Science and Engineering Ethics (2020) 26:2455–2460
https://doi.org/10.1007/s11948-020-00241-1

COMMENTARY

Keeping the "Human in the Loop" in the Age of Artificial Intelligence

Accompanying Commentary for "Correcting the Brain?" by Rainey and Erden

Fabrice Jotterand[1] · Clara Bosco[1]

Published online: 8 July 2020

Abstract
The benefits of Artificial Intelligence (AI) in medicine are unquestionable and it is unlikely that the pace of its development will slow down. From better diagnosis, prognosis, and prevention to more precise surgical procedures, AI has the potential to offer unique opportunities to enhance patient care and improve clinical practice overall. However, at this stage of AI technology development it is unclear whether it will *de*-humanize or *re*-humanize medicine. Will AI allow clinicians to spend less time on administrative tasks and technology related procedures and more time being present in person to attend to the needs of their patients? Or will AI dramatically increase the presence of smart technology in the clinical context to a point of undermining the humane dimension of the patient–physician relationship? In this brief commentary, we argue that technological solutions should be only integrated into clinical medicine if they fulfill the following three conditions: (1) they serve human ends; (2) they respect personal identity; and (3) they promote human interaction. These three conditions form the moral imperative of *humanity*.

Keywords Humanity · Ethical imperatives · Patient–physician relationship · Artificial intelligence · Psychiatry · Neurointerventions

Introduction

The benefits of Artificial Intelligence (AI) in medicine are unquestionable and it is unlikely that the pace of its development will slow down. From better diagnosis, prognosis, and prevention to more precise surgical procedures, AI has the potential to offer unique opportunities to enhance patient care and improve clinical practice

✉ Fabrice Jotterand
fjotterand@mcw.edu

[1] Center for Bioethics and Medical Humanities, Medical College of Wisconsin, 8701 Watertown Plank Road, Milwaukee, WI 53226, USA

overall. However, at this stage of AI technology development it is unclear whether it will *de*-humanize or *re*-humanize medicine. Will AI allow clinicians to spend less time on administrative tasks and technology related procedures and more time being present in person to attend to the needs of their patients? Or will AI dramatically increase the presence of smart technology in the clinical context to a point of undermining the humane dimension of the patient–physician relationship?

It is with this set of questions that we approach the analysis of the article by Stephen Rainey and Yasemin J. Erden. They rightly note that the use of AI in psychiatry could be detrimental to patients in various ways. First, while AI neurotechnological devices may help in the detection and modification of neural activity, the nature of brain interventions in not clear, particularly regarding how they might affect concepts such as free will, agency, responsibility, and perception. To be sure, neuroscience as the science of the brain (neural states) is increasingly interacting with psychiatry (mental states) which could result in "neuropsychiatric accounts of human cognition and behaviour" (p. 4), thus reducing human cognition and behavior to neuroscientific norms. The second important point Rainey and Erden make, is how the clinical encounter could be reconfigured through technological means. They point out that AI could have a role of decision maker, hence pitting "conventional intelligence (HI) and AI in opposition" (p. 14). AI provides valuable and powerful tools for brain interventions aimed at altering neural states and subsequently behavior.

Human Reasons and the "Datafied Brain"

Despite these potential capabilities, there are concerns about the inability of AI to account for the realities of the human condition in its social, cultural, and embodied dimensions. The *modus operandi* of AI-enabled neurotechnology is based on statistical methods that do not include "human reasons" in conceptualizing cognition and behavior (i.e. beliefs, desires, motivations, and intentions). As Rainey and Erden rightly remark "to reduce human reasons to simple gap-filling causes of behaviour is to miss details informed by a rich phenomenological experience of rationality and behaviour" (p. 19). Psychiatry, due to the nature of its practice, is inherently confronted by the realities of the human condition and is always in the process of evaluating how "human reasons" interact with, and shape patients' behavior and identity. Human demeanor cannot be understood nor reduced to mere neural activity. It is certainly the case that technological solutions via neural-activity data can aid in diagnosis and establishing treatment options, but "the aim should always be to include the agent, and to presume agency" (p. 21). Therefore, the datafied brain (i.e., detectable neural activity) is not "the proper brain" (p. 11). As Rainey and Erden explain "diagnosing problems of mind in terms of neurophysical anomaly omits key details about what mindedness consists in" (p. 12). Furthermore, as already discussed, mental states and neural processes alike are not experienced in a vacuum but are part of a rich social context with norms, culture, language, reasoning, other dimensions of human behaviors, and body language. In short, the mind is "an open system" (p. 12) not determined nor explained by mere statistical processes but by complex reasoning capacities that convey agency. Understanding the nature and role of neural states provides a basis to explain human behavior.

The use of AI-enabled neurotechnology in psychiatry, however, adds another layer of explanatory power. Traditional, interpersonal modes of practice in psychiatry are, to a certain extent, challenged by neural explanation based on statistical models. The specificity in diagnosis and prognosis that AI-enabled neurotechnology affords, is unable to integrate "human reasons" into clinical judgments about patient behavior. AI-enabled neurotechnology "by-passes" the agency of the user as mental states are generated through neurointerventions (e.g. a closed neurotechnological device to detect and modify neural activity as stated by Rainey and Erden). It might not be an exaggeration, as Rainey and Erden bluntly conclude, that there is "no human in the loop on this neurotechnology model" (p. 9). Will AI-enabled neurotechnology *de*-humanize or *re*-humane in medicine? It might allow physicians to spend more time with patients, hence *re-humanizing* clinical practice, but the question still remains about whether clinical interventions themselves might *de-humanize* patients by undermining their agency.

The Ethical Imperative of Humanity

AI-enabled neurotechnology will be technologically useful and clinically relevant but also problematic since it has the potential to view psychiatric patients as reducible to neuroscientific norms. Therefore, we argue that the technological solutions should be only integrated into clinical medicine if they fulfill the following three conditions: (1) they serve human ends; (2) they respect personal identity; and (3) they promote human interaction. These three conditions form what we call the moral imperative of humanity.

The ethical framework we suggest is not limited to the ethical imperative of humanity. Rather humanity is the foundational concept from which the other five derive: information, transparency, participation, consensus, accountability. Before we take a deeper dive into humanity, we want to briefly outline the other moral imperatives which are grounded in three main categories. The first category includes *information* and *transparency* to deal with how human beings engage with technology. To anticipate the potentially deleterious implications of AI-enabled neurotechnology, gathering pertinent information about AI must always be at the forefront of any robust analysis. The complexity of AI technologies requires knowledge acquisition about their nature and abilities across relevant disciplines. Further, it is important to maintain *transparency* by informing all stakeholders including society at large since they will be the beneficiary of AI in medicine. To this end, communicating risks and benefits of AI cannot be limited to the context of their particular use, e.g., during the consenting process in the clinical context. Transparency is paramount to ensure a responsible and ethical implementation of AI in the clinical and social contexts. The second category concerns the way the technology might affect patient care and includes *participation* and *consensus*. As stated above, the public, as patients or potential patients, will be affected by the use of AI. Hence, strategies should be implemented to include all stakeholders in the analysis of the ethical, social, and regulatory implications of AI. This effort of course is not without challenges, in particular how to build consensus among stakeholders. Due to the limited scope of this article, it is not our intention to address this issue here. Rather, the importance of creating an environment conducive to develop ethical norms that responsibly guide public policies and establish standards of practice is crucial to harvest the potential benefits of AI-enabled neurotechnologies. This point cannot be stressed enough in light of recent advances in neurotechnology (brain–computer interfaces; consumer neurotechnologies), including the gathering of brain data by various third parties to generate neural profiles. The last category in our ethical framework focuses specifically on health care, that is, how physicians should engage with AI-enabled neurotechnologies and how current and future physicians ought to be trained. Fostering responsible development and implementation of AI in health care will demand that physicians be accountable for patient safety—*accountability*.

We hold that these five ethical imperatives must be supported by *humanity* as the foundation tenet. Humanity encompasses, as already stated, three essential ideas. The first is that technology should always serve human ends. This stance assumes that AI will always be somewhat dependent on human agency. However, considering how technology in general is pervasive in our everyday lives, and even integrated in the human body, it is not obvious that humans are fully in control anymore. Studies have demonstrated the addictive nature of smart phones and there are well-known cases of individuals creating strong psychological bonds with humanoid robots. Our point is not to depict an apocalyptic scenario about the dismal future of humanity where intelligent robots would take over. Rather, our contention is to stress how technology is shaping human beings rather than human beings shaping their surroundings through technology to create a better place for individuals to flourish. Almost 70 years ago, German philosopher Martin Heidegger already anticipated this shift. In his essay *The Question Concerning Technology* (1954/1977) he discusses the essence of technology. He asserts that modern technology is a way of revealing, which in turn is a mode of ordering reality or seeing, or what he calls "enframing" (*Gestell*). Technology thus places humans in a position of orderers of the world, which includes themselves as part of these objects of manipulation and control. This means that technology provides the power to control and manipulate the very essence of human nature, including the brain; or, put another way, human beings have become orderers and orderables through (neuro)technological manipulation.

The second idea underlying humanity is respect for personal identity. Many factors may change personal identity: traumatic events, neurotechnologies, disease, and therapeutic interventions can affect one's sense of self. From a clinical standpoint, the ethical dilemma resides in the use of devices, AI-enabled neurotechnology for instance, to diagnose and treat brain disorders that may change personal identity and character traits in a patient (Jotterand and Giordano 2011). Assuming that these interventions are "technically right" and "ethically good" it is not clear where the decisional burden should be on the physician or the patient. Elsewhere, one of the authors of this article (Jotterand) argues that because neurointerventions in psychiatry are provided in a clinical context, physicians might be ultimately responsible for the evaluation of their risks and benefits: "therapeutic interventions are aimed at preventing and/or reversing any negative changes, by restoring health and normalizing particular functions. To this end, neurotechnological interventions ought to be rendered so as to restore the patient's state of health prior to the occurrence of the disorder (at least as much as possible) by normalizing brain functions" (Jotterand and Giordano 2011, p. 482). This approach assumes a strong emphasis on human interaction and the specific role of the physician in prioritizing the well-being of the patient.

The third dimension of humanity is that AI-enabled technology should promote human interaction. This is an important dimension since ultimately patients' outcomes are at stake and there are practical consequences regarding patient autonomy, safety, and self-conception. The clinical encounter is the cornerstone of doctoring. The physician and the patient meet as two equal moral agents to address a state of dis-ease and determine the best course of action to normalize, as much as possible, brain functions or other ailments. Edmund D. Pellegrino and David C. Thomasma have formulated five imperative defining characteristics of the clinical encounter: (1) inequality of power in the relationship due to the state of dependence and vulnerability of the patient; (2) the fiduciary nature of the relationship which assume trust and the absence of coercion or manipulation; (3) the moral dimension of medical decisions—clinical judgements are a combination of technical and moral factors; (4) the moral nature of medical knowledge which assumes particular obligations on the part of clinicians; and (5) the moral complicity of the physician, that is, the clinical encounter presumes the collaboration between the two parties in a context of trust and collaboration (Pellegrino and Thomasma 1993; see also Jotterand and Giordano 2011 in the context neurotechnology). Each of these imperatives presumes a strong emphasis on human interaction in the clinical context. Clinical practice will be enhanced by AI-enabled neurotechnologies but it is unlikely, at least based on the current stage of AI development, that technology will address a question such as suffering. The phenomenology of suffering appears to be one of the most difficult conundrums to solve in the complex experience of the human condition. What makes this experiential state particularly difficult is uncertainty of what suffering encompasses, how suffering relates to pain, and ultimately what suffering means. Furthermore, the concept of suffering has changed throughout the ages, providing in each stage of human knowledge a new "insight." Primitive cultures tended to perceive human suffering in terms of divine punishment or divine activity whereas from the Greek period on, suffering and its manifestations such as diseases or mental illnesses have been progressively reduced to psycho-biological phenomena and to a social construct. Psychiatric conditions, such as psychological abuse or post-traumatic stress disorder (PTSD) for instance, can be managed by suppressing the emotional surge triggered by certain circumstances. However, any technology will not be able to replace a human connection. Even if a robot would achieve an unprecedented level of sophistication, the machine will never understand the human condition in its biological, psychological, social, and spiritual dimensions.

Concluding Remarks

Keeping the "human in the loop" is a quintessential dimension to clinical practice. It may be the case that AI-enabled neurotechnologies will afford physicians to spend more time with their patients in a meaningful way. However, as Rainey and Erden clearly demonstrate in their article, the reductive potential of these technologies should be of paramount concern. This is not to negate the worthiness of the enterprise to understand the brain, its neural states, and ultimately human behavior as well as causes of mental disorders and abnormal behavior. We argue that any implementation of AI-enabled neurotechnologies should be guided by the ethical imperative of humanity. We hope that our modest attempt to provide an ethical framework will stimulate further constructive debates on the responsible implementation of AI-enabled neurotechnologies in psychiatry.

References

Heidegger, M. (1954/1977). The question concerning technology. In: D. F. Krell (Ed.), *Basic writings*. New York: Harper & Row.

Jotterand, F., & Giordano, J. (2011). Transcranial magnetic stimulation, deep brain stimulation and personal identity: Ethical questions, and neuroethical approaches for medical practice. *International Review of Psychiatry, 23*(5), 476–485.

Pellegrino, E. D., & Thomasma, D. C. (1993). *The virtues in medical practice*. New York: Oxford University Press.

Rainey, S. & Erden, Y. J. (2020). Correcting the brain? The Convergence of neuroscience, neurotechnology, psychiatry, and artificial intelligence.