Artificial Intelligence in Healthcare: Rethinking the Notions of Responsibility, Causal Inference and Empathy

Panagiotis Kormas[1] and **Antonia Moutzouri**[2]
[1, 2] National and Kapodistrian University of Athens

Abstract: Artificial Intelligence (AI) systems have demonstrated precision at similar or superior degree in relation to healthcare professionals. However, several ethical debates have focused on the issues of accountability, explainability and clinician-patient trust. Deep Learning systems generate largely uninterpretable results, thus directly challenging the concept of responsible agency and moral responsibility. Furthermore, epistemologically, it is different to identify a correlation between symptoms and diseases than to demonstrate a causal explanation. The incorporation of causal reasoning seems critical in harnessing all the benefits and surpassing human expert capability in demanding clinical decisions. The physician-patient relationship is also of paramount importance in the therapeutic outcome and how empathy is reproduced in systems may be crucial for the delivery of moral medical care. The dynamics of AI in healthcare urge for a rethinking of notions of responsibility, causal inference and empathy as they are key constructs in framing the proper ethical foundation.

Keywords: artificial intelligence; deep learning; ethical AI; responsible agent; causality; cognitive and emotional empathy; trust

I. Introduction

The technological progress experienced by humanity today is known as the 4th Industrial Revolution, or Industry 4.0. As the term itself indicates, the adaptation of new technologies is accompanied by an abrupt and deep change within the economic systems and the social structures. This era's Industrial Revolution pertains to the development and exploitation of holistic digital systems that have the capacity to integrate the digital, the physical and the biological realms across all sectors.[1]

Among the driving forces for implementing artificial intelligence algorithms in the medical practice are the increasingly digital collection methods of health data, the excellent early results of imaging analysis and the need for fast decision making in the case of extremely urgent and critical conditions. Moreover, the parallel development of personalised solutions in the healthcare domain has increased the interest for AI-driven recommendations.[2] On the ethical implications, the technological advances on the healthcare sector are of particular interest not only because of the sensitivity of private health-related data of individuals but also because of the critical importance of diagnostic and therapeutic decision-making processes.

Artificial Intelligence (AI) is a broad domain that encompasses fields such as Machine Learning (ML), Artificial Neural Network (ANN) and Deep Learning (DL). It is devoted to building artificial entities and, as a self-standing discipline, it has its origins in the mid-20th century.[3] However, it has seen significant development over the last decades while today's importance is mostly understood when referring to intelligent machines endowed with learning, reasoning and adaptation capabilities. ML gives the capability to AI to solve problems based on data acquired from a given context while not demanding explicit programming. ANN is an evolved process of ML inspired

[1] Klaus Schwab, "The Fourth Industrial Revolution: What it Means, How to Respond," *World Economic Forum*, January 14, 2016, https://www.weforum.org/agenda/2016/01/the-fourth-industrial-revolution-what-it-means-and-how-to-respond/.

[2] Adam Bohr, and Kaveh Memarzadeh, "The Rise of Artificial Intelligence in Healthcare Applications," in *Artificial Intelligence in Healthcare*, 25-60 (London: Academic Press, 2020), 25-27.

[3] John McCarthy, et al., "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955," *AI Magazine* 27, no. 4 (2006): 12.

by the model of the human brain.[4] Ultimately, DL is large neural-network-style model with multiple layers of representation.[5]

A plethora of scientific publications underscore the precision of AI medical tools, demonstrating that algorithms can achieve precision at similar or superior degree in relation to humans in detecting skin cancer, heart arrhythmia, and Alzheimer Disease. The hope is that AI will facilitate timely detection, allow for improved diagnosis, and enhance human reasoning and clinical decision-making capacity.[6]

The prevalence of AI technologies in almost all domains of human life and its highly promising potential in healthcare have raised many debates on the ethical implications of its deployment. The clinical setting in particular constitutes a complex environment where AI could be entrusted with life-and-death decisions. The unprecedented technical achievements of AI alongside the dynamic contemporary environment urge for a rethinking of notions of responsibility, causal inference and empathy as they are key constructs in framing the proper ethical foundation.

II. Explainability and responsibility

Many state-of-the-art AI models are constructed on DL techniques which, by nature, enclose inner workings into which it is difficult or even impossible to gain insight. In contrast to more conventional ML approaches, deep neural networks, inspired by the human biological neural system, operate by propagating the input data through multiple layers while not just executing the pre-determined instructions. Thus, within their so-called

[4] Wesam Salah Alaloul, and Abdul Hannan Qureshi, “Data Processing Using Artificial Neural Networks,” in *Dynamic Data Assimilation - Beating the Uncertainties*, ed. Dinesh G. Harkut (London: IntechOpen, 2020).

[5] Brenden M. Lake, et al., “Building Machines that Learn and Think like People,” *The Behavioral and Brain Sciences* 40 (2017): 1.

[6] Angeliki Kerasidou, “Artificial Intelligence and the Ongoing Need for Empathy, Compassion and Trust in Healthcare,” *Bulletin of the World Health Organization* 98, no. 4 (2020): 246.

"black box," they make predictions and reach decisions similar to how humans do but without 'communicating' their reasons to do so.[7] In fact, the established belief that there is a trade-off between accuracy and interpretability[8] may have intensified the development of AI black boxes in the name of increased performance.

On the one end of a neural network there is the input layer which receives data from the outer environment and transfers it in the inner structure of the network while on the other end the output layer produces the results on the basis of the processing conducted by the system. Between input and output, there are intermediate layers, namely hidden layers, which perform the processing of the ANN. Each layer is a linear array compiled of various nodes, similar to neurons, which correspond to the various inputs introduced either by the external environment (for the input layer) or by the previous layer (for any intermediate and the output layer). The number of hidden layers (depth), as well as the number of nodes in each layer (width) together with the designed path determine the network's topology.[9]

Hence, DL advances have led to complicated AI networks that generate inherently uninterpretable models to human users, sacrificing interpretability for prediction accuracy.[10] Nevertheless, there is a consensus among the research community that the concept of responsible agency and – in turn – moral responsibility, is closely related to the degree of explainability of AI algorithms.

Especially in the healthcare sector, how clear the functioning of a model is, possesses a key importance as it is connected to accountability and transparency issues. Logistic or linear

[7] Yavar Bathaee, "The Artificial Intelligence Black Box and the Failure of Intent and Causation," *Harvard Journal of Law and Technology* 31, no. 2 (2018): 893.

[8] Reubern Binns, "Algorithmic Accountability and Public Reason," *Philosophy and Technology* 31 (2018): 553.

[9] Alaloul, and Qureshi.

[10] Mengnan Du, Ninghao Liu, and Xia Hu, "Techniques for Interpretable Machine Learning," *Communications of the ACM* 63, no. 1 (2019): 69.

regression models can be interpreted when a human attempts to understand the relationship between variables as there are certain statistical parameters that one can refer to. In an ANN, how any given output affects the final outcome depends on the complex interaction of values embedded in a highly entangled web of connections system. Human-scale cognition is lacking the capacity to understand how and most of the ANNs arrive at any particular decision.

Several articles have been published on issues of interpretability or explainability, and, despite being two terms that are frequently used interchangeably, they actually describe different features of AI. An interpretable system is one where "a user cannot only see but also study and understand how inputs are mathematically mapped to outputs." Explainability describes the "capability of understanding the work logic in the ML algorithms."[11]

The origins of the requirements for moral responsibility date back to the Greek ancient philosophy; following the Aristotelian requirements for responsibility, namely control and knowledge, we infer that one is responsible if they have a sufficient level of control over an action and be knowledgeable of what is pertaining to the action.[12]

The traditional responsibility ascription cannot be applied in the case of ML algorithms as the developer of the model is, in principle, not capable of intervening in the course of action of the process. This incompatibility between the moral framework of society and the design principles of machine learning models has been characterised as a "responsibility gap."[13]

[11] Amina Adadi, and Mohammed Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," *IEEE Access* 6 (2018): 52141.

[12] Mark Coeckelbergh, "Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability," *Science and Engineering Ethics* 26, no. 4 (2020): 2054.

[13] Andreas Matthias, "The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata," *Ethics and Information Technology* 6 (2004): 177.

Apart from principally being a philosophical issue, since agency is connected to responsibility, the problem of responsibility attribution in the contemporary context is ultimately practical. Building on the concept of the 'responsibility gap,' the more advanced the technology – reaching to the point of carrying intelligence that may be initiated by an algorithmic design but then evolved 'on its own learning,' – the harder it is to ascribe blame to any human or corporate entity along the chain of development, employment and decision-making of the AI system. Furthermore, in the healthcare sector there are multiple actors, among whom the algorithm designer, the data provider, the healthcare institution implementing the AI system and the healthcare professional who uses it. This multiplicity further obscures the attribution of responsibility. What should also be mentioned is that accountability does not exclusively apply in the cases of something going wrong when following AI outputs but also when physicians decide to override the recommendations.[14]

In the last decades, considering the aforementioned landscape and in view of the breadth of practical circumstances in which AI tools and autonomous robots are present in our lives, the concept of artificial or virtual moral agency and responsibility has been proposed and much debated.[15] On the one hand, the argument that the inability of AI to understand the shared moral values among a human community renders it ineligible for moral responsibility, and on the other, the search for less anthropocentric definitions for moral agency,[16] have shaped a dynamic and highly fluid environment in which traditional philosophical concepts have been revisited.

III. Correlation vs. causation

The medical sector is overwhelmed with an ever-increasing amount of biologic, biometric and electronic health data. Big

[14] Kerasidou, 247.

[15] Dorna Behdadi, and Christian Munthe, "Normative Approach to Artificial Moral Agency," *Minds & Machines* 30 (2020): 212.

[16] Mihaela Constantinescu, et al., "Understanding Responsibility in Responsible AI. Dianoetic Virtues and the Hard Problem of Context," *Ethics and Information Technology* (2021): 3.

data in medicine has the potential to reveal formerly unidentified health patterns and ultimately new therapies.[17] Improving predictions as to when an individual is at risk of an acute health event or a chronic disease or even relying on highly accurate digital diagnostic tools is of paramount importance in delivering healthcare. Notwithstanding, it is not the data *per se* but the algorithms encoding reasoning and knowledge that can actually be game-changing in the medical sector.[18]

Beyond the issue of interpretability and explainability, but closely related to the notion of statistical inference, is causal inference. AI models have the capacity to identify patterns within enormous datasets, but its capacity to go beyond data-driven association is now considered instrumental in qualitatively transforming medicine.[19]

The employment of AI, and particularly ML models, may carry the danger of conflating causation with association. In the diagnosis procedure, it is another thing to identify correlations between patient data and disease occurrences and another to determine the underlying cause of a patient's symptoms. The definition of diagnosis is reminder of this distinction: "the identification of the diseases that are most likely to be causing the patient's symptoms, given their medical history."[20] In the scope of the definition of this medical practice, the drawing of a causal model of how a disease relates to the outcomes (symptoms) is fundamental in the subsequent clinical decision-making process.

According to the "ladder of causation," proposed by Judea Pearl, there are three defining levels of cognitive ability – name-

[17] Mary Mallappallil, et al., "A Review of Big Data and Medical Research," *SAGE Open Medicine* 8 (2020): 1.

[18] Ziad Obermeyer, and Ezekiel J. Emanuel, "Predicting the Future – Big Data, Machine Learning, and Clinical Medicine," *The New England Journal of Medicine* 375, no. 13 (2016): 1217.

[19] Yoshihiko Raita, et al., "Big Data, Data Science, and Causal Inference: A Primer for Clinicians," *Frontiers in Medicine* 8, (2021): 11.

[20] Jonathan G. Richens, Ciaran M. Lee, and Saurahb Johri, "Improving the Accuracy of Medical Diagnosis with Causal Machine Learning," *Nature Communications* 11 (2020): 2.

ly seeing, doing and imagining – which, accordingly, entail association, intervention and counterfactuals.[21] In order to perform counterfactual-based tasks, one has to first be able to respond to association and intervention problems. Expert knowledge is what would then make the shift to the upper level; one has to be able to specify the question and to describe the causal structure. In the clinical setting, biological knowledge is necessary to shift from association and intervention to the counterfactual framework, as without it no causal effects could be defined and the causal structure could not be specified.[22]

The truth of counterfactuals denotes a causal link between a 'cause' and an 'effect.' However, causal effects cannot be measured by technology systems that operate exclusively on data alone, even if data are vast and learning algorithms are very deep. Maybe it is for this reason that diagnostic algorithms have not delivered the desired outcomes on what concerns the accuracy in differential diagnosis, one of the most important but also challenging tasks in a physician's clinical practice.[23]

Causality as a concept has been of paramount importance in the long history of human effort to explain and understand phenomena in the universe. It is a concept intimately relating to intellectual understanding and one that has fostered long-standing debates in the history of philosophical literature. Causality stretches back to the times of Aristotle and extends to modern debates in contemporary sciences. Aristotle, in his theory of causality, recognised four causes: the material, the formal, the efficient and the final, all of which are involved in the explanation process and shape the theoretical framework for the study of the natural world. In analysing causation, David Hume in the 1700s acknowledged regularity as the major feature of causation; hence,

[21] Mark J. Bishop, "Artificial Intelligence Is Stupid and Causal Reasoning Will Not Fix It," *Frontiers in Psychology* 11 (2021): 10.

[22] Raita, et al., 6.

[23] Stuart F. Leeds, et al. "Teaching Heuristics and Mnemonics to Improve Generation of Differential Diagnoses," *Medical Education Online* 25, no. 1 (2020): 1.

beyond a cause temporally preceding its effect and being contiguous to it, it is necessary that "all objects similar to the cause are in a 'like relation' to objects similar to the effect."[24] Hume has also identified causation through the notion of counterfactual: "… where, if the first object had not been, the second never had existed." However, it is after the work of David Lewis that the concept became elaborated and more important.[25]

In the healthcare arena, the early 20th century was a period when studies of cancer and chronic diseases shifted the interest from strictly identifying causes of the diseases to recognising patterns and identifying groups of people at increased risk. In fact, epidemiology is described "the study of the distribution and determinants of disease patterns in human populations" in contemporary definitions.[26] This strategic turn was systematic and aimed at more targeted healthcare interventions. These new models of causation may have created an environment where the concept of causation experienced a radical change and the distinction between prediction and causal inference may have been de-emphasized.

However, recent results highlight the importance of counterfactual reasoning in the medical diagnosis field, showing that counterfactual algorithms can be designed that position the accuracy in the top 25% of physicians, contrary to merely associative AI tools which achieved accuracy in the top 48%.[27]

Anyhow, data-driven prediction AI can only indicate towards a decision, but it is causal inference that can support the deci-

[24] Andreas Holger, and Mario Guenther, "Regularity and Inferential Theories of Causation," *The Stanford Encyclopedia of Philosophy,* Fall 2021, ed. Edward N. Zalta, https://plato.stanford.edu/archives/fall2021/entries/causation-regularity/.

[25] Yu-Liang Chou, et al., "Counterfactuals and Causability in Explainable Artificial Intelligence: Theory, Algorithms, and Applications," *Information Fusion*, pre-proof.

[26] Mark Parascandola, "The Epidemiologic Transition and Changing Concepts of Causation and Causal Inference," *Revue d' Histoire des Sciences* 64, no. 2 (2011): 244.

[27] Richens, Lee, and Johri, 2.

sion-making process. The incorporation of causal reasoning, and thereupon the human domain expertise, in machine learning algorithms may be crucial in harnessing all the benefits that AI can provide and in surpassing human expert capability, especially in certain domains.[28, 29]

IV. Humanization and empathy in medical care

One expected benefit from AI in the healthcare sector is that improvement of efficiency will allow clinicians to focus on the human side of care, directly engaging with patients, building a relationship of trust, exercising empathy while using judgment to guide and advise.[30] This is particularly meaningful as it has been shown that establishing a relationship of mutual trust is central for effective medical care while the patient enjoys an improved experience and clinical outcomes.[31] Apart from considerations that have to do with the challenges to actually realise such a potential, e.g. driven by the profit-oriented business models in healthcare,[32] it has been argued that AI inherently lacks the potential to demonstrate empathy characteristics.

In the history of scientific research, empathy has progressed from a predominantly cognitive construction to one that also includes affective, imaginative and relational dimensions.[33] As

[28] Ibid., 7.

[29] Rama K. Vasudevan, et al., "Off-the-shelf Deep Learning Is Not Enough, and Requires Parsimony, Bayesianity, and Causality," *npj Computational Materials* 7, no. 16 (2021): 5.

[30] Alexander L. Fogel, and Joseph C. Kvedar, "Artificial Intelligence Powers Digital Medicine," *npj Digital Medicine* 1, no. 5 (2018): 1.

[31] John M. Kelley, et al., "The Influence of the Patient-Clinician Relationship on Healthcare Outcomes: A Systematic Review and Meta-Analysis of Randomized Controlled Trials," *PLOS ONE* 9, no. 4, e94207 (2014): 1.

[32] Matthew Nagy, and Bryan Sisk, "How Will Artificial Intelligence Affect Patient-Clinician Relationships?" *American Medical Association Journal of Ethics* 22, no. 5 (2020): E397.

[33] Laurence Tan, et al., "Defining Clinical Empathy: A Grounded Theory Approach from the Perspective of Healthcare Workers and Patients in a Multicul-

described in the *Stanford Encyclopaedia of Philosophy*, empathy encompasses "a wide range of psychological capacities that are thought of as being central for constituting humans as social creatures allowing us to know what other people are thinking and feeling, to emotionally engage with them, to share their thoughts and feelings, and to care for their well-being."[34] Although the observer's emotional state is isomorphic with the other person's state, the observer is aware that the other person is the source of their state, thus differentiating empathy from emotional contagion.[35] Compassion and sympathy are analogous terms in so as the representation of the emotions of others is present, however empathy is distinct in that requires the synchronisation of the emotional states.[36] Empathy includes feelings that are similar to what the other feels and not feelings for how the other person feels. Moreover, these concepts represent different neurobiological phenomena.

Empathy is a complex phenomenon, and its contemporary notion is divided into a cognitive (cognitive empathy) and an affective (emotional empathy) element; cognitive empathy relates to the capacity for taking another individual's perspective, also referred to as mentalising, perspective-taking or theory of mind. On the other hand, affective empathy is caused by sharing the emotions of another agent through observation or imagination of their experience.[37, 38] Although emotional and cognitive

tural Setting," *British Medical Journal Open* 11, no. 9, e045224 (2021): 1-2.

[34] Karsten Stueber, "Empathy," *The Stanford Encyclopedia of Philosophy*, Fall 2019, ed. Edward N. Zalta, https://plato.stanford.edu/archives/fall2019/entries/empathy/.

[35] Frederique de Vignemont, and Tania Singer, "The Empathic Brain: How, When and Why?" *Trends in Cognitive Sciences* 10, no. 10 (2006): 435.

[36] Minoru Asada, "Development of Artificial Empathy," *Neuroscience Research* 90 (2015): 43.

[37] Patricia L. Lockwood, et al., "Individual Differences in Empathy are Associated with Apathy-Motivation," *Scientific Reports* 7 (2017): 1.

[38] Meghan L Healey, and Murray Grossman, "Cognitive and Affective Perspective-Taking: Evidence for Shared and Dissociable Anatomical Sub-

aspects are largely acknowledged as distinct processes taking place in separate brain regions, they may engage in a more complicated relationship, as e.g., in cases of metacognition where one can observe their selves from another's perspective.[39] The extent of the empathic experience is further regulated by executive functions, such as attention and self-regulation, resulting in empathic concern, i.e., sympathy.[40]

Empathy is an essential component of healthy human social interactions, stimulating prosocial and caregiving behaviours. It is acknowledged as fundamental in the development of moral behaviour, while absence of it may result in serious social and cognitive dysfunctions and has been associated with psychopathic personality.[41]

Drawing on the features of empathy and on research findings suggesting impairment in the affective but not in the cognitive aspect of empathy in psychopathic criminals, scholars have raised concerns on the risk of manufacturing 'psychopathic,' yet intelligent and cognitively empathic, AI machines.[42]

Cognitive empathy, entailing comprehending rather than feeling, is based on the perception of bodily expressions and behaviours of others and the subsequent process of inference. However, despite laying the ground for the notions of openness and other-directedness to build upon, it has been suggested that cognitive empathy can, in reality, be concurrent or even auxiliary to immorality. It is in this respect that this type of empathy does not work as a facilitator for moral agency. AI, limited to representing the situation of a hypothetical patient and applying a reliable algorithm

strates," *Frontiers in Neurology* 9, no. 491 (2018): 2.

[39] Asada, 45.

[40] Josanne D. M. van Dongen, "The Empathic Brain of Psychopaths: From Social Science to Neuroscience in Empathy," *Frontiers in Psychology* 11 (2020): 3.

[41] Ibid., 2.

[42] Carlos Montemayor, et al., "In Principle Obstacles for Empathic AI: Why we can't Replace Human Empathy in Healthcare," *AI & Society* (2021): 1.

for the rule of inference, is constrained to this type of empathy. On the contrary, affective empathy is more interrelated to moral agency as it is evocative of openness and other-directedness.[43]

In a study using resting-state fMRI, the researchers examined how differences between cognitive and affective empathy are reflected in the brain's intrinsic functional dynamics and found that affective empathy is associated with stronger functional connectivity among social–emotional regions (ventral anterior insula, orbitofrontal cortex, amygdala, perigenual anterior cingulate).[44]

In conclusion, what is concerning on the application of AI in the clinical setting is if and how empathy can be reproduced in the systems. Up to today, it seems that close relatives of empathy, like compassion and sympathy can be demonstrated by AI[45] 'agents' but the issue of affective empathy remains to be elucidated, if not yet accepted as impossible. Real human empathy is absolutely necessary in order to provide genuine healthcare in which a sense of connection is grown between the healthcare workers and patients.[46] Moral medical care cannot be dissociated from demonstrating empathy in response to human suffering.

V. Conclusion

The development of AI systems, especially those employing deep learning technologies is accompanied with several challenges. On the ethical domain, the issues of explainability and causation have raised hard debates on whether AI ought to be understandable or to follow counterfactual reasoning in order to be implemented in the clinical practice. As to today, achieving consensus on the meaning and implications of AI-related responsibility has proven difficult, while newly coined terms challenge traditional con-

[43] Elisa Aaltola, "Varieties of Empathy and Moral Agency," *Topoi* 33 (2014): 247.

[44] Christine L. Cox, et al., "The Balance Between Feeling and Knowing: Affective and Cognitive Empathy are Reflected in the Brain's Intrinsic Functional Dynamics," *Social Cognitive and Affective Neuroscience* 7, no. 6 (2012): 727.

[45] Montemayor, Halpern, and Fairweather, 3.

[46] Tan, et al., 8.

cepts in the domain of philosophy. At the same time, advances in neurocognitive research have revealed that empathy also includes affective, imaginative and relational dimensions, thus suggesting that a moral therapeutic relationship in medicine may not be reached via a machine, albeit intelligent.

References

Aaltola, Elisa. "Varieties of Empathy and Moral Agency." *Topoi* 33 (2014): 243-253.

Adadi, Amina, and Mohammed Berrada. "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)." *IEEE Access* 6 (2018): 52138-52160.

Alaloul, Wesam Salah, and Abdul Hannan Qureshi. "Data Processing Using Artificial Neural Networks." In *Dynamic Data Assimilation – Beating the Uncertainties*, edited by Dinesh G. Harkut. London: IntechOpen, 2020.

Asada, Minoru. "Development of Artificial Empathy." *Neuroscientific Research* 90 (2015): 41-50.

Bathaee, Yavar. "The Artificial Intelligence Black Box and the Failure of Intent and Causation." *Harvard Journal of Law and Technology* 31, no. 2 (2018): 889-938.

Behdadi, Dorna, and Christian Munthe. "Normative Approach to Artificial Moral Agency." *Minds & Machines* 30 (2020): 195-218.

Binns, Reubern. "Algorithmic Accountability and Public Reason." *Philosophy & Technology* 31 (2018): 543-556.

Bishop, J. Mark. "Artificial Intelligence Is Stupid and Causal Reasoning Will Not Fix It." *Frontiers in Psychology* 11, (2021): 1-18.

Bohr, Adam, and Kaveh Memarzadeh. "The Rise of Artificial Intelligence in Healthcare Applications." In *Artificial Intelligence in Healthcare*, 25-60. London: Academic Press, 2020.

Chou, Yu-Liang, Catarina Moreira, Peter Bruza, Chun Ouyang, and Joaquim Jorge. "Counterfactuals and Causability in Explainable Artificial Intelligence: Theory, Algorithms, and Applications." *Information Fusion*, pre-proof.

Coeckelbergh, Mark. "Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability." *Science and Engineering Ethics* 26, no. 4 (2020): 2051-2068.

Constantinescu, Mihaela, Cristina Voinea, Radu Uszkai, and Constantin Vica. "Understanding Responsibility in Responsible AI. Dianoetic Virtues and the Hard Problem of Context." *Ethics and Information Technology* (2021): 1-12.

Cox, L. Christine, Lucina Q. Uddin, Adriana Di Martino, Xavier F. Castellanos, Michael P. Milham, and Clare Kelly. "The Balance Between Feeling and Knowing: Affective and Cognitive Empathy are Reflected in the Brain's Intrinsic Functional Dynamics." *Social Cognitive and Affective Neuroscience* 7, no. 6 (2012): 727-737.

de Vignemont, Frederique, and Tania Singer. "The Empathic Brain: How, When and Why?" *Trends in Cognitive Sciences* 10, no. 10 (2006): 435-441.

Du, Mengnan, Ninghao Liu, and Xia Hu. "Techniques for Interpretable Machine Learning." *Communications of the ACM* 63, no. 1 (2019): 68-77.

Fogel, Alexander, and Joseph C. Kvedar. "Artificial Intelligence Powers Digital Medicine." *npj Digital Medicine* 1, no. 5 (2018): 1-4.

Healey, L. Meghan, and Murray Grossman. "Cognitive and Affective Perspective-Taking: Evidence for Shared and Dissociable Anatomical Substrates." *Frontiers in Neurology* 9, no. 491 (2018): 1-8.

Holger, Andreas, and Mario Guenther. "Regularity and Inferential Theories of Causation." *The Stanford Encyclopedia of Philosophy*. Fall 2021. Edited by Edward N. Zalta. https://plato.stanford.edu/archives/fall2021/entries/causation-regularity/.

Kelley, M. John, Gordon Kraft-Todd, Lidia Schapira, Joe Kossowsky, and Helen Riess. "The Influence of the Patient-Clinician Relationship on Healthcare Outcomes: A Systematic Review and Meta-Analysis of Randomized Controlled Trials." *PLOS ONE* 9, no. 4, e94207 (2014): 1-7.

Kerasidou, Angeliki. "Artificial Intelligence and the Ongoing Need for Empathy, Compassion and Trust in Healthcare." *Bulletin of the World Health Organization* 98, no. 4 (2020): 245-250.

Lake, Brenden M., Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. "Building Machines That Learn and Think Like People." *The Behavioral and Brain Sciences* 40, e253 (2017): 1-58.

Leeds, Stuart F., Kareem M. Atwa, Alexander M. Cook, Katharine A. Conway, and Timothy N. Crawford. "Teaching Heuristics and Mnemonics to Improve Generation of Differential Diagnoses." *Medical Education Online* 25, no. 1 (2020): 1-7.

Lockwood, Patricia L., Yuen-Siang Ang, Masud Husain, and Molly Crockett. "Individual Differences in Empathy Are Associated with Apathy-Motivation." *Scientific Reports* 7 (2017): 1-10.

Mallappallil, Mary, Jacob Sabu, Angelika Gruessner, and Moro Salifu. "A Review of Big Data and Medical Research." *SAGE Open Medicine* 8 (2020): 1-10.

Matthias, Andreas. "The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata." *Ethics and Information Technology* 6 (2004): 175-183.

McCarthy, John, Marvin L. Minsky, Nathaniel Rochester, and Claude E. Shannon. "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955." *AI Magazine* 27, no. 4 (2006): 12-13.

Montemayor, Carlos, Jodi Halpern, and Abrol Fairweather. "In Principle Obstacles for Empathic AI: Why We Can't Replace Human Empathy in Healthcare." *AI & Society* (2021): 1-7.

Nagy, Matthew, and Bryan Sisk. "How Will Artificial Intelligence Affect Patient-Clinician Relationships?" *American Medical Association Journal of Ethics* 22, no. 5 (2020): E395-400.

Obermeyer, Ziad, and Ezekiel J. Emanuel. "Predicting the Future – Big Data, Machine Learning, and Clinical Medicine." *The New England Journal of Medicine* 375, no. 13 (2016): 1216-1219.

Parascandola, Mark. "The Epidemiologic Transition and Changing Concepts of Causation and Causal Inference." *Revue d' Histoire des Sciences* 2, no. 64 (2011): 243-262.

Raita, Yoshihiko, Carlos A. Camargo Jr., Liming Liang, and Kohei Hasegawa. "Big Data, Data Science, and Causal Inference: A Primer for Clinicians." *Frontiers in Medicine* 8 (2021): 1-13.

Richens, Jonathan G., Ciarán M. Lee, and Saurahb Johri. "Improving the Accuracy of Medical Diagnosis with Causal Machine Learning." *Nature Communications* 11 (2020): 1-9.

Schwab, Klaus. "The Fourth Industrial Revolution: What it Means, How to Respond." *World Economic Forum*, January 14, 2016. https://www.weforum.org/agenda/2016/01/the-fourth-industrial-revolution-what-it-means-and-how-to-respond/.

Stueber, Karsten. "Empathy." *The Stanford Encyclopedia of Philosophy*. Fall 2019. Edited by Edward N. Zalta. https://plato.stanford.edu/archives/fall2019/entries/empathy/.

Tan, Laurence, Mai Khanh Le, Chou Chuen Yu, Sok Ying Liaw, Tanya Tierney, Yun Ying Ho, Evelyn Lim, Daphne Lim, Reuben Ng, Colin Ngeow, and James Low. "Defining Clinical Empathy: A Grounded Theory Approach from the Perspective of Healthcare Workers and Patients in a Multicultural Setting." *British Medical Journal Open* 11, no. 9, e045224 (2021): 1-9.

van Dongen, Josanne D. M. "The Empathic Brain of Psychopaths: From Social Science to Neuroscience in Empathy." *Frontiers in Psychology* 11 (2020): 1-12.

Vasudevan, K. Rama, Maxim Ziatdinov, Lukas Vlcek, and Sergei V. Kalinin. "Off-the-shelf Deep Learning Is not Enough, and Requires Parsimony, Bayesianity, and Causality." *npj Computational Materials* 7, no. 16 (2021): 1-6.