Artificial Intelligence will eventually replace psychiatrists

Journal:	*BJPsych*
Manuscript ID	BJPsych-19-0824
Manuscript Type:	In Debate
Date Submitted by the Author:	11-Sep-2019
Complete List of Authors:	Brown, Christian; SWLStG Mental Health NHS Trust, Psychiatry Story, Giles; Max Planck UCL Centre for Computational Psychiatry and Ageing Research, Psychiatry; Barnet Enfield and Haringey Mental Health NHS Trust, Psychiatry Mourão-Miranda, Janaina; UCL, Centre for Medical Image Computing, Department of Computer Science Baker, Justin; McLean Hospital, Department of Psychiatry
Keywords:	Artificial Intelligence, Information Technologies, Silicon Valley, Technology, Biomarkers

Title: Debate: Artificial Intelligence will eventually replace psychiatrists

Author Names: Christian Brown, Giles William Story, Janaina Mourão Miranda, Justin Taylor Baker,

Declaration of Interest: J.T.B. has declared the following in addition: *J.T.B. has received consulting fees from Pear Therapeutics, BlackThorn Therapeutics, Niraxx Therapeutics, and AbleTo, Inc.*

Corresponding Author: Christian Brown, cp.brown@nhs.net, British Journal of Psychiatry, 21 Prescot Street, London, E1 8BB.

Introduction

For

Brown C [1], Story GW [2,3].

1. Specialty Trainee in Forensic Psychiatry, South West London and St George's Mental Health NHS Trust, UK

2. Max Planck UCL Centre for Computational Psychiatry and Ageing, Russell Square House, 10-12 Russell Square, London, WC1B 5EH

3. Barnet, Enfield and Haringey NHS Mental Health Trust, St Ann's Hospital, St Ann's Road, London, N15 3TH

We are told that Artificial Intelligence (AI) could replace most, if not all, human jobs.[1] Here we argue that psychiatry is far from immune to such a coup. We will consider our looming obsolescence in two domains: technical and relational.

It is uncontentious to assert that AI will transform the *technical* aspects of psychiatry. For example, smartphone technology enables the capturing of rich, longitudinal, multi-modal data, the analysis of which promises vastly improved characterisation of illnesses and their trajectories.[2] Such methods have already shown predictive potential in forecasting relapse in bipolar disorder.[3,4,5] Coupled with improved models of treatment efficacy [6], and increasingly naturalistic data-driven taxonomies of mental illness,[7] it seems likely that computers' technical mastery in diagnosis and treatment planning will soon out-do that of humans.

However, psychiatry offers more than the sum of its technicalities. Clinical exchanges in psychiatry involve a dynamic interplay of facts and values, through which patients might find relief in feeling listened to and understood. Trust within the ensuing relationship further enhances treatment effects.[8] Might this richness of communication prove beyond the abilities of AI? If so, could the human psychiatrist survive as skilled mediator between patient and machine – a human front-end to an AI operating system? In our view there are good reasons to believe that AI-led care, even in psychiatry's most relational aspects, could ditch the human middleman.

AI is not yet able to converse with sufficient flexibility to hold a psychiatric interview. However natural language processing is rapidly advancing and conversational agents have already found application in assessing alcohol-drinking habits.[9] Indeed, there is good evidence to suggest that people can build therapeutic bonds with AI agents. For instance, human minds naturally infer emotion and intentionality from limited data-points,[10,11] derive comfort from [12] and even empathise with [13] minimally-animate objects. Evidence suggests that people can be more honest with computers than they are with humans.[14] The intelligent clinician-in-the-app might even foster trust by accurately reporting confidence intervals on its own predictions. Taken together with the creation of highly contextualised models of a person's thoughts and behaviours, it seems very likely that people will readily experience an AI clinician as genuinely caring and understanding. Availed with sufficient data, a future AI will build a deep enough model of a person's responses that its understanding of them surpasses that of their psychiatrist.

We emphasise that the succession of human psychiatrists by AI agents is not without risk. Unsupervised machine learning could lead to insidious exacerbations of existing biases.[15] Value misalignment - the degree to which the goals of AI systems fail to overlap with our own - is a major potential hazard.[16] In psychiatry this problem is prefigured by existing disagreements as to what the objective of mental health care is [17,18,19], in other words *how people ought to live*. These are political questions and will remain a matter for human discourse. As such, we must qualify our argument with the suggestion that the process of *psyche-iatros* (mind-healing) in a broad sense will retain a human component, even in a post-human-doctor world.

Perhaps human patients will want human doctors – with all their quirks and comparative ineptitude – simply because they're human. Cynically, if silicon-shrinks are cheaper and measurably as effective as human psychiatrists, they may find mass employment by economic demands alone. More positively, AI promises considerable advantages for patients. Rather than straitjacketing complex individuals into diagnostic categories, and doling out treatments based on generic guidelines, AI offers truly individualised care. Furthermore unlike the hugely fragmented pathways of care patients are currently subjected to, a personal AI clinician would be available more or less anywhere (from primary care to the inpatient ward) and at any time (cradle-to-grave, night or day). Over time then, patients' trust could be earned and maintained, rather than being repeatedly fractured with every handover. Thus people may well find AI-led care to be paradoxically more humane than the *status quo* of psychiatry today; in their desire to be understood, and treated to the best of scientific understanding, they will willingly choose AI over its flesh-and-blood-clinician alternative.

Against

Baker JT [1,2], Mourão -Miranda J [3,4]

1. Institute for Technology in Psychiatry, McLean Hospital, 115 Mill St. Belmont, MA 02478

2. Department of Psychiatry, Harvard Medical School, Boston, MA 02115

3. Centre for Medical Image Computing, Department of Computer Science, University College London, UK

4. Max Planck University College London Centre for Computational Psychiatry and Ageing Research, University College London, UK

Will the art of mind healing eventually ditch “the human middleman” and come to rely exclusively on data-driven AIs that patients will pour their hearts out to and then optimize their mental health? Don’t count on this rosy tech-driven mental health future coming to pass anytime soon.

While we agree that AI has the potential to significantly transform the *technical* aspects of psychiatry by discovering hidden patterns in rich, longitudinal, multi-modal data which can be predictive of individual risks and outcomes,[20] we’re less convinced that AI-based tools will ever be able to serve as the treatment of choice for the vast majority of humans in distress, at least in a way that fulfils the purpose of the psychiatrist. Of course, many humans find it easier to be more honest with computers about certain things; however, this hardly indicates that the encounter is serving the best interests of the patient, except in a strictly information gathering context. However this is a tiny fraction of a psychiatrist’s role and one often served by less trained individuals or by intake forms. Once computers start responding more like a human concerned for that individual’s long-term well-being, humans may well become more circumspect about what they reveal to computers.

Starting from a technical point of view there are still many challenges for AI to overcome. Currently, the quality (and quantity) of data available to train AI models in psychiatry is very limited, partly driven by untested or even invalid assumptions about the structure of mental illness. As an example, the performance of neuroimaging-based diagnostic AI models tends to decrease as sample sizes increase (e.g. [21,22]), indicating that the models do not perform well for highly heterogeneous large samples. One important reason for this lack of clarity in such studies focused on using AI for diagnostics is the extent of comorbidity in psychiatry, and the limitations of current categorical classifications.[23] Therefore, it is very unlikely that AI systems will be successful on solving any diagnosis task, since psychiatrists often disagree and therefore there is no ground truth to measure model performance against.

Furthermore, artificial intelligence is very different from human intelligence. AI systems often perform badly when presented with data that is very different from the training data or with a new task (different from the one it has been trained to solve). While AI systems can be trained to detect novelty (‘anomaly or outlier detection’ [24]) and transfer knowledge to solve related problems (‘transfer learning’ [25]), making accurate predictions when confronted with uncommon patterns or new tasks is incredibly difficult for an AI system. For example, an “AI psychiatrist” wouldn’t know what to do if a patient were to present with a behaviour completely different from every behaviour it has "seen" before. In such situations, a human psychiatrist is able to take in a far broader and real-world relevant set of observations, and then draw on their own idiosyncratic knowledge base to contextualize those observations, applying a moment-by-moment solution that mutually optimizes for more than just the treatment outcome (e.g., aspects of that individual's current life situation that the AI never sees and thus never models). This ability of human intelligence to draw on "common sense" when needed, and uncommon sense when especially needed (i.e., that of specialists with their unique training data), means that humans will almost certainly remain critical for managing and interacting with even relatively simple psychiatric cases for the foreseeable future.

Looking ahead, we acknowledge that AI systems may eventually learn to incorporate many of these contextual variables and become more efficient in transfer learning across different tasks. And yet, AI systems will require more, better data than currently exist to solve any real-world clinical task faced by psychiatrists. For most of these points in the modern psychiatry workflow (e.g., initial evaluation, establishing a trusting working relationship, differential diagnosis, treatment selection, side-effect and efficacy monitoring) very few if any studies have attempted to measure and compare clinical behaviours that distinguish adequate from inadequate clinical performance. While several large-scale research efforts – such as the UK Biobank – are collecting detailed information from thousands of individuals – including many demographic, lifestyle, behavioural, and imaging-related factors – these data barely scratch the surface of what a psychiatrist would need to perform any of the above tasks competently, let alone provide comprehensive care of a single patient.

Some have even suggested that the lack of group-to-individual generalizability – which is not unique to psychiatry – threatens the entire enterprise of human subjects' research,[26] since most human-derived measures fail to account for important within-person variance that are critical to their interpretation. Therefore, to truly succeed at contextualized AI – able to provide nuanced care to complex, evolving people embedded in complex, evolving environments, – these systems would require data from hundreds or potentially thousands of patients and psychiatrists, gathered across dozens of repeated encounters, and combined with standardized ways to assess observed outcomes in order to validate any autonomously discovered relationships to establish trust in the algorithm. These data sets are possible, and indeed are starting to be collected, but they too require tremendous human capital to collect and understand.

And no matter how sophisticated such systems become, psychiatry will always be about connecting with another human to help that individual integrate all the conflicting signals they are receiving and make the best choices for their life situation. So much of addressing mental health is dealing with challenging, embarrassing issues that people often do not admit even to themselves, and only begin to comprehend as layer upon layer of meaning and data are exposed by skilled professionals. Human psychiatrists (and psychologists) are the ones in the loop who that individual will not be able to fool, and will know how to respond when things go sideways, which they often do. In these cases, how many would feel safe with an autonomous agent helping them and their family make the best use of available resources and data during a complex mental health challenge, and not inadvertently make things worse?

Clinicians nowadays rarely rely on strict categorical diagnostic systems, because they already know to treat constellations of symptoms and avoid side-effect profiles that will intersect with real-life challenges like cost and access to follow-up care. It would seem that entrusting an AI with all this complexity and uncertainty has no less potential for abuse and eventual backlash than any of psychiatry's past treatment approaches (e.g., straightjackets) and diagnostic schemes that eventually were discarded, and even reviled, precisely because they had failed to capture relevant individual nuance and thereby eroded patient trust. This meta-psychiatry problem is a social one and probably one AI systems cannot and should not solve. Human psychiatrists, along with the people they treat, clearly will first need to agree on which problems are even worth addressing with AI in order to move forward in a way that is both pragmatic, ethical, stakeholder-driven, and *iterative* by design. The power of iteration is one place humans could stand to learn a great deal from artificial systems; but with regard to learning what it means to be "*humane*", we humans will be the ones doing the teaching for a good long while.

For: rebuttal

Baker and Mourão-Miranda argue that psychiatry presents significant technical challenges to the AI practitioner of today, and we don't disagree. However, we take the view that there is nothing so mysterious about human behaviour and social interaction that it will never be possible to simulate these by artificial means. As such, the question of technical capability is, for many, not one of *if*, but *when*. We turn therefore to our opponents' non-technical claims made against AI psychiatry in general.

They assert that 'psychiatry will *always* be about connecting with another human'. If AI has *too little* understanding of humans, they point out, there is the potential for people to 'be able to fool' the artificial clinician, or for the silicon shrink to 'inadvertently make things worse'. With *too much* understanding however AI could arouse suspicion. Of course, similar concerns arise with human clinicians. Indeed, this re-telling of Goldilocks, in which people want their AI-psychiatrist to have just the right level of intelligence, reveals a human tendency to prefer their own kind *a priori*, notwithstanding how sophisticated the alternative might be. Among other factors, an AI psychiatrist's authenticity is limited by its lack of a human body, and who knows whether we will ever fully confide in those who are not made of flesh-and-blood like ourselves? If we can't, and continue to nurse our anthropocentric distrust, we stand to miss out on great therapeutic potential.

Our colleagues highlight that AI psychiatry has the potential for abuse and backlash (*à la* the straitjacket). We would go further and acknowledge the immensely serious concerns of many within the world of AI ethics, that these new technologies could give rise to devastating, perhaps *existential* consequences. If human intelligence is what afforded us dominion over nature, superintelligence should rightly be regarded with caution. Baker and Mourão-Miranda highlight in particular a social and 'meta-psychiatric' problem, which they assert that AI *should* not solve, with humans remaining at the helm. We don't see the problem in such binary terms.

There is no clear point at which prevention of mental illness departs from improving the wellbeing of the population. Similarly, the practice of psychiatry on an individual level affects the culture of psychiatric practice, which in turn exists in equilibrium with wider societal values. We acknowledge therefore that AI practitioners will have effects far beyond the clinic room, becoming active participants in our ever-shifting ethical dialogue. Neither humans nor computers ought to entirely dictate this dialogue – rather, various agents will contribute to varying degrees, in an interplay where humans remain able to exert their diverse priorities and values. We do not anticipate a future of ethical homogeneity, artificial or human.

In conclusion, we emphasise that the advance of AI psychiatry is inexorable, and for its advantages (therapeutic and economic) people will readily choose to use it. As such, we should stop debating *whether* it will become sufficiently sophisticated to replace psychiatrists, and instead turn our focus on how best shape this future, and what kind of ethical and regulatory systems are needed to prevent disaster.

Agains: rebuttal

Brown and Story emphasize the inexorable nature of AI eventually replacing psychiatrists, and again we don't disagree that some forms of AI-driven mental health solutions will be integral to the future of mental health care. However, we are not aware of any evidence to suggest the true effectiveness of human- vs. computer-mediated psychiatric disease management is likely, in the end, to come out in favour of the AI agents outperforming actual humans in the complex task of managing humans, or interacting with them at the point of encounter to achieve optimal behavioural health outcomes. It seems to us far more likely, when all is said and done – and the right data are collected and analysed to properly assess the value of both the in-person and remote encounters – that the data will show that computers alone (and even computer-assisted virtual encounters) are not nearly as effective at getting the job done as well-trained humans – at least not today's computers, with their limited interfaces and access to human-relevant metadata. But we agree much of this will change and is changing rapidly.

After all, many individuals *do* seek out indirect human interaction to alleviate their suffering, including online text coaches and therapists (Crisis Text Line, 7cups, WoeBot, etc.), and there is clearly a market for such platforms for many individuals who cannot or prefer not to seek care through conventional means. But again, we strongly suspect – based on the underlying phenomenology of depression and other severe psychiatric conditions – that in-person contact will continue to represent best clinical practice for the vast majority of psychiatric interactions, and that despite the many exciting alternative forms of care delivery designed to improve access and standardize quality, humans will continue to be favored over less direct, less humanizing alternatives. Indeed, we are especially concerned with unintended consequences of expanding access to AI-based platforms that are relatively untested, as they could exacerbate existing disparities by providing those with fewer means access to only lower quality treatment options. We think systems can be designed to avoid this scenario; however, for many individuals who might have been helped by a human with even modest training, the prospect of inexpensive AI-based systems that become the only available option for those in need raises significant ethical concerns that must be addressed.

Given our contention that humans in distress will always be best served by meaningful interactions with other humans, we cannot ignore the potential for immersive technologies like virtual reality (VR) and augmented reality (AR) to help facilitate human interactions at a distance. With VR and AR capabilities built into modern communication devices, immersive technologies are increasingly culturally resonant, such that over the 3-5 year horizon (and certainly within the next decade), realistic in-home agents that replace 2D video conferencing with virtual 3D VR/AR interactions with distant individuals. Here, we are not talking about interacting with virtual avatars [27], but rather connecting with people at distant locations in ways that *feel* real, allowing people to experience deep personal connections with others without technology getting in the way. Just as the first people to experience a musical recording or a film in their own home in ways that felt "just like the real thing," we suspect humans will one day be able to experience one another's presence at a distance in ways that *no longer feel degraded*. While we are not there yet, the technical advances needed to realize this are knowns rather than unknowns, and many are well in process, the remaining engineering and cultural obstacles to achieve embodied interactions with distant humans in realistic virtual or augmented environments is so close at this point as to be both exciting and terrifying, given how it could impact humanity.

In conclusion, while in a narrow sense we stand by our assertion that human psychiatrists will never be replaced by AIs, the technical advances provided by AI will inevitably transform many aspects of psychiatry, particularly in the identification of new biomarkers and approaches to patient stratification, which will likely lead to development of new and more effective therapies. And yet in the end, until technology can enable new ways of connecting with another human being that don't seem degraded by the human-computer interface, psychiatrists will probably remain most effective at the human-*human* interface.

References

[1] Grace K, Salvatier J, Dafoe A, Zhang B, Evans O. When will AI exceed human performance? Evidence from AI experts. *arXiv* [Internet]. 2017 [cited 2019 July 9]; 1705.08807v2 [cs.AI].

[2] Onnela J-P, Rauch SL. Harnessing Smartphone-Based Digital Phenotyping to Enhance Behavioral and Mental Health. *Neuropsychopharmacol* 2016;**41**:1691-1696.

[3] Vanello N, Guidi A, Gentili C, Werner S, Bertschy G, Valenza G, et al. Speech analysis for mood state characterization in bipolar patients. *Conf Proc IEEE Eng Med Biol Soc.* 2012; 2104-7.

[4] Valenza G, Nardelli M, Lanata A, Gentili C, Bertschy G, Paradiso R, et al. Wearable monitoring for mood recognition in bipolar disorder based on history-dependent long-term heart rate variability analysis. *IEEE J Biomed Health Inform* 2014; **18(5)**: 1625-1635.

[5] Gruenerbl A, Osmani V, Bahle G, Carrasco JC, Oehler S, Mayora O, et al. Using smart phone mobility traces for the diagnosis of depressive and manic episodes in bipolar patients. *Proceedings of the 5th Augmented Human International Conference* 2014; 38.

[6] Huys QJM, Maia TV, Frank MJ. Computational psychiatry as a bridge from neuroscience to clinical applications. *Nat Neurosci* 2016; **19**: 404-413.

[7] Grisanzio KA, Goldstein-Piekarski AN, Wang MY, Rashed Ahmed AP, Samara Z, Williams LM. Transdiagnostic Symptom Clusters and Associations With Brain, Behavior, and Daily Function in Mood, Anxiety, and Trauma Disorders. *JAMA Psychiat* 2018; **75**: 201–209.

[8] Krupnick JL, Sotsky SM, Simmens S, Moyer J, Elkin I, Watkins J, et al. The role of the therapeutic alliance in psychotherapy and pharmacotherapy outcome: findings in the National Institute of Mental Health Treatment of Depression Collaborative Research Program. *J Consult Clin Psych* 1996; **64(3)**: 532.

[9] Elmasri D, Maeder A. A Conversational Agent for an Online Mental Health Intervention. In: Ascoli G, Hawrylycz M, Ali H, Khazanchi D, Shi Y, editors. Brain Informatics and Health. BIH 2016. Lecture Notes in Computer Science, vol 9919. Springer, Cham

[10] Heider F, Simmel M. An experimental study of apparent behaviour. *Am J Psychol* 1944; **57**: 243-259.

[11] Scholla BJ, Tremouletb PD. Perceptual causality and animacy. *Trends Cogn Sci* 2000; **4**: 299-309.

[12] Bemelmans R, Gelderblom GJ, Jonker P, de Witte L. Socially Assistive Robots in Elderly Care: A Systematic Review into Effects and Effectiveness. *JAMDA* 2012; **13**: 114-120.

[13] Suzuki Y, Galli L, Ikeda A, Itakura S, Kitazaki M. Measuring empathy for human and robot hand pain using electroencephalography. *Sci Rep UK* 2015; **5**: 15924.

[14] Lucas GM, Gratch J, King A, Morency L-P. It's only a computer: Virtual humans increase willingness to disclose. *Comput Hum Behav* 2014; **37**: 94-100.

[15] Bolukbasi T, Chang K-W, Zou J, SaligramaV, Kalai A. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. *arXiv*. 2016; 1607.06520 [cs.CL].

[16] Bostrom N. Superintelligence: Paths, Dangers, Strategies, Ch. 8. Oxford University Press, 2014.

[17] Szasz, TS. The myth of mental illness. *Am Psychol* 1960; **15**, 113-118.

[18] Kendell RE. The myth of mental illness. In Schaler JA, editor. Szasz Under Fire: The Psychiatric Abolitionist Faces his Critics. Open Court, 2004.

[19] Moutoussis M, Story GW, Dolan RJ. The computational psychiatry of reward: broken brains or misguided minds? F*ront Psychol* 2015; **6**,1445.

[20] Bzdok D, Meyer-Lindenberg A. Machine Learning for Precision Psychiatry: Opportunities and Challenges, Biological Psychiatry: Cognitive Neuroscience and Neuroimaging 2018 Mar; **3**(3):223-230

[21] Varoquaux G. Cross-validation failure: Small sample sizes lead to large error bars. Neuroimage. 2018 Oct 15; **180(Pt A)**:68-77

[22] Janssen RJ, Mourão-Miranda J, Schnack HG. Making Individual Prognoses in Psychiatry Using Neuroimaging and Machine Learning. Biol Psychiatry Cogn Neurosci Neuroimaging. 2018 Sep; **3**(9):798-808

[23] Insel T, Cuthbert B, Garvey M, Heinssen R, Pine DS, Quinn K, et al. Research domain criteria (RDoC): toward a new classification framework for research on mental disorders. Am J Psychiatry. 2010 Jul; **167**(7):748-51

[24] Hodge VJ, Austin J. A Survey of Outlier Detection Methodologies. Artif Intell Rev 2004 Oct; 22:85-126

[25] Pratt LY, Thrun S. Guest Editors' Introduction. Machine Learning. 1997 Jul; **28**(1):5

[26] Fisher AJ, Medaglia JD, Jeronimus BF. Lack of group-to-individual generalizability is a threat to human subjects research. *Proceedings of the National Academy of Sciences*, 2018 Jun; 201711978; DOI: 10.1073/pnas.1711978115

[27] Morency L-P., Rizzo A. SimSensei & MultiSense: Virtual Human and Multimodal Perception for Healthcare Support [web streaming video]. YouTube, 2013 [cited 2019 July 9]. Available from: https://youtu.be/ejczMs6b1Q4