This is the author's version of an article published in the Canadian Medical Association Journal. The final authenticated version is available online at: https://doi.org/10.1503/cmaj.190506

# Deconstructing the diagnostic reasoning of human versus artificial intelligence

Thierry Pelaccia[1], Germain Forestier[2], Cédric Wemmert[3]

Artificial intelligence (AI) is often presented as the future of medical practice. The concept of AI was developed in the 1950s and has been defined as "the use of a computer to model intelligent behaviour with minimal human intervention." [1] It is an alternative to human intelligence, particularly as a replacement for the diagnostic skill of physicians. For several years, the scientific literature and lay media have commented that nonhuman intelligence could equal or even exceed human intelligence in diagnostic tasks [2]. Human intelligence is evident in the concept of clinical reasoning [3], which has been defined as "the internal mental processes that a physician uses when approaching clinical situations" [4]. This central component of physicians' competence, once honed, allows them to make diagnoses [3]. In medicine, clinical reasoning is often understood from the perspective of cognitive psychology's information process theory [4]. Artificial intelligence may refer to several different methods. Most AI diagnostics are based on machine learning algorithms that are "intelligent" enough to handle difficult and complex problems; algorithms rely on human intelligence for their creation [5]. Recently, substantial progress has been made in this field through the resurgence of neural networks — a family of methods of machine learning — and particularly deep neural networks [6]. Herein, we focus mainly on machine learning (specifically deep neural networks). We analyze the differences in the ways humans and AI approach diagnostic reasoning to argue that human reasoning will not become obsolete in medical diagnosis.

## How do humans and AI perform diagnostic tasks and learn to make diagnoses?

Both humans and AI learn through repeated exposure to clinical cases, referred to as "experiences" for human intelligence and "examples" for AI. For both to develop, feedback, based on the intervention of an expert, is important. A physician solves most clinical problems in an intuitive and deductive way, whereas AI problem-solving depends on access to and analytical and deductive processing of large quantities of data that relate to the case.

### Deductive versus inductive; intuitive versus analytical

To learn to make diagnoses, medical students must organize their experiences of many clinical cases in long-term memory [4]. However, in addition to broad-ranging experience, the development of expertise requires understanding of context and the way in which disease is presented in that context; this is crucial to being able to solve new cases through a generalization process [7]. Immediate, appropriate feedback on decision-making consolidates knowledge and enables future clinical reasoning [7].

Physicians mainly use a hypothetico-deductive approach to make diagnoses [8]. After generating diagnostic hypotheses early, they spend most of their diagnostic time testing them by collecting more data. This approach is underpinned by cognitive processes that, according to the dual-process theory, can be either intuitive or analytical [7]. Intuition — sometimes referred to as "pattern recognition" — is a process that works automatically and subconsciously [7, 9]. It allows humans to generate diagnostic hypotheses early by taking a few pieces of information, associating them and comparing the result with patterns stored in long-term memory [7]. These patterns are built through academic and clinical learning experiences, particularly repeated confrontation with similar situations [8]. Intuition allows humans to consider only a few solutions — the most likely in the context — among all those that could be considered given the available data. This approach is essential given the limited capacity of the human brain to process information. Most researchers agree that intuitive processes are the main source of generation of diagnostic hypotheses for humans [10].

Machine learning, however, depends on the development of an algorithm that "learns" important features from a data set known as a "training set" to then make predictions about other unknown data [11]. For the learning to occur, data used for training must be labelled according to their association with the solution; these data are referred to as the "ground truth." For example, a patient's physiologic data must be associated with a label indicating whether the patient is sick or healthy. The ground truth is provided by a human expert (most often a physician), either directly (e.g., image annotations) or through documents (e.g., clinical reports). Thus, unlike humans, who know thousands of small pieces of information (often referred to as "common sense"), AI is limited to the specific information provided for a specific task. Furthermore, for every new task, AI systems must usually start from scratch.

Artificial intelligence systems are composed of a model (representing the learned knowledge), a decision function (making it possible to answer to the problem when a new input is given) and an evaluation metric (to evaluate the quality of

[1]Centre for Training and Research in Health Sciences Education, Faculty of Medicine, University of Strasbourg; Hôpitaux universitaires de Strasbourg, Strasbourg, France
[2]Institute of Research in Computer Science, Mathematics, Automation and Signal (Forestier), Université de Haute-Alsace, Mulhouse, France
[3]The Engineering Science, Computer Science and Imaging Laboratory, University of Strasbourg, Illkirch, France

the answer provided by AI compared with the ground truth). In AI, acquired knowledge can be stored in different ways. Deep neural networks are composed of layers of interconnected artificial neurons forming a "model." The architecture of the network and the weights associated with each connection represent a "decision function." From an input (e.g., a histopathological image), the neural network provides a prediction as an output (e.g., cancer or not cancer). To learn, the algorithm automatically optimizes its solution by calculating an evaluation metric function, which is basically the difference between the output proposed by the algorithm and the ground truth. In deep neural networks, the error computed by the evaluation metric is back-propagated through the layers of the network, and the algorithm modifies the weights of the connections between the neurons. The process is iterated until the algorithm proposes accurate outputs on the training set. Problem solving by AI is thus different from the hypothetico-deductive approach used by humans. Intuitive reasoning is difficult to model or simulate as it is based on experience that bypasses a conscious "orderly sequential analysis" of a situation, which is the core of an algorithm. Therefore, AI uses an analytical approach in an inductive mode (*i.e.*, it systematically moves from data toward the solution) [12]. Although humans understand cause-and-effect relations, these are not yet modelled in AI. This subject has been studied for a long time in AI, but it is only recently that first attempts to define an AI that "thinks like a human" have been proposed [13].

**Data**

Physicians need very few data (i.e., 2 to 4 pieces of contextual or clinical information) to generate diagnostic hypotheses through intuition [7, 14]. Subsequently, and to verify the hypotheses generated, additional data guided by the hypotheses are collected through the interview, clinical examination and additional tests. Human intelligence will transform data collected during the patient interview into something that can be processed through "semantic transformation" [15]. For example, clinicians might transform "the first time" into "inaugural," or "several episodes" into "iterative." Most AI systems do not model intuition and therefore require substantial data to make a relevant diagnosis [12]. This is why AI is presently most effective in situations where all the data of the problem to be solved are immediately accessible, such as in medical imaging. Artificial intelligence also requires data transformation, but in AI this a much more complex and time-consuming process. Through data integration or data preprocessing, the data must be transformed to be computational, which means that all information needs to be digitized and categorized to be interpreted by the machine. This is one of AI's great challenges [16].

**How do humans and AI misdiagnose?**

The rate of diagnostic errors in medical practice is estimated at about 5%–15%, depending on the specialty [17]. This translates into more than 12 million misdiagnoses annually in the United States alone [18]. Cognitive biases are considered to be the cause of most diagnostic errors [19] and many biases have been reported in the medical scientific literature [8]. Premature closure bias (i.e., the tendency to stop considering other hypotheses after reaching a diagnosis) is considered to be the most common [20]. Three other common biases are anchoring bias (the tendency to focus early on 1 or more salient features of the initial presentation of the problem and failure to change this first impression in the light of data gathered later), availability bias (the tendency to consider diagnoses that are easy to remember, often because they have recently been made, as more likely) and confirmation bias (the tendency to consider only confirmatory data in relation to the generated hypothesis, while ignoring or underestimating contradictory data) [8].

In most instances, the error rate for AI can be calculated accurately by comparing the results provided by the AI model to expected results (considered to be the truth) [21]. Errors in AI are not comparable to human errors as they mostly result from problems that arise during the learning step, usually poor training data quality or an irrelevant evaluation metric [22]. Having a data set that expresses the entire variety of the data and the real associations between them, and that does not contain misclassified examples and does not present any bias that could lead the AI to learn false assumptions, is essential. Other sources of errors, imprecisions or uncertainty could include the use of an inappropriate model (e.g., unable to represent the knowledge to learn) or poor experimental design (e.g., stopping learning too early).

**What evidence supports the role of AI in medical diagnosis?**

Artificial intelligence was shown to be capable of classifying skin cancers with a level of performance comparable to that of dermatologists when it was trained using a data set of nearly 130 000 images and then tested on its ability to distinguish between 2 common cancers and between a benign and a malignant lesion [2]. Artificial intelligence was able to detect diabetic retinopathy just as well as 8 ophthalmologists, while providing more consistent interpretation, high sensitivity and specificity, and an instantaneous result, following training using a data set of nearly 130 000 retinal images and validation using 2 further data sets [23]. In an evaluation of more than 30 deep-learning algorithms, 7 diagnostic algorithms were shown to be better than 11 histopathologists at diagnosing breast cancer metastases to lymph nodes in images of tissue sections when human specialists and AI were similarly time constrained [24]. An AI algorithm trained on a data set of more than 100 000 images was better than specialist radiologists at detecting pneumonia using chest radiographs [25]. A machine-learning framework was trained to perform better than emergency medical dispatchers in recognizing cardiac arrest in emergency phone calls [26].

**What are the criticisms of AI in medical diagnosis?**

Many studies conducted in the field of medical AI have been criticized for lack of scientific rigour, an unsatisfactory evaluation process or insufficient information reported in the methods [27]. Moreover, the scientific literature skews toward publishing successful projects, whereas failures are rarely reported

on blogs or consumer articles, if they are reported at all. These concerns undermine trust in AI.

A recent article [28] described 4 essential characteristics for trusting AI systems: fairness (training data and models must be free of bias to avoid unfair treatment of certain groups of patients), robustness (AI systems should be safe and secure), explainability (decisions provided by AI must be understandable by their users) and transparency (AI systems should include details of their development, deployment and maintenance). Explainability is perhaps the most challenging issue to solve. Although it is usually possible to explain physicians' reasoning and the origin of their decisions, many of the most powerful AI methods (e.g., deep neural networks) are often criticized for being a "black box" [29]. Currently, machine learning on medical data most often takes the form of retrospective analysis of large routinely collected data sets with careful scrutiny of the results proposed by the AI.

An active and fast-growing field of AI seeks to make AI decisions explainable and understandable by users, with any preliminary research studies being conducted to reach this goal [30–32]. Another challenge is to propose robust machine-learning methods [33]. Meta-learning [34] and transfer learning [35] are 2 promising avenues of research to help AI "remember" something and to learn "how to learn."

### Future directions

Several studies have shown the extent to which AI can be used to make and support diagnosis in medicine. Since current evidence supports the effectiveness of AI for only a small selection of diagnostic tasks and human experts remain able to learn and diagnose a wide array of conditions, human intelligence would seem to remain essential to diagnosis for now. However, the consistency with which AI can be trained to perform diagnoses when exposed to similar data independent of context — with errors fixable by improving the quality of data supplied for learning — supports the continued development of AI diagnostics. Physicians'reasoning has been shown to be sensitive to factors such as fatigue, sleep deprivation, interruptions, cognitive overload, noise or psycho-emotional status [10], and to be influenced by cognitive biases [17] with human error impossible to eliminate entirely and even difficult to reduce substantially [8]. AI is becoming, and will continue to develop to be, a useful tool to mitigate human error and improve quality in medical practice. Yet the idea that AI is able to learn on its own and will replace physicians is a myth that needs to be deconstructed [36, 37]. The potential of AI in medicine can be realized only if it is designed by the collaborative human intelligence of a physician and a data scientist [38].

Because human and artificial intelligences are different and complementary, it is unlikely that AI will entirely replace the physician in the resolution of clinical problems. Artificial intelligence will be among the tools available to physicians seeking to make a diagnosis, to help with reasoning, reduce diagnostic uncertainty and augment shared decision-making, which also involves other health professionals and the patient. Diagnostic uncertainty is common in medical practice [39]. Artificial intelligence can enable physicians to favour one diagnostic hypothesis over another or to generate hypotheses that they had not previously considered.

The tasks facing stakeholders in the development of AI, among whom physicians will play a central and essential role, will be improving the quality and accessibility of medical data that can be used as a source of learning for AI while carefully respecting ethical considerations; being able to explain the results produced by AI to human intelligence; overcoming physicians'resistance related to fears of being downgraded when certain diagnostic tasks no longer rely solely on their intelligence; and training medical students early on in the integration of AI tools into their diagnostic practice, which implies extracting themselves from a historical and firmly rooted posture of the physician-centred diagnostic process [40]. Under these conditions, AI can assume its place as a routine tool in medical practice.

### Acknowledgment

This article is partly based on a lecture given by the first author at the congress of the French National Society of Internal Medicine on June 6, 2019, and published in the congress proceedings.

## References

[1] Hamet, P. and Tremblay, J. "Artificial intelligence in medicine". In: *Metabolism* 69 (2017), S36–S40.

[2] Esteva, A., Kuprel, B., Novoa, R. A., et al. "Dermatologist-level classification of skin cancer with deep neural networks". In: *Nature* 542.7639 (2017), p. 115.

[3] Norman, G. "Research in clinical reasoning: past history and current trends". In: *Medical education* 39.4 (2005), pp. 418–427.

[4] Durning, S. J., Artino Jr, A. R., Schuwirth, L., et al. "Clarifying assumptions to enhance our understanding and assessment of clinical reasoning". In: *Academic Medicine* 88.4 (2013), pp. 442–448.

[5] Foster, K. R., Koprowski, R., and Skufca, J. D. "Machine learning, medical diagnosis, and biomedical engineering research-commentary". In: *Biomedical engineering online* 13.1 (2014), p. 94.

[6] LeCun, Y., Bengio, Y., and Hinton, G. "Deep learning". In: *nature* 521.7553 (2015), pp. 436–444.

[7] Pelaccia, T., Tardif, J., Triby, E., et al. "An analysis of clinical reasoning through a recent and comprehensive approach: the dual-process theory". In: *Medical education online* 16.1 (2011), p. 5890.

[8] Norman, G. R., Monteiro, S. D., Sherbino, J., et al. "The causes of errors in clinical reasoning: cognitive biases, knowledge deficits, and dual process thinking". In: *Academic Medicine* 92.1 (2017), pp. 23–30.

[9] Norman, G. R. and Eva, K. W. "Diagnostic error and clinical reasoning". In: *Medical education* 44.1 (2010), pp. 94–100.

[10] Croskerry, P. "A universal model of diagnostic reasoning". In: *Academic medicine* 84.8 (2009), pp. 1022–1028.

[11] Nasrabadi, N. M. "Pattern recognition and machine learning". In: *Journal of electronic imaging* 16.4 (2007), p. 049901.

[12] Obermeyer, Z. and Emanuel, E. J. “Predicting the future—big data, machine learning, and clinical medicine”. In: *The New England journal of medicine* 375.13 (2016), p. 1216.

[13] Lake, B. M., Ullman, T. D., Tenenbaum, J. B., et al. “Building machines that learn and think like people”. In: *Behavioral and brain sciences* 40 (2017).

[14] Pelaccia, T., Tardif, J., Triby, E., et al. “Insights into emergency physicians’ minds in the seconds before and into a patient encounter”. In: *Internal and emergency medicine* 10.7 (2015), pp. 865–873.

[15] Bordage, G. and Lemieux, M. “Which medical textbook to read? Emphasizing semantic structures”. In: *Academic Medicine* 65.9 (1990), S23–4.

[16] Smeulders, A. W., Worring, M., Santini, S., et al. “Content-based image retrieval at the end of the early years”. In: *IEEE Transactions on Pattern Analysis & Machine Intelligence* 12 (2000), pp. 1349–1380.

[17] Kuhn, G. J. “Diagnostic errors”. In: *Academic Emergency Medicine* 9.7 (2002), pp. 740–750.

[18] Singh, H., Meyer, A. N., and Thomas, E. J. “The frequency of diagnostic errors in outpatient care: estimations from three large observational studies involving US adult populations”. In: *BMJ Qual Saf* 23.9 (2014), pp. 727–731.

[19] Lambe, K. A., O’reilly, G., Kelly, B. D., et al. “Dual-process cognitive interventions to enhance diagnostic reasoning: a systematic review”. In: *BMJ Qual Saf* 25.10 (2016), pp. 808–820.

[20] Graber, M. L., Franklin, N., and Gordon, R. “Diagnostic error in internal medicine”. In: *Archives of internal medicine* 165.13 (2005), pp. 1493–1499.

[21] Hripcsak, G. and Rothschild, A. S. “Agreement, the f-measure, and reliability in information retrieval”. In: *Journal of the American Medical Informatics Association* 12.3 (2005), pp. 296–298.

[22] Flach, P. “Performance Evaluation in Machine Learning: The Good, The Bad, The Ugly and The Way Forward”. In: *33rd AAAI Conference on Artificial Intelligence*. 2019.

[23] Gulshan, V., Peng, L., Coram, M., et al. “Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs”. In: *Jama* 316.22 (2016), pp. 2402–2410.

[24] Bejnordi, B. E., Veta, M., Van Diest, P. J., et al. “Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer”. In: *Jama* 318.22 (2017), pp. 2199–2210.

[25] Rajpurkar, P., Irvin, J., Zhu, K., et al. “Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning”. In: *arXiv preprint arXiv:1711.05225* (2017).

[26] Blomberg, S. N., Folke, F., Ersbøll, A. K., et al. “Machine learning as a supportive tool to recognize cardiac arrest in emergency calls”. In: *Resuscitation* 138 (2019), pp. 322–329.

[27] “AI diagnostics need attention”. In: *Nature* 555:285 (2018).

[28] Arnold, M, Bellamy, R., Hind, M, et al. “FactSheets: Increasing trust in AI services through supplier’s declarations of conformity”. In: *IBM Journal of Research and Development* 63.4/5 (2019), pp. 6–1.

[29] Wainberg, M., Merico, D., Delong, A., et al. “Deep learning in biomedicine”. In: *Nature biotechnology* 36.9 (2018), p. 829.

[30] Ribeiro, M. T., Singh, S., and Guestrin, C. “Why should i trust you?: Explaining the predictions of any classifier”. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM. 2016, pp. 1135–1144.

[31] Simonyan, K., Vedaldi, A., and Zisserman, A. “Deep inside convolutional networks: Visualising image classification models and saliency maps”. In: *arXiv preprint arXiv:1312.6034* (2013).

[32] Seo, J. D. “Visualizing Uncertainty and Saliency Maps of Deep Convolutional Neural Networks for Medical Imaging Applications”. In: *arXiv preprint arXiv:1907.02940* (2019).

[33] Feurer, M., Klein, A., Eggensperger, K., et al. “Efficient and robust automated machine learning”. In: *Advances in neural information processing systems*. 2015, pp. 2962–2970.

[34] Brazdil, P., Carrier, C. G., Soares, C., et al. *Metalearning: Applications to data mining*. Springer Science & Business Media, 2008.

[35] Pan, S. J. and Yang, Q. “A survey on transfer learning”. In: *IEEE Transactions on knowledge and data engineering* 22.10 (2009), pp. 1345–1359.

[36] Savadjiev, P., Chong, J., Dohan, A., et al. “Demystification of AI-driven medical image interpretation: past, present and future”. In: *European radiology* 29.3 (2019), pp. 1616–1624.

[37] Miller, D. D. and Brown, E. W. “Artificial intelligence in medical practice: the question to the answer?” In: *The American journal of medicine* 131.2 (2018), pp. 129–133.

[38] Pelaccia, T., Forestier, G., and Wemmert, C. “Une intelligence artificielle raisonne-t-elle de la même façon que les cliniciens pour poser des diagnostics ?” In: *Revue de Médecine Interne (La)* 40 (2019), A16–A19.

[39] Hatch, S. “Uncertainty in medicine”. In: *British Medical Journal Publishing Group* 357:j2180 (2017).

[40] Mandl, K. D. and Bourgeois, F. T. “The evolution of patient diagnosis: from art to digital data-driven science”. In: *Jama* 318.19 (2017), pp. 1859–1860.