Reflections
ON HEALTHCARE MANAGEMENT
VOLUME 1 | ISSUE 1 | 2017

MACHINE INTELLIGENCE AND MEDICAL ETHICS

ABSTRACT

The practice of medicine increasingly relies on machine intelligence. Most of the existing technology functions as tools to support human activity, but some emerging uses are much more complex. Automation of complicated tasks requires greater autonomy and better decision-making skills. When lives are on the line, decisions made by machine intelligence must conform to the highest standards of utility and ethical behavior. We discuss some of the near-future medical applications of machine intelligence, and how we can ensure the ethical operation of these devices. We conclude with a review of ethical philosophies that might be used to inform machine medical ethics and the artificial intelligence techniques by which they might be implemented.

PAUL WILKENS

KEYWORDS
machine intelligence, artificial intelligence, machine ethics, machine medical ethics

AUTHOR BIOGRAPHY
Paul Wilkens has a background in biology and scientific computation, and can be reached at wilkens@sbcglobal.net.

doi:10.6083/M44T6HG3

Background: Lessons from History, Future Feasibility

Unrealistic expectations and oversimplified assumptions dominate the history of research into artificial intelligence (AI). Failures outnumber successes by a thousand to one, and the successes are limited to very specific areas and often depend on unexpected developments in other areas of knowledge. For instance, Google's barely-serviceable machine translation capability would have been impossible without the monumental translation effort in Brussels, which yielded twenty year's worth of perfectly matched texts in common European languages (Vasiljevs, 2015). In the Sixties, Chomsky (1965) argued for the existence of a universal grammar, which would allow the extremely limited computers of that period to perform excellent translations. As a result, the early history of AI contains many failed machine translation experiments. Despite decades of research in linguistics, no universal grammar has ever been discovered (Everett, 2005).

Human brains are based on well-known physical and chemical properties. Fundamental physical systems can be described mathematically and simulated by computer. Given sufficient computing resources, the human brain could be simulated. This is the essential argument of "strong AI". Most computer scientists accept this as proof of the possibility of highly intelligent and autonomous machine intelligence, which could make complex ethical decisions (Searle, 1999). If the eventual practice of AI relies on transcribed humans, the ethics of these machines is easy to understand – humans *in silico* would presumably retain the view of ethics they developed when they were flesh and blood. However, this is a problem for the distant future. Although some researchers claim to have simulated significant portions of mammalian brains (Inafuku et al., 2010), their techniques rely on a mathematical oversimplification of the neuron. Successfully simulating human brains will require *biologically realistic* neurons, which are vastly more computationally expensive (Wilkens, 1993). If every computing device on the face of the planet was configured to simulate realistic neurons, and every network connection was devoted to inter-neuronal communication, we would still be several orders of magnitude short of the computational power required to simulate a single human brain in real time. Even with exponential growth in computing resources, strong AI will take many decades to realize.

Near-term Prospects for Machine Intelligence in Medicine

Several thousand research articles on medicine and the related sciences are published every day. No medical professional can possibly keep up with all the developments in their field. Even reading a fraction of one percent of the relevant articles would take several hours a day, but providers *must* stay current. The information in those articles is essential to the practice of evidence-based medicine. Many practitioners feel that they are ethically required to accomplish an impossible task.

Assistance from machine intelligence might help us solve this problem. Many different systems have been proposed, but all seek to inform the provider of context-specific research results and help the provider apply current knowledge to each patient individually. At every encounter, the patient's condition, prescriptions, and co-morbidities will be evaluated and compared to the latest research results, and recommendations will be displayed. For instance, if one of the patient's medications was the subject of a just-released meta-analysis that found it to have significant side effects versus alternatives, the AI assistant will inform the provider immediately.

Many providers hope that this type of technology will rescue the practice of evidence-based medicine from the overwhelming flood of data, which no human can keep up with. However, this technology makes decisions that have ethical implications. Consider a simple example: when a patient visits the provider, the machine assistant analyzes the patient's latest information and outputs a list of important facts in order of relevance. *The act of ordering the list requires a series of ethical decisions.* Should the machine give higher weight to research that suggests a less expensive treatment

alternative? Should it include information about a new treatment that is very expensive but has almost no side effects? And should this change based on whether the patient is experiencing any of those side effects currently? Every decision is potentially controversial.

Machine intelligence may also be taking a greater role in logistics management for medical systems. Computers are very good at tracking tens of thousands of items, monitoring usage rates, and ordering replacements as necessary. These systems will occasionally have to make ethical decisions regarding the apportionment of scarce supplies. In the event of a scarcity conflict, we would design the system to raise an alert and seek human intervention – but what if the operator fails to respond? Delaying the distribution of a scarce supply to all locations while waiting for operator input is probably a worse decision than distributing some supplies to the locations that are most likely to need them, even if the machine is somewhat uncertain about which are "most likely". These systems will also need some ethical mechanism for decision making.

Potential Ethical Bases for Autonomous Decisions

Utilitarianism

Asimov's "three laws of robotics" supported some good storytelling, but they are not a sufficient basis for ethical decision-making. Consider this thought experiment, in the style of Frances Kamm (1993):

> *The robot notices the falling steel girder when it is still 1500 milliseconds from the ground. Two humans are in its shadow, and reliable projection indicates that both will be killed. The robot is aware of the limits of its physical capacity; it can save one but not both. Should it rescue the child in the stroller, or the middle-aged man?*

If you're human, you effortlessly decided what the robot should do before you finished reading the sentence. What guided your decision? Is it simple utilitarianism, because the child has more to lose? Our culture also promotes sacrifice for children. Did that play a role?

> *The child is in the last stage of a terminal illness for which all treatments have failed. Desperate to spend a few more moments together, the father is taking his child for a walk. They don't have much time. There are three more children at home, waiting for dinner. Their mother died last year.*

Did your decision change? Perhaps you were reasoning from utilitarianism. After learning the full context, we base our decision partly on the number of years of life that we can save, and partly on the impact of the loss on the larger family.

Utilitarianism is appropriate for emergency situations that require simple decisions between unequal outcomes. To make the best use of utilitarian principles in machine ethics, we should first concentrate on the immediate problem: how do we recognize when a state of emergency exists?

Utilitarianism does not provide sufficient guidance for larger ethical dilemmas. Utilitarianism is analogous to pure democracy: without constitutionally guaranteed rights, the majority can exercise unlimited control. Every healthy twenty year old contains enough donor organs to save several lives, potentially gaining far more Quality-Adjusted Life Years than are sacrificed. Utilitarianism demands that we make this sacrifice. It makes no provision for individual rights (Jonsen et al., 2010).

Kant

Harvesting a healthy twenty-year-old is clearly a violation of the principle of ends, because it treats the individual as a means (Enders, 2015). It is also a violation of the categorical imperative or Golden Rule, but the Golden Rule is difficult to express in language that a machine can process. Short of strong AI, machines do not have any desires as to how they should be treated.

Despite the complexity of the categorical imperative, some researchers have tried to use it as a basis for machine ethics (Powers, 2011). They design a mechanism for ethical decision-making that depends on a large number of carefully prioritized rules. Theoretically, the rules are derived from core principles, but in practice, cultural preferences are also incorporated. The result is a series of guidelines that describe how most people would like to be treated. The machine can use these guidelines to assign ethical scores to its potential actions.

In practice, rigid systems of rules are vulnerable to gaming and other forms of manipulation. If we know how the guidelines are written and structured, we can design interactions to guarantee certain outcomes. As new manipulations are discovered, the original system can be refined to resist some forms of attack, but these refinements inevitably lead to new vulnerabilities. The inherent limitations of rigid rule-based systems are at the heart of one of the most pressing technological problems of our time – our information technology is vulnerable to hacking, and as long as it's based on rigid rules, it always will be.

When Kant was still alive, Constant asked him if it was ethical to lie to a murderer when he asks you for the location of his next intended victim. For the sake of argument, they constructed a hypothetical situation in which refusing to speak was not possible. In this case, Kant (1797) says you must tell the truth, and thereby turns an otherwise reasonable statement of principles into a murder pact. Most humans would try to tell an effective lie that both keeps the victim safe and makes the murderer's apprehension more likely. In fact, some researchers argue that our human propensity to lie and otherwise violate the rules is one of the reasons that humans can't be hacked (Santos-Lang, 2012; Santos-Lang, 2014). We can be fooled, but not hacked. Con artists are manipulating hopes, dreams, and greed; their victims willfully participate. They are not manipulating a rigid rule-based system to guarantee outcomes contrary to the victim's will.

Rawls

Reasoning from behind the veil of ignorance, we cannot know whether we are rich Westerners or residents of a developing country living on $2.00 per day. In this situation, the correct action is obvious: use whatever resources are necessary to transform the developing world and make it as rich and well-educated as the West. If we lift the veil after this is done, we'll be happy whether we find ourselves in Tokyo or Raqqa.

Most Rawlsians would acknowledge that precise equality is not a realistic goal, but a strict interpretation of Rawlsian ethics suggests that wealthy nations should not be spending government dollars on any non-essential effort that does not directly help the developing world. All space missions, all medical research except infectious diseases, basic research in physics, chemistry, and biology, the national parks, the arts, subsidized sports – it all must go. The money should be spent on uplift. After tremendous effort, all peoples of all nations will experience equal life expectancies and quality-of-life metrics. Once we have reached total global equality, we can revive our stalled research and development and return to the path of progress. This may seem ridiculous, but there are many potential advantages. If all humans were rich and highly educated, our scientific and technological capabilities would be many times greater than they are now. The well-off are also less likely to engage in conflict, because they have more to lose.

Unfortunately, simple observation indicates that none of the wealthy nations have broadly and consistently adopted Rawlsian ethics. This seems to be the biggest practical problem with implementing Rawls at the machine level: *the machines won't know that we're not actually serious.* They will try to make decisions that truly embrace the Rawlsian ideals, which will most likely lead to surprised humans and deactivated machines.

Limited Consequentialism or Preference Utilitarianism

Determining ethical behavior on the basis of expected results is highly problematic in complex situations, in part because ultimate consequences

are unknowable. We may be relatively confident of the first-order results, but second and third-order effects are unpredictable, and unintended consequences are unavoidable. However, in the limited areas where machine ethics is most likely to find early applications, consequences are often surprisingly easy to calculate, and higher order effects are subject to system constraints. Even when the results for all possible decisions are predictable, some utilitarian mechanism must determine desirability. In a highly specialized system, calculation of the utilitarian metric can be designed to match the specialized task, avoiding all issues with applicability. Limited consequentialism is a viable option for weak AI, and has already been implemented in many machine intelligence systems (Tomasik, 2015).

Options for Implementation

Machine intelligence is not yet capable of reliably interpreting natural-language instructions. Ethical guidelines must be coded in a symbolic language that is compatible with the machine's data sources and decision-making mechanism. Depending on the ethical principle, several different implementation options are available.

First, consider the ethical guidelines that humans find most useful: the test of common sense; the test of one's best self; the test of making something public; the gag test. All of these are very difficult to implement. Most attempts will resemble the implementation of the categorical imperative – thousands of rules, written by hand and carefully prioritized, which attempt to approximate the instinctual, emotional, and reasoned aspects of human ethical preferences.

Practical options for the implementation of machine ethics include expert systems, decision trees, rule-based systems, heuristics, and neural networks.

Expert systems are the oldest form of machine intelligence. To develop an expert system, programmers interview human experts and expose them to thousands of scenarios. The human expert describes the guidelines and principles that helped them decide on a course of action, and the programmers transform these into thousands of if-then statements. In some fields, a well-programmed expert system can be extremely effective in mimicking a human expert, and expert systems have found many applications in medicine.

Decision trees are similar to expert systems, but more general and flexible. They do not always require expert input to program; some can be programmed by inputting natural data, like the last fifty years of case law on clinical ethics.

Rule-based systems are amalgamations of carefully prioritized rules that can be used to assign scores to potential actions, like the attempt to computerize the Golden Rule described earlier.

Heuristic systems are designed to look at all possible futures and eliminate as many as possible. IBM's chess champion used a heuristic system to examine all possible outcomes of all possible moves; in practice, it's often possible to eliminate many of the non-productive areas of the search space by applying some rule-based constraints. Heuristic techniques are an excellent implementation option for consequentialist machine ethics.

Neural networks seek to replicate the learning powers of animal brains, and are capable of amazing feats of pattern recognition and complex computation. In practice, the most useful machine intelligence often combines these techniques. For instance, IBM's Jeopardy-winning Watson is a neural network that uses a rule-based system for natural language recognition and heuristics to trim potential responses.

Transparency of ethical decisions can be very important. Of the techniques listed above, only expert systems, decision trees, and rule-based systems are capable of providing total transparency in the decision-making process. They are also easier to game and manipulate. Heuristics provide some transparency, but the traces can be very difficult to interpret. Neural networks are mysterious black boxes – understanding how a decision was made requires special expertise and the results are extremely difficult to explain to non-

experts. Machine ethics for medical technology will also sometimes require privacy. If a machine's ethical decision depended in part on protected health information which must be kept private, then transparency will be necessarily limited, allowing the use of a less-transparent implementation technique.

Conclusion

Which of these ethical principles will be most effective at helping the physician's AI assistant to properly order the list of search results for their patient? For various reasons, utilitarianism and Kantian ethics are incomplete solutions, Rawlsian ethics are often inapplicable, and consequentialism is of limited use for this problem. More research is needed.

When you next run a medical database search for a patient's condition, please take note of the order of the results. Does the relative prioritization match your professional and personal ethics? Did the search designers consider the ethical implications of their weighting algorithm? Health care professionals may not realize the influence that their machine assistants can have on their actions.

References

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. MIT Press.

Enders., J. (2015). *Foundational Concepts in Organizational Business Ethics* (lecture). Portland State University, Portland, OR.

Everett, D. (2005). Cultural Constraints on Grammar and Cognition in Piraha. *Cultural Anthropology*, *46(4)*, 621-646.

Inafuku, J., Lampert, K., Lawson, B., Stehly, S., Vaccaro, A. (2010). Downloading Consciousness. Retrieved from http://cs.stanford.edu/people/eroberts/cs181/projects/2010-11/DownloadingConsciousness/tandr.html

Jonsen, A. R., Siegler, M., Winslade., W. J. (2010). *Clinical Ethics, 7th Ed.* McGraw-Hill.

Kamm, F. (1993). *Morality, Mortality, Vol. 1: Death and Whom to Save From It*. Oxford University Press.

Kant, I. (1797). *On a Supposed Right to Tell Lies from Benevolent Motives.*

Powers, T. M. (2011). Prospects for a Kantian Machine. *Machine Ethics*. New York: Oxford University Press.

Santos-Lang, C. (2012). *Ethics for Artificial Intelligences (version 3)*. Retrieved from https://santoslang.wordpress.com/article/ethics-for-artificial-intelligences-3iue30fi4gfq9-1

Santos-Lang, C. (2014) Moral Ecology Approaches to Machine Ethics *Machine Medical Ethics*. Switzerland: Springer.

Searle, J. (1999). *Mind, language and society*. New York, NY: Basic Books.

Tomasik, B. (2015). *Machine Ethics and Preference Utilitarianism*. Retrieved from http://reducing-suffering.org/machine-ethics-and-preference-utilitarianism/

Vasiljevs., A. (2015). The role of machine translation in Europe's digital single market strategy. *Brussels Times*. Retrieved from http://www.brusselstimes.com/magazine2/3027/the-role-of-machine-translation-in-europe-s-digital-single-market-strategy

Wilkens, P. (1993). *Computational Model of a Realistic Neuron* (unpublished thesis). Reed College, Portland, OR