CHAPTER

18

Ethical issues in the application of machine learning to brain disorders

Philipp Kellmeyer[1, 2, 3]

[1] Neuromedical AI Lab, Department of Neurosurgery, Freiburg, Germany; [2] Institute for Biomedical Ethics and History of Medicine, University of Zurich, Zurich, Switzerland; [3] Freiburg Institute for Advanced Studies (FRIAS), Research Focus: Responsible Artificial Intelligence, University of Freiburg, Freiburg, Germany

18.1 Introduction

With the increasing digitization of health-related data, machine learning is one of the fastest growing areas in biomedical research. This trend is taking place in the context of increasing amounts of potentially useful data becoming available to clinicians and researchers. Not only clinical neurology and psychiatry but also basic clinical research are now awash in biomedical data collected with various methods: (1) neuroimaging, including computer-aided tomography, structural and functional magnetic resonance imaging (MRI), MRI-based diffusion tensor imaging, positron emission tomography, and magnetic resonance spectroscopy; (2) electrophysiological methods including extracranial and intracranial electroencephalography (EEG) and, to a lesser degree, magnetoencephalography; (3) genetic and epigenetic information or, more generally, -omics mapping approaches (from genome and proteome to interactome); and (4) biomarkers from blood, cerebrospinal fluid, and tissue samples. This profound transformation offers great opportunity for advancing our understanding of brain disorders in neurology and psychiatry, for example, by uncovering hitherto unknown neural signatures that could be used as biomarkers to inform diagnostic and prognostic assessment; identifying subtypes or variants of specific disorders,

Machine Learning
https://doi.org/10.1016/B978-0-12-815739-8.00018-3

e.g., depression or dementia; or discovering new nosological entities that were not recognized by existing methods of data analysis and clinical classification systems based on phenomenological presentation. Machine learning methods do, however, pose substantial ethical and legal challenges and tensions in various domains of research and clinical application. Before discussing these ethical tensions, some current and emerging applications to specific brain disorders shall be highlighted to contextualize the main ethical issues.

18.2 Applications of machine learning to brain disorders

Initial attempts to use machine learning methods in brain disorders research have been based on neuroimaging data; these attempts involved, for example, differentiating between people with and without a diagnosis of depression (Patel, Khalaf, & Aizenstein, 2016) or predicting Alzheimer's disease (AD) from MRI images (Moradi, Pepe, Gaser, Huttunen, & Tohka, 2015) or MR spectroscopic images (Su et al., 2016). In light of the increasing flexibility of the latest methods, including deep neural networks (see Chapter 9–11), however, data that contain information in other domains, for example, time and frequency in EEG data (Schirrmeister et al., 2017), or semantic content and syntactic features in text, have been shown to be valuable for machine learning. For example, social media entries could be used to detect and monitor depression (Mohr, Zhang, & Schueller, 2017). This development is leading to an exponential increase in the application of machine learning in brain disorders. For example, aging populations and age-related neurodegenerative diseases pose a global public health challenge (Shah et al., 2016). The failure of existing pharmacological and psychosocial interventions could, in part, be due to the fact that patients receive treatment too late in the disease process. Therefore, using machine learning to detect early stages of AD, the most common of the neurodegenerative diseases, is a promising approach. A number of research groups are actively investigating the feasibility and usefulness of machine learning for identifying early stages of AD, either based on one particular type of data (such as MRI images) or the combination of multiple types of data (Amoroso et al., 2018; Janghel, 2018; Lee et al., 2018). In epilepsy, the brain signal of interest is electrophysiological brain activity measured with extracranial or intracranial EEG. These data pose a substantially more difficult machine learning pattern recognition and feature selection problem than, for example, static 2D images used for the detection of skin cancer or diabetic retinopathy, due to the higher dimensionality (including the time and frequency domain) of the data. Nevertheless, machine learning methods,

particularly deep learning approaches including convolutional neural networks, have now been shown to be well suited for analyzing even raw EEG data in the time and frequency domain to identify pathological features of epilepsy (Acharya, Oh, Hagiwara, Tan, & Adeli, 2018; Schirrmeister et al., 2017; Schirrmeister, Gemein, Eggensperger, Hutter, & Ball, 2017). Eventually, these methods could be used to improve early detection of epileptic seizures; this could be achieved, for example, by building real-time closed-loop systems for seizure control based on brain state−dependent electrical stimulation.

In psychiatry, the application of machine learning is also gaining enormous momentum. Here, the heuristic challenges may perhaps be even greater than in neurological disorders (e.g., stroke or epilepsy), due to more limited nosological understanding. Let us consider, for example, a situation when a researcher or clinician would like to use machine learning methods to detect and/or predict depression at the level of the individual. Here, the choice of the most suitable data is complicated by (1) our current lack of a unified pathophysiological understanding of the illness; (2) our current lack of validated biomarkers; and (3) the heterogeneity of classification systems. Therefore, the current trend in the application of machine learning to depression, as well as other psychiatric disorders, involves using a multitude of data including sensors (e.g., movement patterns from geolocation and gyroscopic data from mobile phones), behavioral data (e.g., semantic content analysis of social media entries), clinical data (e.g., depression rating scales and questionnaires), and neuroimaging data (Barak-Corren et al., 2016; Burns et al., 2011; Dipnall et al., 2016; Kessler et al., 2016; Wager & Woo, 2017). While such a broad and indiscriminate (with respect to the data ontology) approach might be unsatisfactory from the perspective of developing a more nuanced theoretical understanding of the disease, the use of a plethora of data may nevertheless improve accuracy of detection and/or prediction at the individual level.

18.3 Ethical tensions from using machine learning in brain disorders

18.3.1 Overview

Different applications of machine learning to brain disorders raise different ethical tensions, depending on their ultimate purpose and the nature of the disorder in question. Here, therefore, the aim is to use some examples to illustrate the *kind* of ethical tensions that may arise from the increasing popularity of machine learning in neurology and psychiatry.

In light of the increasing versatility of machine learning architectures and their capacity for real-time data analysis, it now seems attainable to build complex decision-support system (DSS) for medical professionals. Let us consider the clinical example of the acute treatment of ischemic stroke. Here, successful treatment of artery-blocking clots with thrombolytic medication is not only time-sensitive—the earlier the treatment is administered, the better the functional outcome of the patient—but also carries substantial risks (e.g., cerebral bleeding). This scenario, therefore, represents a good example of complex decision-making, often under substantial uncertainty. For such a high-stake clinical decision-making problem, we might imagine the following kind of DSS to support neurologists who are treating patients with acute stroke. First, a machine learning model could be trained to recognize an acute cerebral ischemia from MRI images of patients that show acute symptoms of stroke; second, once the presence of acute cerebral ischemia has been established, the machine learning model could be supplemented with auxiliary information (e.g., demographic data, medical history, etc.) to estimate how likely it is that the patient will benefit from thrombolytic treatment and suffer from serious complications. In the following, let us consider some ethical tensions that could arise from the interactions between a human (in this case, a neurologist having to make diagnostic and treatment decisions) and an "intelligent" system (in this case, a DSS).

18.3.2 Challenges in the interaction between humans and intelligent systems

Among the many ethical tensions that can emerge from the close interaction of humans and intelligent systems, we will focus here on the issues of decision-making capacity, agency, and accountability, as these are particularly pertinent to the ethics of medical decision-making. In the case of the machine learning–based DSS for stroke detection and treatment sketched above, as in most other types of such systems, it is critical to consider the precise configuration of the human–machine interaction. As decision-making capacity or agency are transferred from the human to the intelligent system, a situation of "shared agency" may arise where the moral and legal accountability of the human diminishes accordingly (Goering, Klein, Dougherty, & Widge, 2017; Kellmeyer et al., 2016). In this case, however, we do not have philosophical foundations or legal instruments to interpret and adjudicate accountability in cases of severe system failures (Fig. 18.1). While medical insurance may allow for financial compensation, it may not be sufficient to satisfy the human need for ascribing responsibility to discernible human subjects; yet we know

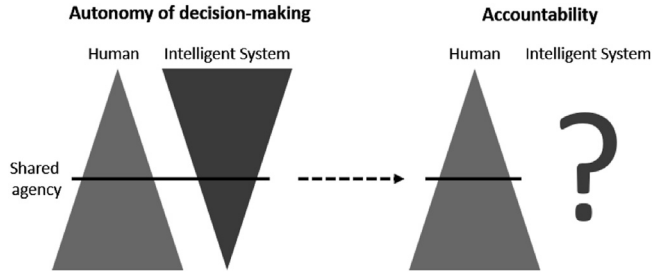

FIGURE 18.1 Illustration of the "accountability gap" that may arise in instances in which a human and an intelligent system work in concert or close interaction. As human autonomy of decision-making and accountability diminishes, who or what is accountable in cases of grave system failures?

that, in many instances, the identification of responsible individuals is required to achieve psychological relief and, ultimately, closure.

In this configuration of human–machine interaction, the accountability gap will be particularly large if the human (in this case the neurologist) is kept *out of the loop*; i.e., the neurologist follows the recommendation of the DSS and their role was relegated to communicating this recommendation to the patient. In contrast, the accountability gap will be small or even absent if the human is kept *in the loop*; i.e., the neurologist keeps autonomy of decision-making and agency and, ultimately, maintains accountability in cases of grave system errors. The patient is of course an important part of the *overall* decision process and, critically, their contribution to this decision might be influenced by whether the recommendation (to treat or not) is perceived as coming from the doctor, the DSS, or a hybrid of both. In other words, the relational aspects of the interaction between patients and health care professionals are likely to be affected by the inclusion of an intelligent DSS in the clinical decision-making. Therefore, the empirical study of human–computer interaction, including aspects such as patients' attitudes and trust toward technology ("technology acceptance") (Kellmeyer, Mueller, Feingold-Polak, & Levy-Tzedek, 2018), should be included in the overall assessment of an intelligent DSS. These observations raise the question: how much decision-making autonomy should we grant to an intelligent DSS? Critically, different applications of machine learning to brain disorders have different benefits and risks (and benefit/risk proportions) and may operate under very different constraints (e.g., time-sensitive or not). Therefore, careful and evidence-based risk assessment is an important first step when deciding how much decision-making autonomy should be granted to an intelligent system—not just in the clinic, but in all contexts in which humans and intelligent DSS work in close interaction.

18.3.3 Data security, patient privacy, the ontology of biomedical data, and mental privacy

18.3.3.1 Data security and patient privacy

With the current trend toward the digitalization of biomedical data, leading to the proliferations of electronic health records, society faces an enormous challenge to protect these data from unwarranted access and illegitimate use (data security) and to preserve the privacy of individual patients (patient privacy) (Amunts, 2018; Kellmeyer, 2018; Yuste et al., 2017). In addition to digitalized biomedical data that are generated in the more or less well-regulated (depending on the sociopolitical system) health-care system, an expanding range of personal data is also generated outside this system. This includes, for example, the digital information that is created by (1) personal smartphones as people go about their daily life; such information includes measures of *mobility* such as time spent at home, distance walked, distance cycled, distance traveled, significant location entropy, etc., and measures of *sociability* such as number of texts, text length, texting reciprocity, texting responsiveness, number of calls, call duration, number of social apps used, social app use duration, etc., fitness trackers, geolocation data from mobile phones, etc., or (2) individuals themselves by using "social" media and other digital "services." This development has spawned much interest in the industry and in academia in harvesting these data for creating countless "mobile health technology" (mHealth) or "electronic health technology" (eHealth) applications, specifically in the area of mental health and well-being (Kreitmair, Cho, & Magnus, 2017; Marzano et al., 2015). Increasingly, the traditional distinction between biomedical data generated and stored within the traditional health-care system and data created in the social realm via everyday use of digital technologies is becoming porous. It is not difficult to envisage a situation where biomedical data, such as blood biomarkers or MRI images, are combined with other types of data, for example, an individual's entries in their social media feed, to infer diagnosis and predict clinical outcome within a multimodal machine learning framework. Such enormous stockpiles of personal data of various origins, however, are vulnerable with respect to data security and patient privacy. Many technological solutions are available (or in development) to protect these aspects, such as the use of distributed and decentralized machine learning methods (e.g., federated learning) in which data are analyzed locally and then the resulting features are aggregated collectively (Winograd-Cort, Haeberlen, Roth, & Pierce, 2017). Less discussed, however, is the issue of group privacy, i.e., the capacity of machine learning methods to identify distinctive (and often discriminating) features of a particular group—e.g., gender, ethnicity, or other identifiers—in biomedical data. Here, the potential danger is that private health-care

providers or insurance companies could use these methods to discriminate against particular groups in society; this poses a significant political challenge in terms of preventing such misuse of machine learning technologies.

18.3.3.2 On the ontology of biomedical data

Biomedical data tend to have clearly defined legal status and governance within a given health-care system. This raises the ontological question of what types of data actually *are* biomedical data and which are not. Be it explicitly or implicitly, many working definitions of "biomedical data" focus on the origin of the data—biomedical data are measurements or other empirical data of human anatomy, physiology, or behavior—rather than its use for medical purposes. This traditional definition of biomedical data would not normally include data generated via personal smartphones. From the perspective of machine learning, the distinction between biomedical and nonbiomedical data is not particularly meaningful, as the two types of data are treated equally within a multimodal machine learning framework. From the perspectives of data security and patient privacy, however, this distinction becomes important because of their different legal status and governance. Let us consider a situation where personal entries on social media are being aggregated on cloud-based computing servers and analysed with machine learning methods for creating targeted advertising, consistent with most end user license agreements; in this case the data would be regarded as personal rather than biomedical data. Let us now consider a situation where the same data are being used internally by the company for large-scale studies in social science or psychology (Wilson, Gosling, & Graham, 2012) or, increasingly, for investigating medical problems such as predicting depressive episode or spotting suicidality. In these cases, the ontological status of the data would need to be reconsidered. This fluidity and elasticity of the definition of biomedical data means that, in several circumstances, the legal status and governance of a dataset can be uncertain. It is important to initiate a broad conversation on these issues that includes multistakeholder perspectives.

18.3.3.3 Mental privacy and the question of "neurorights"

With the increasing abilities of machine learning methods to extract informative features from large amounts of neuroimaging data, another prominent issue in neuroethics and related fields is the question of mental privacy, i.e., the privacy of first-person subjective experience and thoughts. While existing methods are far from being able to decode specific thoughts or "inner speech" from extracranial or intracranial neuroimaging data, it might nevertheless become possible to discern more general states of mind, such as levels of wakefulness, and identify

neural patterns associated with specific cognitive operations, e.g., motor imagery (Aflalo et al., 2015; Derix, Iljina, Schulze-Bonhage, Aertsen, & Ball, 2012; Völker, Schirrmeister, Fiederer, Burgard, & Ball, 2018). For example, let us consider a closed-loop neural implant (such as a micro-electrode grid for recording electric brain activity) that can be used to predict impending epileptic seizures in patients with epilepsy (Kohler et al., 2017). Now, let us imagine that the real-time application of machine learning to the data collected via this neural implant allows detection of whether a patient experiences pain or not. How to deal with a situation where the doctor asks the patient "Are you in pain?" and the patient replies "Yes!" but the output of the machine learning algorithm indicates that the patient does not experience pain? Previously, the privacy of first-person subjective experience and thoughts protected the patient from having their experiences involuntarily shared with another person. Now, neuroimaging techniques threaten to undermine the notion of mental privacy, providing access to a range of measures that could be of great interest to the judicial system (e.g., as evidence in courtrooms) (Kellmeyer, 2017). This, in turn, raises questions on whether existing normative concepts, civil liberties, and legal instruments suffice to justify and govern the use of machine learning technologies for this purpose, or whether we need additional "neurorights," for example, a basic human right for mental privacy (Ienca & Andorno, 2017).

18.3.4 Transparency, interpretability, and biases of machine learning

18.3.4.1 Transparency of machine learning

An often voiced critique of machine learning methods, particularly the most advanced approaches such as deep learning, emphasizes the black box aspects of self-learning algorithms. In other words, owing to the complexity of deep learning approaches, i.e., the high number of intricate transformations to the input data, the learning process lacks transparency and it becomes challenging to extract meaningful information from these models, such as which features are providing the greatest contribution to classification (see Chapter 9) (Castelvecchi, 2016). In many scenarios, this inherent black box aspect of some machine learning architectures might not cause a particularly salient ethical problem. Let us imagine, for example, a deep learning model for real-time analysis of EEG data aimed at understanding and improving experience in people who are playing a computer game: in this case, failure of the algorithm to achieve this aim would not have catastrophic effects and the black box nature of the program would most likely not be considered to pose a particular ethical problem. If, on the other hand, the program was used for monitoring performance in pilots or for controlling an electric wheelchair by a

paralyzed person, we would recognize the necessity to be able to understand the operations of the algorithm in cases of catastrophic failures. Therefore, the answer to the question of how transparent a medical device, software, medical robot, or DSS that uses machine learning ought to be depends on the context including the proportion between risks and benefits (Yuste et al., 2017).

18.3.4.2 Interpretability of machine learning and the problem of bias

A further potential ethical issue with machine learning methods relates to the interpretability of the distinctive features and patterns that are extracted from the data. The question of interpretability of the machine learning framework's output—as well as the danger of perpetuating existing biases—is closely connected to the type of learning scenario that is used to train the network. Important concepts in this context are supervised and unsupervised learning (see Chapter 1). In supervised learning, an algorithm trains with labeled data (e.g., patients vs. controls). Following learning, the algorithm is applied to new observations to predict whether these belong to one class or another. Unsupervised learning, in contrast, does not use labels. Here, the algorithm receives no input on the categories within a dataset but explores the data and finds recurring patterns by itself, for example, by finding clusters of "similar" brain images. Now, in an unsupervised learning scenario, there is the possibility that a particular algorithm might find patterns, classes, or distinctive features in a dataset that cannot be readily interpreted by a human observer or necessarily judged against established gold standards or "ground truth" (which might not be known in the particular machine learning challenge at hand). For example, raw imaging measurements, whether from a computed tomography or MRI scan, can include up to 65,000 shades of gray (Kimpe & Tuytschaever, 2007). Commercial medical displays, however, usually only distinguish 256 shades of gray (in specialized displays up to 1000 shades), whereas the human visual processing system can only distinguish around 30 shades of gray, depending on the lightning conditions (Kreit et al., 2013). This means that human expert raters, such as neuroradiologists attempting to identify brain abnormalities, might find very different patterns and clusters of abnormalities in the data than an unsupervised machine learning algorithm. Yet, how do we rationally evaluate the performance of the human raters against the machine learning rating if both operate on very different heuristics and capabilities? In a supervised learning scenario, a machine learning algorithm could be trained with labeled examples of normal versus pathological MRI images based on an assessment made by neuroradiologists; here the main problem would not be a mismatch of the heuristics and the interpretability of the machine learning diagnostics, but

rather the problem that inherent biases of the human raters might skew the algorithm to reproduce these biases in its own performance. This problem of bias is ubiquitous and pervasive, as all data that are collected, labeled, and processed based on human capacities and existing ontologies will invariably replicate the inherent cognitive biases in these capacities and ontologies (in terms of defining disease categories, classes, and so on) (Ioannidis, 2011; Kahneman & Tversky, 1977; Palminteri, Lefebvre, Kilford, & Blakemore, 2017). Unfortunately, research in cognitive psychology and medical decision-making has demonstrated that it is nearly impossible to effectively (and lastingly) debias humans (Smith & Slack, 2015; Croskerry, Singhal, & Mamede, 2013a; 2013b; Ehrlinger, Gilovich, & Ross, 2005). The problem of how to effectively debias intelligent systems—apart from a current lack of recognition of the depth and pervasiveness of the problem in human–computer interaction—will require coordinated efforts by computer scientists, cognitive scientists, and social scientists and may even be impossible to "solve" in the sense in which other problems in science are solvable (Courtland, 2018; Devlin, 2016; Knight, 2017).

18.4 Conclusion

Machine learning offers exciting and promising opportunities for increasing our understanding of brain disorders and using this information to improve diagnosis and treatment of neurological and psychiatric diseases. With medicine in general and clinical neuroscience in particular becoming an ever more data-rich environment, however, there are significant ethical and legal challenges that require careful consideration. In scenarios that involve close interaction between a human and an intelligent system (i.e., a system that employs machine learning), such as our example of a neurologist who is aided by a DSS for treating acute stroke, we might face particular ethical tensions. In cases in which the autonomy of decision-making is transferred from the human to the intelligent DSS, for example, we might encounter an accountability gap, both morally and legally, when a grave error occurs and individual agency or causation cannot be established with sufficient certainty. The datafication and digitalization of medicine, also taking place in neurology and psychiatry, requires the storage and processing of large amounts of data, which in turn can be subject to inherent vulnerabilities with respect to data security and the privacy of individuals and particular social groups. Furthermore, the increasing ability of advanced and adaptive machine learning methods to infer mental states from neuroimaging data may create ethical tensions and legal challenges with respect to mental privacy and possibly require an extension of existing legal instruments

and rights ("neurorights"). Finally, the transparency and interpretability expected of machine learning methods should be proportionate to the ratio between risks and benefits. Parallel to this, the ubiquitous problem of biased data structures and ontologies that are used to train machine learning algorithms should be more widely recognized, and interdisciplinary research on how to effectively debias intelligent DSS for medical use should be promoted. These and many other ethical tensions point to the fact that proactive and interdisciplinary research as well as multistakeholder discourse and deliberation are required to ensure effective and responsible development and implementation of these emerging technologies.

18.5 Key points

- Machine learning offers exciting and promising opportunities for increasing our understanding of brain disorders and devising new strategies for diagnosis and therapy.
- Transferring autonomy of decision-making to an intelligent system may open an accountability gap in cases of grave system failures.
- The growing amounts of biomedical data and their processing in cloud-based computing environments require strong data security safeguards at the software and hardware level.
- The transparency and interpretability of machine learning methods in clinical neurology and psychiatry should be prioritized and the problem of bias should be recognized.

References

Acharya, U. R., Oh, S. L., Hagiwara, Y., Tan, J. H., & Adeli, H. (2018). Deep convolutional neural network for the automated detection and diagnosis of seizure using EEG signals. *Computers in Biology and Medicine, 100*, 270–278. https://doi.org/10.1016/j.compbiomed.2017.09.017.

Aflalo, T., Kellis, S., Klaes, C., Lee, B., Shi, Y., Pejsa, K., et al. (2015). Decoding motor imagery from the posterior parietal cortex of a tetraplegic human. *Science, 348*(6237), 906–910. https://doi.org/10.1126/science.aaa5417.

Amoroso, N., Diacono, D., Fanizzi, A., La Rocca, M., Monaco, A., Lombardi, A., et al. (2018). Deep learning reveals Alzheimer's disease onset in MCI subjects: Results from an international challenge. *Journal of Neuroscience Methods, 302*, 3–9. https://doi.org/10.1016/j.jneumeth.2017.12.011.

Amunts, K. (2018). Big-data studies need to be part of policy discussion. *Nature Human Behaviour, 1*. https://doi.org/10.1038/s41562-018-0292-9.

Barak-Corren, Y., Castro, V. M., Javitt, S., Hoffnagle, A. G., Dai, Y., Perlis, R. H., et al. (2016). Predicting suicidal behavior from longitudinal electronic health records. *American Journal of Psychiatry, 174*(2), 154–162. https://doi.org/10.1176/appi.ajp.2016.16010077.

Burns, M. N., Begale, M., Duffecy, J., Gergle, D., Karr, C. J., Giangrande, E., et al. (2011). Harnessing context sensing to develop a mobile intervention for depression. *Journal of Medical Internet Research, 13*(3), e55. https://doi.org/10.2196/jmir.1838.

Castelvecchi, D. (2016). Can we open the black box of AI? *Nature News, 538*(7623), 20. https://doi.org/10.1038/538020a.

Courtland, R. (2018). Bias detectives: The researchers striving to make algorithms fair. *Nature, 558*(7710), 357. https://doi.org/10.1038/d41586-018-05469-3.

Croskerry, P., Singhal, G., & Mamede, S. (2013a). Cognitive debiasing 1: Origins of bias and theory of debiasing. *BMJ Quality and Safety, 22*(Suppl. 2), ii58–ii64. https://doi.org/10.1136/bmjqs-2012-001712.

Croskerry, P., Singhal, G., & Mamede, S. (2013b). Cognitive debiasing 2: Impediments to and strategies for change. *BMJ Quality and Safety, 22*(Suppl. 2), ii65–ii72. https://doi.org/10.1136/bmjqs-2012-001713.

Derix, J., Iljina, O., Schulze-Bonhage, A., Aertsen, A., & Ball, T. (2012). "Doctor" or "darling"? Decoding the communication partner from ECoG of the anterior temporal lobe during non-experimental, real-life social interaction. *Frontiers in Human Neuroscience, 6*, 251. https://doi.org/10.3389/fnhum.2012.00251.

Devlin, H. (2016). *Discrimination by algorithm: Scientists devise test to detect AI bias*. The Guardian. Retrieved from https://www.theguardian.com/technology/2016/dec/19/discrimination-by-algorithm-scientists-devise-test-to-detect-ai-bias?CMP=Share_AndroidApp_Gmail.

Dipnall, J. F., Pasco, J. A., Berk, M., Williams, L. J., Dodd, S., Jacka, F. N., et al. (2016). Fusing data mining, machine learning and traditional statistics to detect biomarkers associated with depression. *PLoS One, 11*(2), e0148195. https://doi.org/10.1371/journal.pone.0148195.

Ehrlinger, J., Gilovich, T., & Ross, L. (2005). Peering into the bias blind spot: People's assessments of bias in themselves and others. *Personality and Social Psychology Bulletin, 31*(5), 680–692. https://doi.org/10.1177/0146167204271570.

Goering, S., Klein, E., Dougherty, D. D., & Widge, A. S. (2017). Staying in the loop: Relational agency and identity in next-generation DBS for psychiatry. *AJOB Neuroscience, 8*(2), 59–70. https://doi.org/10.1080/21507740.2017.1320320.

Ienca, M., & Andorno, R. (2017). Towards new human rights in the age of neuroscience and neurotechnology. *Life Sciences, Society and Policy, 13*, 5. https://doi.org/10.1186/s40504-017-0050-1.

Ioannidis, J. P. A. (2011). Excess significance bias in the literature on brain volume Abnormalities. *Archives of General Psychiatry, 68*(8), 773–780. https://doi.org/10.1001/archgenpsychiatry.2011.28.

Janghel, R. R. (2018). Deep-learning-based classification and diagnosis of Alzheimer's disease. In *Feature dimension reduction for content-based image identification* (pp. 193–217). IGI Global.

Kahneman, D., & Tversky, A. (1977). *Intuitive prediction: Biases and corrective procedures*.

Kellmeyer, P. (2017). Ethical and legal implications of the methodological crisis in neuroimaging. *Cambridge Quarterly of Healthcare Ethics: CQ: The International Journal of Healthcare Ethics Committees, 26*(4), 530–554. https://doi.org/10.1017/S096318011700007X.

Kellmeyer, P. (2018). Big brain data: On the responsible use of brain data from clinical and consumer-directed neurotechnological devices. *Neuroethics*, 1–16. https://doi.org/10.1007/s12152-018-9371-x.

Kellmeyer, P., Cochrane, T., Müller, O., Mitchell, C., Ball, T., Fins, J. J., et al. (2016). The effects of closed-loop medical devices on the autonomy and accountability of persons and systems. *Cambridge Quarterly of Healthcare Ethics: CQ: The International Journal of Healthcare Ethics Committees, 25*(4), 623–633. https://doi.org/10.1017/S0963180116000359.

Kellmeyer, P., Mueller, O., Feingold-Polak, R., & Levy-Tzedek, S. (2018). Social robots in rehabilitation: A question of trust. *Science Robotics, 3*(21), eaat1587. https://doi.org/10.1126/scirobotics.aat1587.

Kessler, R. C., van Loo, H. M., Wardenaar, K. J., Bossarte, R. M., Brenner, L. A., Cai, T., et al. (2016). Testing a machine-learning algorithm to predict the persistence and severity of major depressive disorder from baseline self-reports. *Molecular Psychiatry, 21*(10), 1366–1371. https://doi.org/10.1038/mp.2015.198.

Kimpe, T., & Tuytschaever, T. (2007). Increasing the number of gray shades in medical display systems—how much is enough? *Journal of Digital Imaging, 20*(4), 422–432. https://doi.org/10.1007/s10278-006-1052-3.

Knight, W. (2017). *Biased algorithms are everywhere, and no one seems to care*. MIT Technology Review. Retrieved from https://www.technologyreview.com/s/608248/biased-algorithms-are-everywhere-and-no-one-seems-to-care/.

Kohler, F., Gkogkidis, C. A., Bentler, C., Wang, X., Gierthmuehlen, M., Fischer, J., et al. (2017). Closed-loop interaction with the cerebral cortex: A review of wireless implant technology. *Brain-Computer Interfaces, 4*(3), 146–154. https://doi.org/10.1080/2326263X.2017.1338011.

Kreitmair, K. V., Cho, M. K., & Magnus, D. C. (2017). Consent and engagement, security, and authentic living using wearable and mobile health technology. *Nature Biotechnology, 35*(7), 617–620. https://doi.org/10.1038/nbt.3887.

Kreit, E., Mäthger, L. M., Hanlon, R. T., Dennis, P. B., Naik, R. R., Forsythe, E., et al. (2013). Biological versus electronic adaptive coloration: How can one inform the other? *Journal of The Royal Society Interface, 10*(78), 20120601. https://doi.org/10.1098/rsif.2012.0601.

Lee, J. S., Kim, C., Shin, J.-H., Cho, H., Shin, D., Kim, N., et al. (2018). Machine learning-based individual assessment of cortical atrophy pattern in Alzheimer's disease spectrum: Development of the classifier and longitudinal evaluation. *Scientific Reports, 8*(1), 4161. https://doi.org/10.1038/s41598-018-22277-x.

Marzano, L., Bardill, A., Fields, B., Herd, K., Veale, D., Grey, N., et al. (2015). The application of mHealth to mental health: Opportunities and challenges. *The Lancet Psychiatry, 2*(10), 942–948. https://doi.org/10.1016/S2215-0366(15)00268-0.

Mohr, D. C., Zhang, M., & Schueller, S. M. (2017). Personal sensing: Understanding mental health using ubiquitous sensors and machine learning. *Annual Review of Clinical Psychology, 13*(1), 23–47. https://doi.org/10.1146/annurev-clinpsy-032816-044949.

Moradi, E., Pepe, A., Gaser, C., Huttunen, H., & Tohka, J. (2015). Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects. *NeuroImage, 104*, 398–412. https://doi.org/10.1016/j.neuroimage.2014.10.002.

Palminteri, S., Lefebvre, G., Kilford, E. J., & Blakemore, S.-J. (2017). Confirmation bias in human reinforcement learning: Evidence from counterfactual feedback processing. *PLoS Computational Biology, 13*(8), e1005684. https://doi.org/10.1371/journal.pcbi.1005684.

Patel, M. J., Khalaf, A., & Aizenstein, H. J. (2016). Studying depression using imaging and machine learning methods. *NeuroImage: Clinical, 10*, 115–123. https://doi.org/10.1016/j.nicl.2015.11.003.

Schirrmeister, R., Gemein, L., Eggensperger, K., Hutter, F., & Ball, T. (2017). Deep learning with convolutional neural networks for decoding and visualization of EEG pathology. In *Signal Processing in Medicine and Biology Symposium (SPMB), 2017 IEEE* (pp. 1–7). IEEE.

Schirrmeister, R. T., Springenberg, J. T., Fiederer, L. D. J., Glasstetter, M., Eggensperger, K., Tangermann, M., et al. (2017). Deep learning with convolutional neural networks for EEG decoding and visualization. *Human Brain Mapping, 38*, 391–5420. https://doi.org/10.1002/hbm.23730.

Shah, H., Albanese, E., Duggan, C., Rudan, I., Langa, K. M., Carrillo, M. C., et al. (2016). Research priorities to reduce the global burden of dementia by 2025. *The Lancet Neurology, 15*(12), 1285–1294. https://doi.org/10.1016/S1474-4422(16)30235-6.

Smith, B. W., & Slack, M. B. (2015). The effect of cognitive debiasing training among family medicine residents. *Diagnosis, 2*(2), 117–121. https://doi.org/10.1515/dx-2015-0007.

Su, L., Blamire, A. M., Watson, R., He, J., Hayes, L., & O'Brien, J. T. (2016). Whole-brain patterns of 1H-magnetic resonance spectroscopy imaging in Alzheimer's disease and dementia with Lewy bodies. *Translational Psychiatry, 6*(8), e877. https://doi.org/10.1038/tp.2016.140.

Völker, M., Schirrmeister, R. T., Fiederer, L. D. J., Burgard, W., & Ball, T. (2018). Deep transfer learning for error decoding from non-invasive EEG. In *2018 6th international conference on bBrain-Computer Interface (BCI)* (pp. 1–6). https://doi.org/10.1109/IWW-BCI.2018.8311491.

Wager, T. D., & Woo, C.-W. (2017). Imaging biomarkers and biotypes for depression. *Nature Medicine, 23*(1), 16–17. https://doi.org/10.1038/nm.4264.

Wilson, R. E., Gosling, S. D., & Graham, L. T. (2012). A review of facebook research in the social sciences. *Perspectives on Psychological Science, 7*(3), 203–220. https://doi.org/10.1177/1745691612442904.

Winograd-Cort, D., Haeberlen, A., Roth, A., & Pierce, B. C. (2017). A framework for adaptive differential privacy. *Proceedings of the ACM on Programming Languages, 10*(29), 1–10, 1 (ICFP) https://doi.org/10.1145/3110254.

Yuste, R., Goering, S., Arcas, B. A. y, Bi, G., Carmena, J. M., Carter, A., et al. (2017). Four ethical priorities for neurotechnologies and AI. *Nature News, 551*(7679), 159. https://doi.org/10.1038/551159a.

Further reading

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., et al. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature, 542*(7639), 115–118. https://doi.org/10.1038/nature21056.

Mackenzie, C., & Stoljar, N. (2000). *Relational autonomy: Feminist perspectives on automony, agency, and the social self*. Oxford University Press.

Yu, K.-H., Zhang, C., Berry, G. J., Altman, R. B., Ré, C., Rubin, D. L., et al. (2016). Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. *Nature Communications, 7*, 12474. https://doi.org/10.1038/ncomms12474.