden corrective to the operationalization of mental disorders.

Hermeneutical injustice in theory-driven culture

While the goals of theory-driven models are laudable, the reliability and ethical significance of their explainability remain to be evaluated, especially given the gap in collective hermeneutic resources (i.e., the epistemic and linguistic resources that social members use to understand and communicate about the world)*(5)*. Taking the aforementioned reinforcement learning model of addiction as an example, addiction is further explained as a process in which goal orientation is transformed into habit control. As an appropriate language and conceptual tool, addiction has become an indispensable hermeneutic resource in scientific research and clinical work and is widely comprehensible to social collectives*(26)*. This concept involves well-defined criteria and has an objective pathophysiology definition. However, the problem with the reinforcement learning model is that the extended concept of addiction is defined in terms of metrics that are not shared. Consideration of some parameters in the model, such as thresholds that allow systems to define the degree to which behavior is quantified as goal-directed or habit-controlled, has not been explicit and transparent to the stakeholders directly involved in and affected by these systems. The implications of the explanatory mechanisms provided by the model may therefore deviate significantly from widely shared trajectories, and stakeholders are unlikely to modify this shift. Especially when explaining certain aspects of the disease, patients are more inclined to use lay terms to express the harm they experience subjectively*(37)*. However, this information with epistemic value may be excluded from the knowledge production process by models because

it is more difficult to quantify than information expressed in biomedical terms. Hermeneutical injustice occurs when individuals do not have the hermeneutical resources to understand or accept the conceptual meaning of habitual control and relate it to their experiences*(5).*

Systemic problems related to the unequal participation of people with substance use disorders in the production of shared hermeneutical resources may involve two factors. First, the intrinsic contribution condition is a certain degree of the inability to share experiences (i.e., communication difficulties) that is inherent in the process of transmitting subjective experiences by people with substance use disorders*(5).* Second, and more importantly, the decisive role of the reinforcement learning model in shaping shared hermeneutical resources is an unwarranted epistemic privilege*(38)* Distinct from the epistemic privileges necessary for physicians in the psychiatric diagnostic process, the epistemic privileges accorded to models are akin to standardized protocols that strictly limit the testimony of people with substance use disorders to their subjective experiences. These protocols significantly influence which forms of knowledge are acknowledged and used in decision-making*(39).* This situation becomes particularly problematic if the rigor of these protocols leads subjective experience to be treated as an illegitimate source of knowledge, especially when these experiences cannot be easily quantified. Standardized protocols may acquire epistemic privilege devoid of human intervention and become a key determinant of epistemic engagement. The ultimate goal of the reinforcement learning model is to find the best parameters to quantify the degree of transition from addiction to habit control, but an extreme case may be that the model itself is distorted in its definition of addictive behavior. This means that as long as the model free parameters can be shown to be optimal during the fitting process and show habit control and brain dopamine increases in people with substance use disorders, the model is considered valid. Although this is a value-based choice, it does not suggest that individuals with substance use disorders are actually at risk for compulsive behavior. Thus, when it is impossible to define whether a reinforcement learning model follows the implications of addiction itself, the model takes unwarranted epistemic privilege and does not allow the results to be disputed.

A theory-driven culture provides information knowledge in a one-way manner in clinical decision-making in psychiatry, thereby creating a climate conducive to hermeneutical injustice. To avoid bias resulting from hermeneutical resources being dominated by model assumptions, potential solutions may be appropriate supervision strategies for reinforcement learning algorithms and receiving valuable feedback information from users. Such supervision should track the logic of the algorithm to make corrections to parameter weights and compare the results of algorithms trained on groups of different addiction subtypes. Additionally, user experience should be incorporated into the knowledge-generation process as the starting point for the theoretical mechanism constructed by the model. Some innovative pilot studies are currently considering incorporating the above measures into theory-driven computational models, such as measuring and modeling momentary subjective feelings during decision-making to elucidate the affective processes that are influenced by mental disorders*(40, 41).* Such measures are necessary to maximize interpretability and reduce decision bias, but it is currently unclear whether model algorithms and patient experience are indeed related, and the relevant evaluation mechanisms remain to be verified.

Conclusion

The widespread use of CP in scientific research and clinical decision-making may carry the risk of context-specific epistemic injustice. Currently, CP is divided into two cultures based on research orientation: data-driven and theory-driven. Considering the corresponding targets of each culture, we show the nuances of the forms in which epistemic injustice manifests in both. The former suffers from testimonial injustice due to the dominance of naturalism, while the latter suffers from hermeneutical injustice due to the un-shareability of resources. It should be noted that the two are not parallel but are intertwined. When testimonial injustice occurs

repeatedly due to the researcher's naturalistic preferences, epistemic subjects may gradually lose epistemic self-trust. In the long run, this process causes a lack of conceptual resources and a hermeneutical marginalization of patients. We argue that there is a need to take a critical look at how computational models compare to humans in terms of the accuracy, reliability, and interpretability of knowledge representation. It is even more important to encourage greater inclusion of patients and advocates in the knowledge generation process, where their first-person testimonies serve as a valuable corrective to reduce the risk of epistemic loss.

Acknowledgments

This study was supported by the Chinese Academy of Sciences Talent Project (USTC-BR-2022-08).

Authors' contributions

Min Wang wrote the first draft of the manuscript. Zhoukang Wu and Liangjiecheng Huang involved in the conception of the idea. Xiaochu Zhang and Xiaosong He edited the manuscript. All authors contributed to and approved the final manuscript.

Conflicts of interest

The authors declare that no competing interests exist.

References

1. Montague PR, Dolan RJ, Friston KJ, Dayan P. Computational psychiatry. *Trends Cogn. Sci.* 2012; 16(1): 72-80.
2. Starke G, De Clercq E, Borgwardt S, Elger BS. Computing schizophrenia: ethical challenges for machine learning in psychiatry. *Psychol. Med*. 2021; 51(15): 2515-21.
3. Wiese W. From the Ethics of AI to the Ethics of Consciousness: Ethical Aspects of Computational Psychiatry. *Psychiatr. Prax.* 2021; 48: S21-S5.
4. Wiese W, Friston KJ. AI ethics in computational psychiatry: From the neuroscience of consciousness to the ethics of consciousness. *Behav. Brain. Res.* 2022; 420:12.
5. Fricker M. *Epistemic injustice: Power and the ethics of knowing*: Oxford University Press; 2007.
6. Bennett D, Silverstein SM, Niv Y. The Two Cultures of Computational Psychiatry. *JAMA Psychiatry.* 2019; 76(6): 563-4.
7. Huys QJM, Browning M, Paulus MP, Frank MJ. Advances in the computational understanding of mental illness. *Neuropsychopharmacology*. 2021; 46(1): 3-19.
8. Maia TV, Frank MJ. From reinforcement learning models to psychiatric and neurological disorders. *Nat. Neurosci.* 2011; 14(2): 154-62.
9. Eronen MI. The levels problem in psychopathology. *Psychol. Med.* 2021; 51(6): 927-33.
10. Breiman L. Statistical modeling: The two cultures. *Stat. Sci.* 2001; 16(3): 199-215.
11. Huys QJM, Maia TV, Frank MJ. Computational psychiatry as a bridge from neuroscience to clinical applications. *Nat. Neurosci.* 2016; 19(3): 404-13.
12. Zhou ZR, Wu TC, Wang BK, Wang HY, Tu XM, Feng CY. Machine learning methods in psychiatry: a brief introduction. *Gen. Psychiat.* 2020; 33(1): 3.
13. Jiang T, Gradus JL, Rosellini AJ. Supervised Machine Learning: A Brief Primer. *Behav. Therapy.* 2020; 51(5): 675-87.
14. Wang XJ, Krystal JH. Computational Psychiatry. *Neuron*. 2014; 84(3): 638-54.
15. Mak KK, Lee K, Park C. Applications of machine learning in addiction studies: A systematic review. *Psychiatry Res.* 2019; 275: 53-60.
16. Lichenstein SD, Scheinost D, Potenza MN, Carroll KM, Yip SW. Dissociable neural substrates of opioid and cocaine use identified via connectome-based modelling. Mol Psychiatr. 2021; 26(8): 4383-93.
17. Yip SW, Scheinost D, Potenza MN, Carroll KM. Connectome-Based Prediction of Cocaine Abstinence. *Am. J. Psychiat.* 2019; 176(2): 156-64.
18. Ding XY, Yang YH, Stein EA, Ross TJ. Combining Multiple Resting-State fMRI Features during Classification: Optimized Frameworks and Their Application to Nicotine Addiction. *Front. Hum. Neurosci.* 2017; 11: 14.
19. Mete M, Sakoglu U, Spence JS, Devous MD, Harris TS, Adinoff B. Successful classification of cocaine dependence using brain imaging: a generalizable machine learning approach. *BMC Bioinformatics*. 2016; 17: 13.
20. Rish I, Bashivan P, Cecchi GA, Goldstei RZ, editors. *Evaluating Effects of Methylphenidate on Brain Activity in Cocaine Addiction: A Machine-Learning Approach.* SPIE Biomedical Applications in Molecular, Structural and Functional Imaging Conference; 2016 Mar 01-03; San Diego, CA. BELLINGHAM: Spie-Int Soc Optical Engineering; 2016.
21. Sun JW, Bi JB, Chan G, Oslin D, Farrer L, Gelernter J, et al. Improved methods to identify stable, highly heritable subtypes of opioid use and related behaviors. *Addict Behav.* 2012; 37(10): 1138-44.
22. Oliva JD. Dosing Discrimination: Regulating PDMP Risk Scores. *Calif. Law Rev.* 2022; 110(1): 47-115.
23. Murray JD, Demirtas M, Anticevic A. Biophysical Modeling of Large-Scale Brain Dynamics and Applications for Computational Psychiatry. *Biol. Psychiat-Cogn. Neurosci. Neuroimag.* 2018; 3(9): 777-87.
24. Paulus MP. Decision-making dysfunctions in psychiatry - Altered homeostatic processing? *Science.* 2007; 318(5850): 602-6.
25. Daw ND, Gershman SJ, Seymour B, Dayan P, Dolan RJ. Model-Based Influences on Humans' Choices and Striatal Prediction Errors. *Neuron.* 2011; 69(6): 1204-15.
26. Luscher C, Robbins TW, Everitt BJ. The transition to compulsion in addiction. *Nat. Rev. Neurosci.* 2020; 21(5): 247-63.
27. Vandaele Y, Ahmed SH. Habit, choice, and addiction. *Neuropsychopharmacology.* 2021; 46(4): 689-98.
28. Wilson RC, Collins AGE. Ten simple rules for the computational modeling of behavioral data. *eLife.* 2019; 8: 33.
29. Kidd IJ, Spencer L, Carel H. Epistemic injustice in psychiatric research and practice. *Philos Psychol.* 2022: 29.
30. Drozdzowicz A. Epistemic injustice in psychiatric practice: epistemic duties and the phenomenological approach. J. *Med. Ethics*. 2021; 47(12): 5.

31. Boorse C. Health as a theoretical concept. *Philos. Sci.* 1977; 44(4): 542-73.
32. Bueter A. Epistemic Injustice and Psychiatric Classification. *Philos. Sci.* 2019; 86(5): 1064-74.
33. Pozzi G. Testimonial injustice in medical machine learning. *J. Med. Ethics.* 2023; 49(8): 536-40.
34. Crichton P, Carel H, Kidd IJ. Epistemic injustice in psychiatry. *BJPsych bulletin.* 2017; 41(2): 65-70.
35. Sartre J-P. *Being and nothingness: An essay in phenomenological ontology*: Taylor & Francis; 2022.
36. Longino HE. Science as social knowledge*: Values and objectivity in scientific inquiry*: Princeton university press; 1990.
37. Pot M. Epistemic solidarity in medicine and healthcare. *Med. Health Care Philos.* 2022; 25(4): 681-92.
38. Pozzi G. Automated opioid risk scores: a case for machine learning-induced epistemic injustice in healthcare. *Ethics Inf. Technol.* 2023; 25(1): 12.
39. Moes F, Houwaart E, Delnoij D, Horstman K. Questions regarding 'epistemic injustice' in knowledge-intensive policymaking: Two examples from Dutch health insurance policy. *Soc. Sci. Med.* 2020; 245: 9.
40. Kao CH, Feng GW, Hur JK, Jarvis H, Rutledge RB. Computational models of subjective feelings in psychiatry. *Neurosci. Biobehav. Rev.* 2023; 145: 14.
41. Gu XS, FitzGerald THB, Friston KJ. Modeling subjective belief states in computational psychiatry: interoceptive inference as a candidate framework. *Psychopharmacology.* 2019; 236(8): 2405-12.

Received: 22 December 2023
Accepted: 16 January 2024