The Philosophy and Ethics of Artificial Intelligence

Dr. Sanjay Modgil
N7.08 Bush House
Office hours:

Overview of Course

- Introduction
- Intelligence
- Consciousness
- Reasoning and Communication
- Ethics and Morality
- Algorithms
- For Good or bad ? AI in Medicine, AI in War
- Superintelligence and the Value Loading Problem
- AI and Human Society
- Review/Revision

For Good or bad ? AI in Medicine, AI in War

-
- Lethal Autonomous Weapons: Arguments for and Against Banning them
- AI in Medicine. Examples and Ethical Issues
- Towards a Professional Code of Ethics

Lethal Autonomous Weapons Systems (LAWS)

- The issue of whether to continue with, or ban, Autonomous Weapons reveals a number of key ethical issues that arise when algorithms make decisions in autonomous systems

Definition

A weapon that can complete an entire engagement cycle on its own. It searches for targets, identifies them, makes a decision about whether or not to attack them, and then starts the engagement and carries through the engagement all by itself.

So there's *no human in the loop* - the cognitive loop of sensing and deciding and acting out on the battlefield.

LAWS – Current Developments

- R&D in autonomous weapons especially in US, UK, China, Russia, France, Israel, and South Korea
- By and large LAWS do not yet exist, but there are exceptions, e.g., Israeli *Harpy* drone which targets radar signatures and not humans (although a person standing next to the radar will most likely be killed).
- However in 2016 reports of the second generation *Harpy* 2 used in conflict between Azerbaijan and Armenia to destroy and kill a bus full of Armenian soldiers.
- *Harpy* 2 does have a 'human in the loop mode' although unclear whether there is a fully autonomous version.

LAWS – Current Developments

- Many organisations and scientists are calling for an outright ban on the development and deployment of LAWS e.g.

- *The Future of Life Institute* (FLI) (includes Stephen Hawkings, Elon Musk, Stuart Russell ….) initiated a pledge (2015) not to participate in nor support the development, manufacture, trade, or use of lethal autonomous weapons.

  Signed by Google Deep Mind, UCL, the European Association for AI, robotics companies and leading AI researchers.

- in 2017, 137 CEOs of AI companies asked the UN to ban LAWS

To ban or not to ban, that is the question

One answer is to wait and see: too early to know where technology will go.

However we need to think now about these issues

Before development gets fixed on a particular path and before design architecture fixed and difficult to change so as to take account of the law and ethics that ought to inform and govern autonomous weapons systems

Before ethical and legal understanding of LAWS gets fixed in the eyes of governments, so that we can propose and defend a framework for evaluating them that advances strategic and moral interest

Arguments **For** Banning LAWS

- **Ability to control development:** Development will lead to an AI arms race → increase pressure to find shortcuts around safety precautions
- **Ability to control use:**
  - materials to build weapons cheap and easily obtainable (c.f. nuclear) →

  1) mass production and proliferation amongst military powers
  2) available on the black market so may get into the hands of terrorists, rogue states, warlords …
  3) weapons may be hacked making control by military more difficult. People with bad intentions may remove ethical safeguards in LAWS

➔ could lead to a rise in destabilizing assassinations, ethnic cleansing, and greater global insecurity. E.g. North Korea going to great lengths to get atomic weapons. Much easier to obtain autonomous weapons. If South Korea attacked by autonomous weapons and not able to defend themselves adequately, they may escalate and we may end up in nuclear conflict.

- **Increased prospects for conflict:** Possible reduction in loss of human life (soldiers and civilians) will mean more incentive to go to war.

Arguments **For** Banning LAWS

***But**, history of warfare has never seen advances in technologies increasing likelihood of war – quite the opposite. Reduction in deaths a good thing, How much war occurs and at what intensity depends on much more significant factors, ranging across law, politics, diplomacy, the effectiveness of international institutions, the nature of threats, and many other things.*

- **Technological Challenges:** In short term (10 years) significant technical challenges:

  1) how to program reasoning and judgement that respects international law
  2) how to distinguish between civilians and soldiers
  3) how to understand and respond to complex and unanticipated situations on the battlefield
  4) how to verify and validate LAWS

I.e., concerns we are still a long way away from human level AGI, and machines that are morally capable. Note that those pro-banners do not argue that these technical problems could ***never be*** solved.

***Indeed,** what about the long term ? Hasn't recent history of AI demonstrated at least in narrow domains, super-human performance ?*

Arguments **For** Banning LAWS

- **Taking Humans out of Loop:** A core conviction of many in favour of banning LAWS is that algorithms *cannot have* empathy, sympathy, compassion and therefore *cannot have* human morality, and *so cannot* judge **proportionality** (balance between military advantage and civilian harm) or make *humane* context-based kill/don't kill decisions.

***But**: are empathy, sympathy, compassion necessary for making moral judgements ? What matters is ability to consistently behave in a certain way (that leads to good **consequences**), and not whether **virtues** of empathy, sympathy … are used in making moral decisions.*

Core conviction of many pro-banners - a machine, no matter how good, cannot be a true moral agent – is based on moral emotions/instincts (just like in abortion debate). Shouldn't we instead consider whether LAWS will result in better consequences ?

*Future driverless cars / robot surgery will make life and death decisions and will be in use because they prove to be safer. Our core convictions about machine and human decision-making will evolve. We will accept machine decisions because they lead to better outcomes (indeed we might think morally wrong **not** to use LAWS).*

Arguments **For** Banning LAWS

- **Accountability** By taking the human out of the loop, we fundamentally dehumanize warfare and obscure who is ultimately responsible and accountable for lethal force.

  Its unclear who, if anyone, could be held accountable and/or responsible if a lethal autonomous weapon causes unnecessary and/or unexpected harm.

  Again, a core conviction of many in favour of banning LAWS

***But**: But It would be unfortunate to sacrifice reduced battlefield harm through machine systems (assuming there are such gains) simply to satisfy principle that a human be held accountable. It would be better to adapt mechanisms of collective responsibility borne by a "side" in war (recall discussion about how we should judge appropriate penalty if an AI systems causes harm)*

Arguments **Against** Banning LAWS

- **Unilkely that nations will adhere to Ban**: No formal ban is going to prevent people from building LAWS.

But: *Bans against landmines, cluster bombs, blinding lasers, biological weapons, chemical weapons … **have been largely successful** and have at least had a major effect in reducing use of these weapons.*

- **In principle LAWS can be better at discrimination:** Better at distinguishing civilians (in terms of technical recognition and immunity to emotions that affect human judgment e.g. panic, vengeance …) and so selectively targeting soldiers, ➔ better outcome in terms of less civilian deaths. Note that currently banned weapons are typically banned because they are *indiscriminate*.

- **In principle LAWS will be better at calculating proportionality:** Better at weighing up military advantage versus civilian deaths. Again, immune to emotions that 'cloud' human judgment

So in certain situations, LAWS might cause less harm than conventional weapons

Summing Up

- Arguments against ban point to *potential* better ability to discriminate and weigh up proportionality. No *in principle* reason to assume that LAWS cannot make appropriate moral judgments. But a serious concern that this is not achievable technically, especially in the short/medium term.
- Ability to control developments and use are perhaps overriding arguments for ban
- Arguments against ban are utilitarian. But **pragmatic utilitarianism** might suggest that we respect deeply held convictions that we should not dehumanise warfare and so keep humans in the loop. That is to say, we would on balance be happier to live in a world in which there was always a human in the loop than a world in which LAWS reduced civilian casualties.
- **Semi-autonomous** weapons might realise benefits of better identification and targeting, while leaving kill decision to human (acknowledging, as a pragmatic utilitarian should do, that the greater good is served by respecting moral conviction that human should be in loop)
- However, there is psychological research suggests that keeping a human in the loop may not be as effective as many hope, given human tendencies to be over-reliant on machines, especially in emergency situations.

For Good or bad ? AI in Medicine, AI in War

-
- Lethal Autonomous Weapons: Arguments for and Against Banning them
- AI in Medicine. Examples and Ethical Issues
- Towards a Professional Code of Ethics

Ethical Dimensions of Using AI in Healthcare

Case Study 1

AI algorithm based on a trained neural network that identifies presence of cancer cells in images of tissue samples (biopsies). Neural network algorithm able to learn from its mistakes: it is designed to have the ability to improve its scan sensitivity over time with continued use.

Images scanned in fractions of seconds with 92.4% sensitivity as compared with 73.2% sensitivity for human eye

Diagnostic sensitivity: the probability (P) that, given the presence of disease (D), an abnormal test result (T) indicates the presence of disease; that is P(T/D).

Should the algorithm be used to replace/confirm human diagnosis ?

Given higher success rate, algorithm should at the very least be used to support human pathologist. Could also be used to train pathologists.

Indeed, hospital might be sued if it were known such a program existed and would have spotted cancer cells missed by a pathologist and patient died as a result.

Ethical Dimensions of Using AI in Healthcare

Two possible ethical concerns:

1. The black box issue: Not possible to explain how neural network identifies cancerous cells.

On the other hand a pathologist's job could include trying to identify which visual features cancerous (or normal) cells have in common and verbalizing what it is about cancerous cells that prompts the AI program to identify them as cancerous. This would 'illuminate' the black box, and the transparency achieved could enable pathologists and patients to feel more comfortable relying on the AI program.

Given considerably higher sensitivity, and prospect of learning from mistakes resulting in sensitivity approaching 100%, should we care how it does it ?

All we should care about is its success rate and hence increase benefits for patients (utilitarianism)

Ethical Dimensions of Using AI in Healthcare

Recall the discussion about LAWS (*What matters is ability to consistently behave in a certain way (that leads to good **consequences\outcomes**), and not whether **virtues** of empathy, sympathy … are used in making moral decisions*)

Contrast with self-driving cars: lives are at stake when a pedestrian suddenly appears on a road in a tunnel and the car "chooses" between swerving left or right and so possibly injuring the driver, or continuing straight on, braking, but with insufficient distance to avoid colliding with the pedestrian.
Not understanding why a self-driving car does what it does would not be ethically acceptable. We want the basis for such "decisions" to be made understandable

2. Automation Bias: Concerns that complacency sets in when a job once done by a health care professional is transferred to an AI program (health care professional gets lazy about using his/her clinical judgement skills)

Again, ethically this may not matter ? All that matters is the success rate. Unless the degrading of such skills has other negative consequences.

Ethical Dimensions of Using AI in Healthcare

Case Study 2

IBM Watson can be used as a decision support system (DSS) to *support* and not necessarily replace a clinicians diagnosis and treatment decisions.

It uses natural language processing, information retrieval, semantics analysis, automated reasoning and machine learning.

It famously beat champion Jeopardy game show contestants in 2011

Ethical Dimensions of Using AI in Healthcare

It works as follows:

Fed with massive amounts of data from clinical literature, health records, and test results.

Clinician poses query describing patient symptoms and other related info. Watson parses input to identify most important info and then mines patient data to find relevant facts about the patient's clinical and hereditary history.

Watson then examines available data (previously inputted) to form and test hypotheses and finally lists individualised, confidence-scored recommendations.

The system can then describe the supporting evidence in text form for its ranked responses. Because information is constantly being fed to Watson, the system can learn over time to optimize its recommendation

Ethical Dimensions of Using AI in Healthcare

IBM Watson HealthTM presently offers commercialized applications of Watson for genomics, drug discovery, health care management, and oncology.

IBM has partnered with several academic and private institutions to apply Watson to patient care research and treatment.

Some key ethical and legal questions are:

1. Should Watson ever replace the clinical judgment of a doctor ?

2. What are the liability concerns of professionals who use Watson ?

3. What are the limitations of Watson and ethical implications of these limitations ?

Ethical Dimensions of Using AI in Healthcare

1. Watson's Role

According to IBM, Watson is intended to **assist** and **enhance** the decision making of health care professionals by giving them greater confidence in their diagnostic and treatment decisions for their patient.

It is not intended to replace the judgment of health care professionals. Nor should it be viewed as any kind of authoritative decision-making tool.

In the US the Food and Drug Administration (FDA) regulates the safety and effectiveness of devices and drugs. Because Watson is considered a management tool under the control of physicians and not a device, the system does not presently require regulations to control its use !

However, regulatory requirements could change as Watson and other emerging AI systems are used to make diagnoses or treatment decisions with little or no supervision from clinicians

Ethical Dimensions of Using AI in Healthcare

2. Liability

Use of the system as an assistant has potential to **increase** liability for health care professionals and organizations

E.g. Watson could recommend treatment that doctor decides to pursue while ignoring other contraindicating patient data, because doctor assumes Watson (or any other DSS like it) had evaluated that information. Such a scenario could result in a malpractice claim against the doctor .

Use of Watson could *prematurely* contribute to a higher legal standard of care that could put health care professionals at greater risk for negligence. Because expectations of the standard of care can change while the impact of the technology on health outcomes is not yet **fully** known.

E.g. if Watson is shown to improve diagnostic accuracy and treatment recommendations for leukemia, then expectations that doctors who consult Watson will get the diagnosis and treatment recommendation “right” could be raised to a higher level.

Ethical Dimensions of Using AI in Healthcare

3. Understanding Watson's Limitations

Possible that Watson might make a recommendation that is inconsistent with current clinical standards or that contradicts what a doctor considers to be the appropriate decision.

Eg a clinical standard might be always to prescribe a particular medication for a particular diagnosis, but Watson could recommend an alternative (eg, a nonstandard medication or no medication at all). In such a scenario, doctors must be able to support their decision to follow or not to follow the alternative and to understand the potential clinical and legal consequences.

➔Need an audit trail of how decisions are made, especially when they contradict accepted clinical standards or clinicians' recommendations
However black box issue means that ML based decisions cannot be explained !

Note that Watson can and should be designed to use a rule base limiting recommendations to current clinical standards, so ensuring recommendations are consistent with treatment guidelines and currently accepted practices.

Ethical Dimensions of Using AI in Healthcare

Despite ethical concerns:

The need for intelligent systems such as Watson is clear given the exponential expansion and complexity of clinical data.

E.g., IBM has suggested that a person will generate 1 million gigabytes of health-related data in a lifetime—which is equivalent to more than 300 million books !

Given the amount and complexity of patient data, physicians ***ought*** to consult intelligent systems such as Watson.

In the future, it may be considered unethical (and create liability) **not** to consult Watson or intelligent systems like it, for a second opinion, assuming that such systems prove effective in what they claim to do.

Ethical Dimensions of Using AI in Healthcare

Case Study 3

Mazor Robotics Renaissance Guidance System technology uses AI software to analyze images and plan placement of surgical tools in spinal surgery.

Suppose a surgeon recommends use of such robot assisted surgery to a patient. A number of ethical concerns arise:

1. Informed Consent and the Black Box Problem

Patient must agree (consent to) medical interventions, and this consent must be "informed". That is to say, informed consent is valid only if appropriate information has been disclosed to a competent patient who is permitted to make a voluntary choice.

AI raises a number of challenges for the informed consent process:

- presentation of information can be complicated by patient and doctor fears, overconfidence or confusion
- requires doctor to be sufficiently knowledgeable to explain to patient how AI works. But black box problem, so difficult to explain how decisions made or why errors occurred.

Ethical Dimensions of Using AI in Healthcare

At the very least the doctor should:

i) carefully distinguish the roles that the doctor/nurses play and the roles of the AI/ robotic system, when obtaining consent
(e.g. she should explain that she is responsible for the planning before the operation, whereas the Renaissance Guidance System will manually guide placement of tools or implants)

ii) clearly state the potential harms that might result from either human or robotic errors

2. Patient Perceptions of AI

A 2016 survey of 12 000 people across 12 European, Middle-Eastern, and African countries found that only 47% of respondents would be willing to have a "robot perform a minor, non-invasive surgery instead of a doctor," with that number dropping to 37% for major, invasive surgeries.

In addition to distinguishing roles, doctor should therefore address patient fears by describing risks and benefits (i.e., not just saying that system has been used in past but also studies comparing robotic system with human surgeons)

Ethical Dimensions of Using AI in Healthcare

3. Medical Errors and the Problem of "Many Hands"

The problem of assigning moral responsibility, and perhaps legal liability, when errors occur and the cause of the harm is distributed amongst multiple actors, technologies, and perhaps organisations. Black box issue makes problem worse.

Individuals might use a many hands argument in an attempt to avoid personal responsibility for bad outcomes.

A consequentialist/utilitarian approach would be to focus on penalties that:
1) minimise the possibility of future errors; 2) incentivise serious R&D effort to ensure that:

- Coders and designers document what they create and, to extent that it is possible, make technology and underlying processes (and hence errors) explainable
- Companies clearly state conditions for successful application of an AI technology, such as the quality of diagnostics, imaging, and preparation for surgical procedures. They should also detail types of errors and side effects, their likelihood and severity, and differences in predictive accuracy and error rates across demographic subgroups, health conditions, and patient histories.

Ethical Dimensions of Using AI in Healthcare

- Doctors are responsible for acquiring basic understanding of the AI devices they use and the types and likelihood of errors across subgroups, to the extent that this information is available.

  They should communicate relevant information to patients and health care teams and ensure adherence to standards provided by device companies.

  If a medical error occurs because instructions were not followed, the primary responsibility could lie with the doctor (or team); however, if a medical error occurs because adequate instructions or training were not provided by the company, the primary responsibility could lie elsewhere.
  In either case, assigning responsibility and the subsequent penalties should incentivise proper adherence to instructions/adequate instructions/training.

- Organisations should providing training, protocols, and best practices related to AI use

For Good or bad ? AI in Medicine, AI in War

- Lethal Autonomous Weapons: Arguments for and Against Banning them
- AI in Medicine. Examples and Ethical Issues
- Towards a Professional Code of Ethics

Towards a Professional Code of Ethics for Research and Development of AI

Professional codes of ethics are needed because members of a profession possess certain skills, knowledge and capacities that their clients and the general public lack. This means that professionals have power over their clients and the public who are in a vulnerable position.

Codes of ethics mitigate the harmful effects and/or misuse of such power.

For example a developer creates an application for uploading electricity meter readings. A code of ethics should dictate that the application is accessible for use by clients who may be visually impaired.

Challenges for specifying a Professional Code of Ethics for AI R&D

One problem is diversity.

1. There is a concentration of intellectual and commerical power amongst those involved in developing AI.
2. Those working in AI have a certain demographic profile. They are not immune to group think, belief polarisation. …

Other problems particularly relevant to AI are:

1. As AI becomes more and more intelligent and autonomous, professionals themselves may be vulnerable with respect to AI (AIs may have power over researchers and developers)

2. The black box problem - researchers and developers may themselves not understand how their products operate.

However, there are attempts at enumerating comprehensive codes of ethics. For example:

Asimolar Principles

Developed by Future of Life Institute (https://futureoflife.org) with the first major conference on beneficial AI (https://futureoflife.org/bai-2017/) (high profile industrialists working in AI – e.g Elon Musk - and researchers e.g Stuart Russell)

Research Issues

1) **Research Goal:** The goal of AI research should be to create not undirected intelligence, but beneficial intelligence.

2) **Research Funding:** Investments in AI should be accompanied by funding for research on ensuring its beneficial use, including thorny questions in computer science, economics, law, ethics, and social studies, such as:
- How can we make future AI systems highly robust, so that they do what we want without malfunctioning or getting hacked?
- How can we grow our prosperity through automation while maintaining people's resources and purpose?
- How can we update our legal systems to be more fair and efficient, to keep pace with AI, and to manage the risks associated with AI?
- What set of values should AI be aligned with, and what legal and ethical status should it have?

Asimolar Principles

3) **Science-Policy Link:** There should be constructive and healthy exchange between AI researchers and policy-makers.

4) **Research Culture:** A culture of cooperation, trust, and transparency should be fostered among researchers and developers of AI.

5) **Race Avoidance:** Teams developing AI systems should actively cooperate to avoid corner-cutting on safety standards.

Asimolar Principles

Ethics and Values

6) **Safety:** AI systems should be safe and secure throughout their operational lifetime, and verifiably so where applicable and feasible.
7) **Failure Transparency:** If an AI system causes harm, it should be possible to ascertain why.
8) **Judicial Transparency:** Any involvement by an autonomous system in judicial decision-making should provide a satisfactory explanation auditable by a competent human authority.
9) **Responsibility:** Designers and builders of advanced AI systems are stakeholders in the moral implications of their use, misuse, and actions, with a responsibility and opportunity to shape those implications.
10) **Value Alignment:** Highly autonomous AI systems should be designed so that their goals and behaviors can be assured to align with human values throughout their operation.
11) **Human Values:** AI systems should be designed and operated so as to be compatible with ideals of human dignity, rights, freedoms, and cultural diversity.

Asimolar Principles

12) **Personal Privacy:** People should have the right to access, manage and control the data they generate, given AI systems' power to analyze and utilize that data.
13) **Liberty and Privacy:** The application of AI to personal data must not unreasonably curtail people's real or perceived liberty.
14) **Shared Benefit:** AI technologies should benefit and empower as many people as possible.
15) **Shared Prosperity:** The economic prosperity created by AI should be shared broadly, to benefit all of humanity.
16) **Human Control:** Humans should choose how and whether to delegate decisions to AI systems, to accomplish human-chosen objectives.
17) **Non-subversion:** The power conferred by control of highly advanced AI systems should respect and improve, rather than subvert, the social and civic processes on which the health of society depends.
18) **AI Arms Race:** An arms race in lethal autonomous weapons should be avoided.
.

Asimolar Principles

Longer Term Issues

19) **Capability Caution:** There being no consensus, we should avoid strong assumptions regarding upper limits on future AI capabilities.
20) **Importance:** Advanced AI could represent a profound change in the history of life on Earth, and should be planned for and managed with commensurate care and resources.
21) **Risks:** Risks posed by AI systems, especially catastrophic or existential risks, must be subject to planning and mitigation efforts commensurate with their expected impact.
22) **Recursive Self-Improvement:** AI systems designed to recursively self-improve or self-replicate in a manner that could lead to rapidly increasing quality or quantity must be subject to strict safety and control measures.
23) **Common Good:** Superintelligence should only be developed in the service of widely shared ethical ideals, and for the benefit of all humanity rather than one state or organization.