CURRENT THEMES IN THE PHILOSOPHY OF PSYCHIATRY

(forthcoming in *Crítica*, vol. 56, no. 167, 2024)

Federico Burdman

Universidad Alberto Hurtado, Chile

fgburdman@uba.ar

https://orcid.org/0000-0002-8506-9271

Philosophy has always been concerned with mental health and mental disorder, and many prominent historical figures have held views that have drawn attention to these issues. In recent years, however, there has been a surge of interest in philosophical issues surrounding psychiatry. As a result, the philosophy of psychiatry has emerged as a distinct, well-established field of inquiry.

In the current state of play, conceptual work on the theory of mental health and disorder remains a central issue, with many views attempting to move beyond the apparent impasse between naturalistic and normativist approaches. Other classical themes also remain central, such as the conceptual and methodological problems of psychiatric classifications, including the question of what sort of kinds mental disorders are, and questions about the validity and utility of standard nosological constructs. The proper place of psychiatry in relation to (the rest of) medicine and its adherence to the 'medical model' also continues to be hotly debated, as does the question of the relationship between mental disorders and brain disorders.

The most striking development in recent years, however, has been a significant broadening of the range of topics that are now the subject of lively philosophical investigation. The seemingly distinctive nature of causal and mechanistic explanations in psychiatry has come under close scrutiny. Extensive work is being done on the

phenomenology of various mental conditions, especially (though not exclusively) from an enactive perspective. We also find a renewed awareness of theoretical problems related to the therapeutic encounter and patient-therapist relationships, as well as the complexity of patient identities, with implications for how we think about the ethics of care, the ideal of patient autonomy, the capacity to consent to treatment, and other pressing ethical issues. The use of newly available technologies in the context of research and care is at the center of several important debates. Much attention has also been given to the implications of mental disorder for fitting attributions of moral responsibility. Beyond such general issues, which speak to problems concerning the very notion of mental disorder or the practice of psychiatry as a whole, philosophers have also become much more involved in theorizing about particular psychiatric conditions, drawing attention to conceptual problems both in the definition of particular diagnostic labels and in the most widely accepted scientific approaches to their etiology and treatment.

This special issue is a testament to the breadth of this burgeoning field. The papers in this issue can be seen as a representative —though, of course, not exhaustive—sample of the diversity of topics, as well as the range of theoretical and methodological approaches, that characterize current work in the philosophy of psychiatry. Walter Sinnott-Armstrong, Laura Soter, and Jesse Summers focus on the problem of the unity of mental disorders (i.e., what are the unifying features of the conditions commonly considered to belong to this domain?) and propose a novel account of mental disorders as failures of attention. Virginia Ballesteros and Ana Batalla discuss the harmful dysfunction analysis of mental disorder and present an objection to standard formulations of the harm criterion, inspired by consideration of the case of philosopher Jean Améry, a Holocaust survivor who resisted the medicalization of his condition on moral grounds. Michelle Maiese draws attention to an equally general issue, bringing an enactive perspective to account for the distinctive contribution of patient expertise in the context of therapeutic encounters. The remaining three papers in the special issue focus on philosophical issues related to particular psychiatric conditions. Alice Kelley discusses negative and positive looping effects following the introduction of the new diagnosis of Prolonged Grief Disorder. Quinn Hiroshi Gibson looks at theories of addiction with an eye on integrating neural and agent-level stories about the central features of the condition. And Pablo López-Silva and Emmanuel

Méndez discuss a problem related to the (apparent) failure of self-adscription of mental states that occurs in delusions of thought insertion in the context of schizophrenia.

A brief look at the main ideas developed in each of these articles will help to give a sense of the overall shape of the special issue.

The paper by Walter Sinnott-Armstrong, Laura Soter, and Jesse Summers addresses the central question: What is mental disorder? Their starting point is the most recent definition of the term in the DSM, according to which mental disorders involve significant disturbances in cognition, emotion, or behavior (American Psychiatric Association, 2022, p. 14). The disjunctive nature of this definition, the authors note, is not sufficiently informative as to what the conditions listed in the manual have in common. The paper is a bold attempt to provide a novel answer to this question. The key idea is to think of mental disorders as intimately related to failures of attention. The first part of the paper is devoted to detailing the conceptual contours of the proposal. The second part illustrates how the approach can be applied to several standard diagnostic labels, each representative of a major nosological category.

Sinnott-Armstrong, Soter, and Summers begin by presenting their conceptual machinery. To attend to something, according to a common understanding going back to William James's *Principles of Psychology*, is to think about it or experience it to the partial or total exclusion of other things. Defined in this way, attention is commonly thought of as a capacity that is fundamentally in the realm of cognition and/or perception. However—and again following James—, Sinnott-Armstrong, Soter, and Summers move beyond this restrictive picture to highlight a "consistent and reliable" connection between attention and agency: what we pay attention to when confronted with a situation plays a crucial role in shaping how we will respond. Moreover, the deployment of attention is also critical for fine-tuning our performance. Thus, the reasons we have to act one way or another translate into reasons to attend to one aspect of a situation or another.

This agential framing of attention underscores a crucial connection between attention and reasons. On the approach proposed by Sinnott-Armstrong, Soter, and Summers, agents have reason to attend to things when doing so is likely to contribute to patterns of thought and action that are beneficial to the agent. Of course, someone may

have reasons to attend to one thing and, at the same time, reasons to attend to other things that are incompatible with the first. But reasons can be weaker or stronger, so there will typically be one thing that can be said to be the thing that the agent has most reason to attend to. The account is meant to be neutral between an objective and a subjective reading of what it means for an agent to have reason to attend to something, provided that the relevant reasons have at least some psychological grip on the agent.

In line with this framework, a failure of attention is a failure to distribute attention in the way one has most reason to. This can be either by failing to pay attention to what one has most reason to attend to, or by failing to shift attention away from things one doesn't have most reason to attend to. It can result from a failure of ability (when the agent is unable to distribute her attention in accordance with the relevant reasons) or a failure of tendency (when the agent repeatedly fails to do this successfully, even though she scores well on standardized tests of attentional capacity). In the authors' view, such failures may, in some cases, reflect a failure of bottom-up attentional mechanisms. However, the sort of failure that is central to mental disorders usually involves top-down exercises in attentional deployment. Of course, not all failures of attention are associated with mental disorder. They are, however, when the failure in question is persistent and harmful enough to be clinically significant.

Sinnott-Armstrong, Soter, and Summers's view is explicitly not meant to account for *all* conditions that go under the name of mental disorder. They are, however, committed to the claim that a sufficiently wide range of mental disorders can be meaningfully thought of as centrally involving failures of attention. The second part of the paper looks at several particular conditions to illustrate how the proposed view can shed light on standard diagnostic labels. It focuses on the three main constructs in the DSM definition (cognition, emotion, and behavior), to show how failures to sustain attention or to keep attention away from something are central to disturbances in each of these domains. The discussion focuses on ADHD and anorexia nervosa, phobias and depression, narcissism and delusions, addiction, OCD, and psychopathy. In the final section, the authors briefly discuss some implications of their view, including the promising possibility that techniques for

overcoming failures of attention can be the object of training to improve patients' treatment prospects.

The paper by Virginia Ballesteros and Ana Batalla engages with the well-known and highly influential Harmful Dysfunction Analysis of mental disorder, first introduced by Jerome Wakefield in (1992) and since then refined and developed in an extensive series of publications. In brief, the HDA proposes to think of mental disorders as conditions that are the result of dysfunctional processes and lead to harmful consequences. As Ballesteros and Batalla note, much more attention has been paid to the question of what is the relevant sense of dysfunction than to the seemingly simpler question of how to characterize the harm that is a central feature of disorder. One lesson from the debates over the exclusion of homosexuality from psychiatric nosology in the 1970s is that not just any kind of harm is relevant to mental disorder. Harm that is a proper consequence of mental disorder needs to be distinguished from the kind of harm that is secondary to modifiable societal responses to certain people or certain conditions. Ballesteros and Batalla propose an analogous argument that focuses on a different dimension of harm, namely moral value. Their proposal stems from the thought that a satisfactory rendering of the normative element in hybrid theories such as Wakefield's must accommodate the intuition that labeling a condition a disorder implies that it is disvalued. They then argue that certain conditions appear to satisfy the harm criterion and yet are morally valuable in a way that makes it intuitively wrong to see them as proper instances of mental disorder. Standard formulations of the harm criterion, they conclude, cannot capture the full complexity of the normative dimension of mental disorder and therefore need to be revised or supplemented.

Ballesteros and Batalla build their argument around the case of philosopher Jean Améry. Améry, himself a theorist of mental disorders, was a Holocaust survivor. The depth of the horror he experienced in the concentration camps left a profound mark on him, which manifested itself in a set of 'symptoms' prima facie amenable to PTSD. Améry, however, resisted the medicalization of his condition. Importantly, he did not mean to deny that there was something dysfunctional about it, nor that he suffered greatly from it. Améry's point, which Ballesteros and Batalla make their own, was that his clinging to the past and his refusal to 'heal' and move on represented a *moral* stance. Moreover, he claimed that this

stance was valuable both to him as a survivor and to society at large, since the continuing pain of the victims was a living testimony to the horrors they had endured and to the need for reparation and justice that, he felt, had not been fully served in post-war Germany. Recognition of the value of this moral stance, Ballesteros and Batalla argue, would be foreclosed by portraying it as the result of a medical condition. If this is correct, then Améry's case illustrates that a condition can meet both the dysfunction criterion and the harm criterion and still not be properly considered a mental disorder.

If standard formulations of the harm criterion do not pass this test, then we need a different articulation of the normative element in the theory of mental disorder. In their final section, Ballesteros and Batalla explore a possible solution to this problem. In their view, harm may be part of the picture, but it is not the whole picture. A more promising approach, they suggest, is to cast the normative criterion in terms of human flourishing and the ability to lead a meaningful life. Under this approach, a condition such as PTSD can be properly viewed as a disorder if it has a negative impact on an individual's ability to live a meaningful and flourishing life. A crucial caveat, however, is that on the 'non-essentialist' reading they propose, meaning and flourishing should be seen as to some extent relative to the individual's perspective. The key to Améry's case, they argue, is that the most meaningful life for him was one in which harmful and dysfunctional psychological processes played a key role.

The paper by Michelle Maiese is not about the theory of mental disorder, but about an issue that is equally relevant to the field of psychiatry as a whole: the dynamics of patient-doctor relationships. The traditional way of thinking about these relationships is arguably in relatively unidirectional and paternalistic terms. In such a view, the clinician possesses knowledge that is both necessary and sufficient to diagnose and impart treatment directives, while the patient's role is conceived as essentially passive. Many have suggested that such a view is flawed on several accounts. One might think, for instance, that placing patients in a fundamentally passive role is detrimental to treatment effectiveness, or that it is crucial to approach the patient-doctor relationship in a way that does not undermine patient autonomy and dignity. Maiese's paper articulates a different though related insight, namely that the traditional view of the clinical encounter, which places knowledge squarely

on the side of the physician, has no place for patients' expertise about their condition. The paper develops the concept of patient expertise and highlights the crucial role it can play in constructive therapeutic relationships, resulting in both epistemic and agential benefits that are critical to positive clinical outcomes. Incorporating patient expertise into decision-making processes, Maiese argues, is a way to take advantage of patients' unique epistemological position while also promoting their autonomy, itself an important goal of treatment.

On Maiese's view, patients are experts in several interrelated ways. They have first-hand knowledge of what symptoms feel like and how they affect their lives—knowledge that is typically forever beyond the experiential reach of treating physicians. In addition, they have a uniquely intimate knowledge of their medical history, particularly what types of treatments have worked in the past. More broadly, patients know what it is like to use mental health services and what it is like to experience the broader social repercussions of mental disorder (often in the form of stigmatizing social attitudes) that are to some extent elusive from the clinician's perspective. Furthermore, unlike clinicians, patients' knowledge is holistic in that it encompasses the various dimensions of mental disorders (phenomenological, behavioral, social, cultural, and physiological).

Maiese's account of patient expertise motivates the search for a new approach to patient-doctor dynamics. In this context, the central move in Maiese's proposal is to frame the understanding of patient-doctor relationships using the conceptual toolkit of enactivism. Key to this project is the notion of *participatory sense-making* (De Jaegher & Di Paolo, 2007). This makes for a conceptually well-developed way of thinking about clinical relationships as contexts in which new insights emerge from coordination and negotiation among participants. Importantly, this way of framing patients' epistemic contribution to the joint patient-therapist effort does not require one to conceive of doctor-patient reciprocity in symmetrical terms, as a "like for like" exchange. Moreover, and crucially, participatory sense-making does not require individuals to give up their autonomy. On the contrary, Maiese emphasizes how the kind of relationship she describes fosters patient autonomy. Philosophically, this is tantamount to giving pride of place to the relational dimension of autonomy over the kind of autonomy that focuses on an individualistic perspective.

The remaining three papers in the special issue break the pattern of discussing problems relevant to psychiatry as a whole and focus on philosophical issues related to particular psychiatric conditions.

Alice Kelley brings a fresh perspective on the introduction of the diagnosis of Prolonged Grief Disorder (PGD) in the DSM-5-TR (American Psychiatric Association, 2022). The inclusion of this new diagnostic label has been the subject of much debate. Is PGD sufficiently distinct from MDD to warrant a new label? How does one distinguish normal-range grief from the sort of excessive grief that can be clinically significant? Is PGD an etiologically valid entity, or is it merely a shorthand to refer to a group of cases with certain similar characteristics that does not track any real disorder type out there in the world?

Kelley's paper discusses another set of concerns, surrounding the potentially troubling effects that the introduction of this new diagnosis may have on the experiences of those who are grieving themselves. Several years ago, Ian Hacking drew attention to an intriguing feature of mental disorder classifications: receiving a diagnosis can significantly alter the experience of diagnosed individuals, possibly leading to changes in their feelings and behavior as a result of being classified as falling under a particular diagnostic type. Over time, this can lead to the need to introduce new modifications to the classification in order to keep up with the newly emerging characteristics of the phenomenon that is the object of the classification, the characteristics of which are then *partly* an effect of the use of the classification tool itself (Hacking, 1999). Hacking coined the term *looping effects* to refer to these complex dynamic interactions. When it comes to PGD, some have expressed concern that the medicalization of grief through this new diagnosis may create looping effects that are fundamentally detrimental to people struggling to overcome grief, based on how viewing their grief through the lens of clinical psychiatry may lead grievers to adopt a more passive stance and identify themselves as having a medical condition. Furthermore, conceptualizing their experience through a psychiatric label may fundamentally alter the profound significance of the experience for the griever.

Kelley explores another possibility: that the inclusion of PGD may generate looping effects that are positive in nature. First, she argues that in assessing whether the effects of

introducing a new diagnostic category are positive or negative, we should measure them against the alternative of not introducing that category. She notes that one likely result of not having a diagnostic label that serves the purpose of identifying clinically problematic grief is that people experiencing the relevant symptoms are likely to receive another, less specific diagnosis —perhaps MDD or PTSD— rather than not having their experience medicalized. To the extent that one worries about the potential for clinical diagnosis to distort the experience of grievers, less specific diagnostic labels (paired with less specific treatment strategies) may make for a worse, not better, scenario. Second, Kelley suggests that there are ways in which the introduction of PGD may actually lead to properly positive looping effects, not just be less harmful than available alternatives. Viewing their experiences through the lens of the PGD label may, in some cases, promote a healthy engagement with grief. Kelley's argument exploits a further noteworthy observation, namely that the looping effects that result from the medicalization of a particular condition depend in part on one's view of medicine and medical practice. She suggests that if we think of health and disease as *institutional* concepts (Kukla, 2014), this can provide patients diagnosed with PGD with a valuable tool for engaging with their condition in an authentic way. In this way, the argument indirectly provides support for the institutional framework itself.

Quinn Hiroshi Gibson's paper focuses on addiction. He addresses what he calls the *integration problem*: how to bring together the best available scientific knowledge about the mechanisms behind addiction —mostly couched in the subpersonal language of neuroscience— and the personal-level understanding of the condition with which folk psychology operates. The task is important, among other things, because what matters most to us from an ethical point of view concerns personal-level phenomena. The problem arises because the scientific and manifest images, though overlapping to some extent, use conceptual frameworks whose relationship to each other is far from transparent. The issue is not unique to addiction, of course, and Gibson's discussion of it may hold lessons relevant to psychiatry more broadly.

Gibson's strategy is to start with a consideration of scientific evidence on addiction and try to work his way up to the realm of moral psychology. The first result of this

endeavor is to narrow the field of theoretical options. An extreme compulsion view of addiction is inconsistent with evidence suggesting that, under certain conditions (e.g., given the right sort of incentive structure), people with addiction can refrain from using. An extreme pharmacological view is also inconsistent with epidemiological data and evidence from animal studies. Furthermore, a consideration of recent advances in neuroscience suggests that an understanding of addiction that has no place for physiological factors is equally untenable. In this context, Gibson discusses two of the most influential theories in the neuroscience of addiction: incentive sensitization theory (e.g., Robinson & Berridge, 2008) and prediction error theory (e.g., Redish, 2004). Both theories offer different explanations of the way in which the processing of the reward value of drugs is anomalous in addiction. An important contribution of Gibson's paper is to argue that the explanations offered by the two theories are not incompatible, first appearances notwithstanding. Prediction error theory focuses on representational dysfunction, while incentive sensitization theory emphasizes motivational dysfunction, but the insights of the two theories are fundamentally complementary. The upshot is that to make progress on the integration problem, we need a theory of the psychology of addiction that can strike the right balance between the representational and motivational aspects of the condition.

Agent-level theories of addiction have often focused on the motivational side of the coin, and many are framed fundamentally as theories of desire. Gibson discusses some of the problems with such theories before offering a theory of his own. The view he puts forward—an innovative contribution of this paper— focuses on what he calls *hybrid intentions*, a sui generis kind of motivational state that combines features we usually find in desires and intentions. Hybrid intentions seem to have the sort of close connection to action that ordinary intentions have, even though they are not formed in the usual way and seem to "merely assail" people with addiction. "The result", as Gibson puts it, "is a state which drives action in accordance with one's internal states, but which is not responsive to the will". Importantly, this motivational story is consistent with the insight that representational dysfunction plays a crucial role in addiction: the computational defects described by prediction error theory are mapped onto the personal level under the guise of unstable representations and cognitive distortions. In the final sections of his paper, Gibson

discusses the implications of his view for questions of moral responsibility, suggesting that these are more aptly framed if we focus on attributability rather than accountability.

The paper by Pablo López-Silva and Emmanuel Méndez focuses on the puzzling phenomenon of delusions of thought insertion in the context of schizophrenia. People with delusions sometimes report that thoughts occur in their minds that are not their own but were inserted there by an external agent. The precise etiology, phenomenology, and clinical significance of thought insertion are all topics of lively debate (see López-Silva & McClelland, 2023). In this paper, López-Silva and Méndez address a different issue, namely whether thought insertion can be aptly considered as a counterexample to the widely held claim that it is simply not possible to be mistaken about which thoughts are one's own. Such an idea is of broadly Wittgensteinian inspiration, but its best-known formulation is due to a classic paper by Sydney Shoemaker, who called it 'immunity to error through misidentification' (Shoemaker, 1968). Years later, in another famous paper, John Campbell discussed Shoemaker's proposal in the light of delusions of thought insertion in the context of schizophrenia (Campbell, 1999). López-Silva and Méndez take up this classic debate and argue that thought insertion (as discussed by Campbell) does not falsify Shoemaker's principle.

López-Silva and Méndez make two distinct points in support of this claim. First, they reconstruct the history of the controversy, identifying precisely the views held by both Shoemaker and Campbell. The kinds of immunity to error discussed by Shoemaker and Campbell, they argue, differ in important ways. On their reading, Campbell's discussion of immunity to error cannot be a rebuttal of Shoemaker's view, simply because the claim that Campbell discusses differs in certain key respects from the one put forward by Shoemaker. Their second move is to distinguish two different elements in the self-adscription of mental states. One concerns the location, so to speak, where the relevant states occur. In the context of thought insertion reports, the relevant thoughts are said to occur in the mind of the reporter. The second element concerns ownership per se, which is what reporters deny in cases of thought insertion, attributing the relevant thoughts to someone else as their owner. However, insofar as the relevant thoughts are said to occur in the reporter's mind,

López-Silva and Méndez argue that they still represent a partially successful case of self-adscription of mental states, albeit an atypical one.

References

American Psychiatric Association. (2022). *Diagnostic and statistical manual of mental disorders: DSM-5-TR*. American Psychiatric Association.

Campbell, J. (1999). Schizophrenia, the Space of Reasons, and Thinking as a Motor Process. *Monist*, *82*(4), 609–625. https://doi.org/10.5840/monist199982426

De Jaegher, H., & Di Paolo, E. (2007). Participatory sense-making: An enactive approach to social cognition. *Phenomenology and the Cognitive Sciences*, *6*(4), 485–507. https://doi.org/10.1007/s11097-007-9076-9

Hacking, I. (1999). Madness: Biological or Constructed? In I. Hacking, *The Social Construction of What?* Harvard University Press.

Kukla, R. (2014). Medicalization, "normal function", and the definition of health. In J. D. Arras, R. Kukla, & E. Fenton (eds.), *Routledge companion to bioethics* (pp. 515–530). Taylor and Francis.

López-Silva, P., & McClelland, T. (2023). *Intruders in the Mind: Interdisciplinary Perspectives on Thought Insertion*. Oxford University Press.

Redish, A. D. (2004). Addiction as a Computational Process Gone Awry. *Science*, *306*(5703), 1944–1947. https://doi.org/10.1126/science.1102384

Robinson, T., & Berridge, K. (2008). The incentive sensitization theory of addiction: Some current issues. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1507), 3137–3146. https://doi.org/10.1098/rstb.2008.0093

Shoemaker, S. S. (1968). Self-Reference and Self-Awareness. *The Journal of Philosophy*, *65*(19), 555. https://doi.org/10.2307/2024121

Wakefield, J. C. (1992). The concept of mental disorder: On the boundary between biological facts and social values. *American Psychologist*, *47*(3), 373–388. https://doi.org/10.1037/0003-066X.47.3.373