A Virtue Ethics Analysis of the Development of IBM Watson in Oncology

STS Research Paper
Presented to the Faculty of the
School of Engineering and Applied Science
University of Virginia

By

Nathan Patton

Spring, 2024

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

ADVISOR

Benjamin J. Laugelli, Assistant Professor, Department of Engineering and Society

Introduction

The IBM AI supercomputer, Watson, hailed as a breakthrough in cancer treatment recommendations, promised to sift through extensive data sets encompassing patient histories and medical protocols. Starting back in 2007, IBM set out to create a computer system capable of taking on human champions in Jeopardy. Just four years later, it achieved this feat against the all-time winner, Ken Jennings. Yet, Watson's journey didn't stop there. It began delving into the complexities of the healthcare industry (“IBM Watson”).

Introducing Watson into healthcare promised oncologists instant access to the latest research and insights. Initially trained by oncologists at the Memorial Sloan-Kettering Cancer Center (MSKCC) in New York City in optimizing treatments for lung cancer, the aim was to extend Watson's knowledge to encompass other cancer types and diseases. Chris Hickey, a spokesperson for MSKCC, believed that the real benefit of using Watson throughout the healthcare system is to even expertise so that oncologists who see a handful of patients a year can have the same level as those who see thousands (Miller, 2013, p. 1). With its natural language processing abilities, Watson could understand human queries effortlessly. For instance, a simple statement like "Watson, I have a headache and my eyes hurt" could trigger a wealth of information and treatment suggestions drawn from extensive literature and data.

Unfortunately, doubts surfaced regarding the reliability and precision of IBM Watson and its oncology treatment suggestions. One of the major limitations of a system of this type is not what it can do for the 99.9% of people, but what it can fail to do for the rest (Miller, 2013, p. 2). While some scholars have dissected the saga, scrutinizing the use of IBM Watson in healthcare, little attention has been paid to the ethical dimensions of actions carried out by IBM. The ethical discourse surrounding the implementation of Watson into oncology remains underexplored,

notably the absence of vital characteristics expected from tech company leadership. Delving into these ethical considerations offers a lens through which to assess the moral stance of the organization.

Applying a virtue ethics framework, which emphasizes cultivating moral attributes, I aim to support this argument. Specifically, I will provide insights into the root causes behind the shortcomings of IBM Watson by demonstrating that their actions were morally unacceptable due to their lack of three categories of character traits: informative communication and commitment to quality, openness to correction and perseverance, and competence. These character shortcomings become apparent through repeated actions and poor decision making during the development of Watson.

Literature Review

A significant amount of research already exists analyzing the uses of IBM Watson For Oncology (WFO). These analyses typically underscore the significance of understanding and addressing specific contextual factors and potential biases associated with the use of AI systems like WFO in oncology practice. They fail to make any significant judgment regarding the morality of the IBM individuals responsible for developing this technological device.

In the *Concordance Study Between IBM Watson for Oncology and Clinical Practice for Patients with Cancer in China,* the author concluded that varying levels of concordance were exhibited for different cancer types. Concordance was analyzed by comparing the treatment decisions proposed by WFO with those of the multidisciplinary tumor board, and it was divided into four categories: recommended, for consideration, not recommended, and Physician's choice. For concordance to be achieved, it must fall within the "recommended" or "for consideration" categories. Lung cancer, breast cancer, and ovarian cancer reached a concordance exceeding

80%, while rectal cancer surpassed 60%, with gastric cancer behind at only 12%. The study later suggests that it is necessary to accelerate the localization of WFO before concordance can be rapidly adapted in other countries, such as China, due to differences in factors such as incidence rates and pharmaceutical availability (Zhou et al., 2019, p. 817-819). While the authors do discuss the reliability of the concordance metric and the localization of WFO, there is no judgment of the morality of IBM.

Similarly, the second article, *Concordance as evidence in the Watson for Oncology decision-support system*, raises concerns about WFO being developed based on US treatment guidelines from MSKCC and thus does not consider differences in cancer incidence or risk across different populations. It then goes even further to mention that concordance alone may not be a reliable metric for validating treatment options. They argue that using it as a form of evidence of quality or trustworthiness is problematic. "Concordance says little about the outcomes of chosen treatment protocols, which would be of greater value for determining which options to choose (Tupasela & Di Nucci, 2020, p. 816)." Again, this article focuses on the reliability of a metric used for validating treatment options, without touching on any moral judgment of IBM.

Further analysis is justified to assess the reliability metrics for validating treatment options. However, it is equally important to consider the ethical implications of the actions carried out by IBM and how it plays a role in the reliability of their product. Delving into the immorality of these steps will shed light on broader considerations in healthcare technology.

Conceptual Framework

To evaluate the moral conduct of the IBM Watson developers, I will employ a virtue ethics framework. Virtue ethics is currently one of the three major approaches in normative

ethics, and it emphasizes moral character, rather than duties or rules (deontology) or consequences of actions (consequentialism). Unlike Utilitarianism and Kantian theory, which prioritize universal principles, virtue ethics delves into the character of the individual, highlighting the traits deemed morally commendable. This approach integrates ethics and psychology, emphasizing the cultivation of virtuous attributes (van de Poel & Royakkers, 2011, p. 95).

Originating in ancient Greece with philosophers like Socrates, Plato, and Aristotle, virtue ethics centers on achieving "the good life" or embodying goodness as a person, while not referring to a happy circumstance that brings pleasure. Aristotle, one of its early proponents, laid the groundwork for virtue ethics, emphasizing the importance of character development. He believed in practical wisdom, or that a wise man can see what he has to do in the specific and often complex circumstances in life. In other words, people make the right choice for action concerning what is good and useful for a successful life. Virtue is mean or middle between two vices, which depends on two extremes: excess and deficit. An example of this is courage, which is the balance between cowardice and recklessness (van de Poel & Royakkers, 2011, p. 96-98).

Some virtues are quite generic, such as integrity, honesty, courage, and self-sacrifice, which apply not only to most professions but to non-professional life as well. In this case, since I am interested in the virtues associated with the development of IBM Watson, engineering-specific virtues might be more applicable. Such virtues include but are not limited to the following: competence, ability to communicate clearly and informatively, perseverance, willingness to compromise, openness to correction (admitting mistakes, acknowledging oversight), and commitment to quality (Pritchard, 2001). Pritchard notes that lacking these virtues “detracts from responsible engineering practice in general, and exemplary practice in

particular." He then goes on to mention how having these dispositions is a fundamental part of morally commendable engineers.

Thus, in examining the actions of stakeholders involved in creating Watson, I will scrutinize instances where key virtues were lacking. This includes deficiencies in informative communication and commitment to quality, openness to correction and perseverance, and competence. Such an analysis aims to shed light on the ethical dimensions of the project and the moral responsibilities of those involved.

Analysis

As previously noted, the use of concordance as a metric for assessing the reliability of IBM Watson has been extensively studied. IBM and its clinical partners have published multiple studies on their website demonstrating Watson's high level of "concordance" with the treatment recommendations of oncologists. However, internal documents presented by Dr. Andrew Norden, an oncologist and deputy health chief, said otherwise (Ross & Swetlitz, 2018, p. 3). An analysis of the actions taken by IBM executives and partners within these documents reveals a lack of many important virtues, including informative communication and commitment to quality, openness to correction and perseverance, and competence. The following sections focus on each of these categories of virtues and how IBM is noticeably lacking them based on their actions. These deficiencies raise concerns for the development and deployment of their product, and addressing them is necessary for IBM to regain trust, reliability, credibility, and confidence with their customers within the health sector.

<u>Informative Communication and Commitment to Quality</u>

IBM struggled with clear, informative communication and commitment to quality, which proved to be a significant issue during the implementation into oncology. Communication

involves the unbiased and professional exchange of information to promote understanding, encourage action, and stimulate thought or ideas (Danielson, 2020). Quality encompasses various facets but, in this context, refers to a degree of excellence or superiority. It's about having distinguishing attributes that set a product or service apart ("Definition of Quality," 2024). These are two complementary processes that help build trust with customers. In this regard, IBM failed to represent these virtues, which deviates from the norm and raises questions about their morality.

Executives made several misleading statements, presented conflicting information, and responded poorly to customer dissatisfaction. In 2017, the senior vice president for the IBM cognitive solutions division, which includes Watson, made statements at an IBM event that implied that Watson for Oncology is "going fabulously" and that its recommendations were based on data from real patients data. He said that Watson has "ingested all of the MSKCC data, historic patients, and results." IBM CEO, Ginni Rometty, commented "You must be transparent about it because it matters in these decisions." General manager of IBM Watson, Deborah DiSanzo, stated in an interview that physicians endorse it; some have even remarked that if it were removed, they would consider revolting. However, there were discrepancies between the information that was presented by IBM marketing to the public and their internal documents (Ross & Swetlitz, 2018, p. 2).

Contradicting the success presented by IBM Watson executives, Martin Kohn, who came to IBM with a medical degree from Harvard and an engineering degree from MIT, was the chief medical scientist for IBM research who once said that the company fell into a common trap: "Merely proving that you have powerful technology is not sufficient. Prove to me that it will actually do something useful - that it will make my life better, and my patients' lives better."

Kohn mentioned how he has been waiting to see peer-reviewed papers in medical journals demonstrating that AI can improve patient outcomes and save health systems money. “To date, there’s very little in the way of such publications, and none of the consequences for Watson,” he says. Before this, however, Kohn believed that Watson had the potential to overcome the complexities of tackling the language of medicine (Strickland, 2019). Based on these quotes, it seemed that IBM promised something that they could not deliver through bad press. They were too ambitious with their product, and thus disappointed clients and created skepticism around the quality of their other products (Mearian, 2018).

Additionally, several key points highlight the lack of commitment to quality: transparency and communication, training data issues, inadequate testing and validation, and customer feedback and concerns. Using synthetic cases rather than real patient data raised questions about the validity of the training data. This approach may have ultimately compromised the quality of the device and undermined the effectiveness of providing accurate treatment guidelines. Additionally, IBM had a lack of transparency regarding these training methods, which not only speaks to their ability to communicate clearly and informatively but also their quality. It demonstrates their failure to uphold quality standards in communication with the public and healthcare providers. On top of all of this, customers, including doctors at hospitals who were promoting the product, indicated that they were dissatisfied with the system's performance. It was reported as being error-prone and unsafe. “A machine learning tool is only as good as the data that goes into it (Konam, 2022).” This is suggesting that a machine learning tool cannot be quality if the data is not quality.

As stated previously, IBM lacked a commitment to quality due to various reasons, one of them being training data issues. However, some might argue that hand-crafted or synthetic data is

not necessarily bad, and therefore there is not enough evidence to say that their commitment to quality is lacking. There is no clear consensus regarding a unified definition of synthetic data, leading to inconsistent use of the term and interpretations that vary across contexts, which affects the reproducibility and transparency of research (Giuffrè & Shung, 2023, p. 1). However, the closest working definition is "data that has been generated using a purpose-built mathematical model or algorithm, with the aim of solving a (set of) data science tasks(s) (Jordan et al., 2022, p. 5)." Jordan et al. convey how synthetic data of this type is a technology with significant promise and has the potential to accelerate development. They go on to list negatives associated with synthetic data, such as it not being automatically private, nor being a replacement for real data. In this case, Watson did generate data using a mathematical algorithm, but they did not account for the fact that this data reflected the doctors' own biases and blind spots. Therefore, this data was not necessarily generalizable to all patient cases. The internal documents revealed that IBM conducted studies on the software to demonstrate its usefulness. However, these studies were designed to generate favorable results, rather than rigorously testing the system (Ross & Swetlitz, 2018, p. 3).

IBM failed to prioritize ethical conduct, which was apparent in their lack of commitment to quality and transparency in communication with customers. They neglected in maintaining high-quality, real-world data, which led to issues with the reliability and effectiveness of the product. This disregard for virtues undermined the trustworthiness of the IBM Watson and contributed to its shortcomings.

Openness to Correction and Perseverance

The second group of virtues IBM failed to represent was being open to correction and perseverant. Similarly to being open to correction, what does it mean to be open-minded? " To be

open-minded is… to be critically receptive to alternative possibilities, to be willing to think again despite having formulated a view, and to be concerned to defuse any factors that constrain one's thinking in predetermined ways (Hare, 2003, p 4-5).” To extend on that, William Hare, the leading proponent of the view, contends, “the test of open-mindedness is… whether or not we are prepared to entertain doubts about our views (Hare, 1987, p 99).” Perseverance is described as “to persist in a state, enterprise, or undertaking in spite of counter influences, opposition, and discouragement (“Perseverance Definition & Meaning,” 2024).” These are both concepts that are crucial for the company's drive towards innovation and excellence, making it easier to compete in the rapidly growing market. They failed to persist in such a promising field of work, once faced with challenges.

Thus, reason number two for why IBM acted immorally in the development of their product is because they were not receptive to alternative possibilities or prepared to entertain doubts about their views of the product from their customer. IBM received several forms of feedback from customers regarding the product's inadequacies and limitations, but despite this, they continued to not meet their needs. Similarly, the internal documents revealed that they were slow to adapt to new research findings and the changing treatment guidelines. Despite the internal documents revealing that unsafe and incorrect treatment recommendations were given by Watson, IBM continued to promote that the product was highly effective and based on real patient data (Ross & Swetlitz, 2018, p. 3).

On top of this, they failed to deliver on their promises about its potential to revolutionize healthcare through its advanced analytics, decision support, and personalized medicine. There were no tangible results that were sufficient enough for healthcare providers and their patients. The main indicator of their lack of perseverance was the discontinuation of the project in its

entirety back in 2020. Collaborators with IBM Watson, such as Herbert Chase, a professor at Columbia University who collaborated with health care efforts, believed that Watson could keep up - and if turned into a tool for "clinical decision support,"it could enable doctors to keep up too. Eliot Siegel, a professor of radiology at The University of Maryland, also collaborated with IBM on diagnostic research. He believed that AI-enabled tools will be "indispensable" to doctors within a decade (Strickland, 2019). Even with such positive outlooks on the potential of this product, the business was sold for an undisclosed price to Francisco Partners, a private investment firm. More than $4 billion was spent just to acquire companies with medical data, yet IBM was looking to sell the company business for about $1 billion. The senior vice president in charge of software business at IBM, Tom Rosamilia, said the move was necessary for IBM to narrow down its focus to a platform-based hybrid cloud and AI strategy (Lohr, 2021) All in all, IBM lacked a strong long-term vision. They were not committed to long-term goals, learning, and adaptation for a more successful product.

The absence of perseverance and acknowledgment for flaws within the IBM culture demonstrated immoral behavior, preventing them from overcoming obstacles. This culture prevented the company from making meaningful improvements to their product, which limits their potential for success in the future. Without a commitment to improvement and openness to correction, IBM was unable to adapt to the evolving healthcare marketplace.

Competence

Putting the two reasons together, IBM lacked competence. Competence, as defined by the Merriam-Webster dictionary, entails having sufficient knowledge, judgment, skill, or strength for a particular duty or in a specific context ("Definition of Competence," 2024). A lung cancer specialist at MSKCC, Mark Kris, said it best, "I don't think anybody had any idea it would take

this long or be this complicated." To MSKCC and IBM, it sounded like a great product. However, Kris believed that nobody knew what they were in for. The lack of experience of IBM in healthcare, and the complexity of AI, was apparent from the beginning of their journey.

Firstly, there are substantial differences between the structured data of medical literature and the unstructured nature of patients' electronic health records, which IBM did not anticipate. Watson learned how to quickly scan articles about clinical studies and determine some basic outcomes, but it never would be able to read the way a doctor would. Watson's thinking is based on statistics, so all it can do is gather statistics about the main outcomes, and that is not how doctors work. Additionally, Watson could not mine information from patients' electronic records as expected. Watson did have phenomenal NLP skills, as seen in Jeopardy, but in medical records, data is not as organized. Data might be written down ambiguously, out of order, or entirely missing (Strickland, 2019).

In addition, as previously mentioned, IBM executives claimed that Watson provided accurate treatment recommendations based on real patient data, but this was not the case. There was a clear discrepancy between the perception that IBM had of their product capabilities and the reality of its performance. This gap between their claims and actual outcomes revealed their lack of understanding and knowledge in a complex medical field.

Overall, this protracted timeline underscored their lack of judgment and failure to grasp the magnitude of taking on such an intricate field of healthcare technology. The deficiency that IBM had in competence underscores their unpreparedness for the challenges they faced. This lack of readiness ultimately led to their immoral actions and failure to accurately and effectively provide cancer treatment recommendations in oncology. These consequences emphasized the ethical imperative to prioritize competence when working in critical domains such as healthcare.

Conclusion

In conclusion, while much attention has been given to the technical challenges associated with this product, and the reliability of its validation metrics, a deeper examination of the decisions made by the various actors involved in its creation reveals a critical absence of key virtues essential for company executives. Specifically, their ability to communicate effectively and prioritize quality, persevere and be open to correction, and be competent for the task, were notably lacking in this case. From the perspective of virtue ethics, the decisions made by certain individuals should be regarded as ethically questionable.

Ultimately, part of IBM Watson's downfall can be attributed to their failure to carry out moral actions by neglecting the necessary groundwork of virtue ethics in technology. As developers of a healthcare product, it is imperative to integrate virtue ethics into decision-making processes. This approach emphasizes the importance of embodying virtuous qualities, which are essential for building trust and ensuring ethical conduct in the development and deployment of healthcare technologies (Frownfelter, 2021).

References

Danielson, R. (2020, July 1). 14.3: What is an informative message? *Social Sci LibreTexts*. https://socialsci.libretexts.org/Courses/Lumen_Learning/Book%3A_Business_Communication_Skills_for_Managers_(Lumen)/14%3A_Module_11-_Communicating_Different_Messages/14.03%3A_What_is_an_Informative_Message

Merriam-Webster. (2024, March 7). *Competence*. https://www.merriam-webster.com/dictionary/competence

Merriam-Webster. (2024, February 27). *Quality*. https://www.merriam-webster.com/dictionary/quality

Frownfelter, J. (2021, April 8). Why IBM Watson Health could never live up to the promises. *MedCity News*. https://medcitynews.com/2021/04/why-ibm-watson-health-could-never-live-up-to-the-promises/

Giuffrè, M., & Shung, D. L. (2023). Harnessing the power of synthetic data in healthcare: Innovation, application, and privacy. *Npj Digital Medicine, 6*(1), 186. https://doi.org/10.1038/s41746-023-00927-3

Hare, W. (1987). Open-mindedness in moral education: Three contemporary approaches. *Journal of Moral Education, 16*(2), 99–107.

Hare, W. (2003). The ideal of open-mindedness and its place in education. *Journal of Thought, 38*(2), 3–10.

IBM. (n.d.). *IBM Watson*. Retrieved from https://www.ibm.com/watson?utm_content=SRCWW&p1=Search&p4=43700074359379220&p5=e&gad_source=1&gclid=CjwKCAiAi6uvBhADEiwAWiyRdr2XuXE276JgPV

YszrNUc8oVxmA5JIIFmOMEvGXvgjie9k-WDOkOpBoC5AAQAvD_BwE&gclsrc=aw.ds

Jordon, J., Szpruch, L., Houssiau, F., Bottarelli, M., Cherubin, G., Maple, C., Cohen, S. N., & Weller, A. (2022). Synthetic data—What, why and how? *arXiv*. https://arxiv.org/abs/2205.03257

Konam, S. (2022). Where did IBM go wrong with Watson Health? *Quartz*. https://qz.com/2129025/where-did-ibm-go-wrong-with-watson-health/

Lohr, S. (2021). IBM is selling Watson Health to a private equity firm. *The New York Times*. Retrieved January 21, 2022, from https://www.nytimes.com/2022/01/21/business/ibm-watson-health.html

Mearian, L. (2018, November 14). Did IBM overhype Watson Health's AI promise? *Computerworld*. https://www.computerworld.com/article/3321138/did-ibm-put-too-much-stock-in-watson-health-too-soon.html

Miller, A. (2013). The future of health care could be elementary with Watson. *Canadian Medical Association Journal, 185*(9), E367–E368. https://doi.org/10.1503/cmaj.109-4442

Persevere. (n.d.). In *Merriam-Webster.com dictionary*. Retrieved March 8, 2024, from https://www.merriam-webster.com/dictionary/persevere

Pritchard, M. (2001). Responsible engineering: The importance of character and imagination. *Science and Engineering Ethics, 7*(3), 391–402.

Ross, C., & Swetlitz, I. (2018, July 25). IBM's Watson supercomputer recommended "unsafe and incorrect" cancer treatments, internal documents show. *STAT*.

https://www.statnews.com/2018/07/25/ibm-watson-recommended-unsafe-incorrect-treatments/

Strickland, E. (2019, April 2). How IBM Watson overpromised and underdelivered on AI health care. *IEEE Spectrum*. https://spectrum.ieee.org/how-ibm-watson-overpromised-and-underdelivered-on-ai-health-care

Tupasela, A., & Di Nucci, E. (2020). Concordance as evidence in the Watson for Oncology decision-support system. *AI & Society, 35*(4), 811–818. https://doi.org/10.1007/s00146-020-00945-9

van de Poel, I., & Royakkers, L. (2011). Ethics, technology, and engineering: An introduction. *Blackwell Publishing*.

Zhou, N., Zhang, C.-T., Lv, H.-Y., Hao, C.-X., Li, T.-J., Zhu, J.-J., Zhu, H., Jiang, M., Liu, K.-W., Hou, H.-L., Liu, D., Li, A.-Q., Zhang, G.-Q., Tian, Z.-B., & Zhang, X.-C. (2019). Concordance study between IBM Watson for Oncology and clinical practice for patients with cancer in China. *The Oncologist, 24*(6), 812–819. https://doi.org/10.1634/theoncologist.2018-0255