Life in the hands of a computer:
Ethical challenges of using AI in medical triage

David Zebedee Kleinsorge

October 10, 2024

Life in the hands of a computer: Ethical challenges of using AI in medical triage

The ethical and practical challenges inherent in medical triage have persisted throughout the history of healthcare. In moments of crisis, where the need for care exceeds the resources available, the weight of decision-making falls upon medical practitioners who must balance urgency, fairness, and the sanctity of human life. The ceaseless and rapid technological advancement that characterizes this day and age does not ameliorate these problems, but actually compounds them. That is, the latest and greatest medical technology has become an invaluable resource that, in many cases, may mean life or death. With this realized, the surrounding ethical debates and controversies are a natural consequence. Furthermore, we should expect to see an increase in this category of dilemma just as surely as humanity continues to innovate. It is our responsibility as innovators to continually deliberate about the ethical implications of that which our innovative spirit produces. The next rung in the ladder of technology in medicine is the application of Artificial Intelligence (AI), the deliberation and implementation of which has already begun. TriageGO, for example, is an active AI-powered triage system in use at Johns Hopkins today (Johns Hopkins Technology Ventures, 2022). Algorithmic and AI-powered triage systems like these are impactful and growing in popularity but require conscientious consideration and tactful implementation.

This essay will focus on AI's potential impact on triage in particular. The use of AI in triage settings offer significant benefits, such as speed, consistency, and scalability. However, their implementations raise ethical challenges, including the risk of undermining human judgment, learned biases, lack of transparency, and the impact on medical professionals. To address these challenges, it is essential to consider safeguards like the thoughtful establishment of ethical frameworks, bias audits, and human oversight. Looking to the future, as AI models become more adaptive and embedded in healthcare systems, our responsibility is to ensure that these tools remain aligned with the values of equity, compassion, and accountability that define the practice of medicine.

The Making of a Moral Machine

It is important to understand that AI as we know it today is not able to "solve" the ethical dilemmas the humankind grapples with (presuming that there is a solution). Any computer system that we implement today will only be as ethically attuned as the maxims we provide. This of course means that the ethical maxims we ultimately provide any AI should be thoroughly considered. We will consider the crisis standards of care as outlined by the National Academy of Medicine (Institute of Medicine, 2012), informed by the following ethical principles:

- The duty to care (beneficence): Prioritizing actions that maximize patient well-being.
- Fairness (justice): Ensuring equitable access and avoiding discrimination.
- The duty to steward resources (utility): Maximizing the overall benefit derived from limited resources.

These principles are foundational to the ethical frameworks that guide crisis standards of care. However, these concepts are too abstract for AI (or human intelligence) to use effectively. To serve as decision-making criteria, they must be operationalized. Ethicists recognize the value of creating and analyzing hypotheticals as a method for developing more concrete ethical principles. The following scenarios, a variation of the trolley problem, highlight the myriad of variables that medical practitioners and institutions must consider when deciding on crisis standards of care. For example, consider a single scarce medical resource, such as a mechanical ventilator, and two hypothetical patients:

- One patient is 19 years old and the other is 91 years old.
- One patient has many comorbidities, the other has few.
- One patient is paying to be treated, the other is not.
- One patient is pregnant, the other is not.
- One patient is a convicted criminal, the other has an unblemished record.
- One patient is a convicted violent criminal sentenced to die by lethal injection in a month, the other has an unblemished record.
- One patient has been receiving treatment and continues to require it, but their recovery is slow because the patient suffers from comorbidity x. Three more patients show up who don’t suffer from comorbidity x and therefore are expected to benefit greatly and quickly from treatment.

While these simplified examples control for all variables except one, and therefore poorly reflect the complexities of reality, they are valuable for constructing ethical maxims. They illuminate the difficult questions and trade-offs involved in allocating resources during crises. Once a set of guiding principles have been formulated using this method, they can be used as the foundation for a decision-making AI.

Such systems are built using advanced machine learning models trained on extensive datasets of medical cases, patient outcomes, and resource availability. The aforementioned moral laws, beneficence, justice, and utility, are translated into the AI's decision-making process through weighted algorithms and constraints within optimization models. The system is then tested and refined using simulations and real-world scenarios to evaluate its performance and adherence to ethical guidelines.

Benefits of Using Algorithms in Triage

The theoretical and practical benefits of algorithms in medical triage are undeniable. Chief among them is their ability to process vast amounts of patient data with a speed and precision unattainable by human practitioners alone. In a high-stakes environment where seconds can mean the difference between life and death, algorithms equipped with predictive capabilities have proven invaluable.

For instance, during the COVID-19 pandemic, hospitals relied on algorithmic models to prioritize the allocation of ventilators and ICU beds as explored by Cardona et al. (2020). Personalized predictive models, such as those described by Wollenstein-Betech, et al. (2020), utilized patient preconditions and health data to determine the likelihood of hospitalization, mortality, and the need for intensive care. Expanding on these applications, Ortiz-Barrios, et al. (2023) highlighted the role of artificial intelligence in capacity management, combining AI and discrete-event simulation to optimize ICU usage during the pandemic. Their study underscores the evolving capacity of AI to go beyond static decision-making, offering dynamic insights that improve both the efficiency and equity of resource allocation. These tools not only optimized resource distribution but also illustrated the capacity of algorithms and AI to synthesize complex datasets into actionable insights, allowing clinicians to make informed decisions under extreme pressure. In these moments, algorithms and AI served as amplifiers of human expertise, extending the reach of care in ways previously unimaginable.

Another significant advantage of AI systems lies in their consistency in decision-making, which surpasses that of human practitioners who may be influenced by fatigue, stress, or implicit biases. Large Language Models like ChatGPT maintain uniformity across cases, reducing the likelihood of oversight and enhancing the overall quality of care. This potential is supported by Korean researchers Kim and colleagues (2024) who compared the accuracy of human assessments with those made by different versions of ChatGPT. Their findings suggest that the latest versions of ChatGPT could serve as reliable decision-making support systems in performing triage tasks in emergency departments (Kim et al.). Yet, as with all resources, their consistency is contingent upon the integrity of their design—a subject that invites ethical scrutiny and underscores the importance of human vigilance.

Ethical Challenges and Limitations

Despite the allure of AI in triage, the incorporation of such a tool begets a number of difficulties. Medical practitioners, though imperfect, bring a depth of understanding that extends beyond metrics and probabilities. They interpret not only symptoms but also the lived experiences of patients, incorporating nuances that an AI may not know how to identify. For example, in a case where two patients present with identical clinical profiles, a clinician might prioritize the individual with fewer social supports, recognizing that their survival hinges more critically on medical intervention. An algorithm, bound by its parameters, might lack the capacity to make such a distinction, thereby reducing care to a series of calculations divorced from context. While algorithms may excel in situations of clarity and abundance, their limitations become stark in the ambiguity that define many triage scenarios.

It also remains an issue that, left unchecked, learning algorithms have the proclivity to perpetuate or even exacerbate existing biases. Unlike human practitioners, who may recognize and address systemic biases within a healthcare system, algorithms are limited to the data on which they are trained. If this data reflects historical disparities—such as the underrepresentation of certain demographics in medical research—the algorithm may unintentionally encode these injustices, delivering care that is efficient but inequitable.

Another ethical limitation in AI-supported triage lies in the issue of transparency. Algorithms, particularly those built on machine learning models, often operate as opaque systems whose decision-making processes are difficult, if not impossible, to fully understand. This lack of interpretability raises questions about accountability. If an algorithm's recommendation results in harm, who bears the ethical and legal responsibility? The clinician who implemented the recommendation? The developers who designed the system? Or the institution that chose to deploy it? In a discipline as sensitive as medicine, where trust forms the cornerstone of patient care, these ambiguities pose a significant problem.

The considerations above focus on the patients, but it's important to recognize the potentially negative impact of AI integration on the healthcare providers as well. Dr. Lauris Kajian notes that “the requirements of triage protocols that strive to maximize outcomes across populations may create serious ethical tensions for clinicians who see it as their professional responsibility to provide the best available care to each of their individual patients.” In this context, replacing the decision maker with AI may either ameliorate or exacerbate this concern depending on the circumstances and personalities of those affected. On the one hand, having triage decisions made “by a computer” might be reassuring as they are less prone to mistakes and have greater access to relevant data. On the other hand, these decisions could be viewed by practitioners as cold and inconsiderate, especially if a decision is made to rescind care. Concerns like these highlight the need for a thoughtful, careful, and conscientious implementation of AI.

Proposed Safeguards and Solutions

The integration of algorithms into medical triage requires a robust framework of safeguards. First and foremost, the development of ethical frameworks specific to algorithmic triage is paramount. Bioethicists, healthcare professionals, and technologists must collaborate to establish hierarchies of ethical principles that algorithms can apply consistently. These frameworks should be robust, accommodating diverse cultural contexts and evolving with advancements in technology.

Another critical safeguard is the rigorous auditing of algorithms for bias. Historical data, while rich in clinical insights, often reflects societal inequities. Algorithms trained on such data may inadvertently perpetuate these disparities, disadvantaging vulnerable populations. Regular bias audits, combined with the use of diverse and representative datasets, are essential to mitigate this risk (e.g. Aldrees, et al, 2022) Predictive models should be tested across demographic groups to identify and correct patterns of inequitable outcomes before they are implemented in real-world settings.

Human oversight is another indispensable safeguard. Algorithms should function as decision-support tools rather than autonomous decision-makers. By placing clinicians at the center of the triage process, we preserve the capacity for intuition and compassion—qualities that remain critical in ambiguous or ethically complex situations. For example, an algorithm might prioritize patients based on survival probabilities, but a clinician can integrate contextual factors, such as a patient's social support system or personal values, to make a more holistic decision.

Equally important is the need for transparency in algorithmic design and decision-making processes to ensure proper responsibility and accountability. Medical practitioners must understand the logic and limitations of the tools they use, particularly in high-stakes scenarios. To achieve this, developers should prioritize the creation of interpretable models, ensuring that clinicians can trace and question the rationale behind a given recommendation. Transparency not only fosters trust but also allows for accountability, a cornerstone of ethical medical practice.

In the same vein, stakeholders will need to consider augmenting medical education and training. As AI becomes an inevitable collaborator in clinical settings, training programs must prepare practitioners to work effectively with intelligent systems. Beyond technical skills, clinicians will need a deep understanding of the ethical complexities these technologies introduce, equipping them to navigate the blurred boundaries between human judgment and algorithmic recommendations. In this way, we will meet AI in the middle.

Conclusion

The integration of algorithms into medical triage represents a profound shift in the way healthcare decisions are made. These tools hold the potential to revolutionize triage by enhancing speed, consistency, and efficiency, particularly in high-stakes, resource-limited scenarios. Yet, their power also demands caution. Algorithms are not neutral actors; they are shaped by the data and principles we provide, reflecting both our ethical strengths and our blind spots. Their adoption, therefore, must be guided by a commitment to fairness, transparency, and human oversight.

The future of algorithms in medical triage is not a question of "if" but "how." The rapid rise of AI technology underscores the urgency of deliberate preparation. By addressing challenges such as bias, privacy, and the erosion of human-centered care, we can build systems that complement, rather than replace, the judgment of clinicians. Bioethicists, technologists, and medical professionals must collaborate to ensure that every line of code and every decision-making model reflects the values that define the practice of medicine: equity, beneficence, and respect for human dignity. By doing so, we can shape a future where innovation and ethics coexist.

References

Aldrees, A., Poland, C., Irshad, S.A. (2022). Auditing Algorithms: Determining Ethical Parameters of Algorithmic Decision-Making Systems in Healthcare. In: Lossio-Ventura, J.A., et al. *Information Management and Big Data. SIMBig 2021. Communications in Computer and Information Science*, vol 1577. Springer, Cham. https://doi-org.proxy.lib.umich.edu/10.1007/978-3-031-04447-2_20

Cardona, M., Dobler, C. C., Koreshe, E., Heyland, D. K., Nguyen, R. H., Sim, J. P. Y., Clark, J., & Psirides, A. (2021). A catalogue of tools and variables from crisis and routine care to support decision-making about allocation of intensive care beds and ventilator treatment during pandemics: Scoping review. *Journal of Critical Care, 66*, 33–43. https://doi.org/10.1016/j.jcrc.2021.08.001

Institute of Medicine. (2012). Crisis standards of care: A systems framework for catastrophic disaster response. The National Academies Press. https://doi.org/10.17226/13351

Johns Hopkins Technology Ventures (2022, October 11). Digital health startup that assists emergency department decision making acquired. https://ventures.jhu.edu/news/stocastic-beckman-coulter-acquisition-digital-health/

Kim, J. H., Kim, S. K., Choi, J., & Lee, Y. (2024). Reliability of ChatGPT for performing triage task in the emergency department using the Korean Triage and Acuity Scale. *Digital Health, 10*(1), 1–9. https://doi.org/10.1177/20552076241227132

Ortiz-Barrios, M., Arias-Fonseca, S., Ishizaka, A., Barbati, M., Avendaño-Collante, B., & Navarro-Jiménez, E. (2023). Artificial intelligence and discrete-event simulation for capacity management of intensive care units during the Covid-19 pandemic: A case study. *Journal of business research, 160*, 113806. https://doi.org/10.1016/j.jbusres.2023.113806

Wollenstein-Betech, S., & Cassandras, C. G., et al. (2020). Personalized predictive models for symptomatic COVID-19 patients using basic preconditions: Hospitalizations, mortality, and the need for an ICU or ventilator. *International Journal of Medical Informatics, 141*, 104258. https://doi.org/10.1016/j.ijmedinf.2020.104258