CSCI 1800 Cybersecurity and International Relations

AI and Ethics

John E. Savage

Brown University
Outline

• Birth of Artificial Intelligence
• Development of AI technologies
• AI Winter and Renewal
• Challenges of autonomous machines
• Social impacts of AI
• Ethical dimensions introduced by AI
• Ethics in the news
What is Artificial Intelligence?

• Software designed to exhibit intelligence
  – Intelligent machines long sought, e.g. Frankenstein

• What is intelligence?
  – Is it symbol manipulation?
  – Does it involve creativity?

• 1956 Dartmouth AI conference a landmark event
  – Conference report entitled *Automata Studies*
‘56 Dartmouth Workshop Topics*

1. Automatic Computers
2. How Can a Computer be Programmed to Use a Language
3. Neuron Nets
4. Theory of the Size of a Calculation
5. Self-Improvement
6. Abstractions
7. Randomness and Creativity

* See [http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html](http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html) for the fascinating proposal written by John McCarthy for the two-month summer research program on AI held in 1956.

© JE Savage
Defining Automata

• Finite state machine (FSM)
  – Memory stores fixed no. bits
  – Logic circuit computes state from previous state & input

• Turing Machine
  – Control unit is an FSM
  – FSM reads cell contents
  – Makes state transition
  – Replaces cell contents
  – Moves head left or right
Computing with Automata

• Finite state machine (FSM)
  – FSM in fixed start state
  – Some states labeled accept
  – Input accepted if FSM lands in accept state after last input

• Turing Machine
  – String written on tape
  – Head over first cell
  – If FSM enters halt state, string on the tape is output
  – Most powerful computer!
Automata Studies Paper Topics*
C.E. Shannon, J. McCarthy, Editors

• Finite Automata
  – Nerve nets, robots, logic gates, black-box analysis

• Turing Machines (TMs)
  – Universal TMs with 2 states, universal TMs, probabilistic TMs, inversion of functions

• Synthesis of Automata
  – Intelligence amplifier, conditional probability machines, epistemology of automata

• [Amazon link](https://www.amazon.com/Automata-Studies-Annals-Mathematics-Studies/dp/0691079161)
Historical Developments of AI

• LISP programming language designed for AI
  – Introduced by John McCarthy in 1958
  – For processing lists, based on Lambda Calculus
  – 2nd oldest high-level programming language
  – Fortran is the oldest

• Very expressive language
  – It can describe strings accepted by Turing machines
  – Well suited to logic and knowledge representation
Historical Developments of AI*

• ‘58 Predictions led to massive AI research funding
  – Chess playing machines forecasted by 1968!
  – AI said would match human intelligence in 25 years!
• 1973 – hype not achieved, funding gone – AI winter
• 1980 – Japanese AI initiative restarts funding
• Late 1980s – Funding drops again
• Early 2000s – Deep learning resuscitates field
• Today, revolutions are again predicted
• Are these claims believable?

History of AI Advances

• Problem solving (i.e. games) via search with backtracking
  – Unfortunately, the combinatorial explosion can’t be overcome

• Natural language understanding
  – Deducing meaning is very hard, but Eliza is a sensation
    http://www.manifestation.com/neurotoys/eliza.php3

• Micro-worlds
  – Small world permits shocking human-robot communication

• Robotics
  – Simple humanoid robots appear
History of AI Advances

• Perceptron, early neural net, introduced
  – One layer of neurons
  – Nice idea, but extremely limited
• Deductive systems of logic introduced
  – OK on small problems but proofs very time consuming
History of AI Advances

• Expert systems
  – Rule based, e.g. if-then-else, on limited domain
  – First-order logic based, e.g. Prolog introduced in ‘72
    • Very successful, e.g. diagnosing infectious diseases

• AI needs massive amounts of knowledge
  – Experiments demonstrate this

• 1980 Japanese Fifth-Generation Project
  – Nations challenged
History of AI Advances

• Intelligent agents
  – Interact with their environment, e.g. robots*

• Probability & decision theory absorbed into AI
  – AI becomes more rigorous

• Deep learning – neural nets re-emerge
  – Multiple hidden layers
  – Backpropagation – learns by adjusting weights
  – Speech recognition
  – Language translation

* [https://www.youtube.com/watch?v=rVlhMGQgDkY](https://www.youtube.com/watch?v=rVlhMGQgDkY)
Neural Networks

- Nodes values are integers, edges have weights
- Values multiplied by weights, passed through non-linear activation function, giving integer values
- Weights adjusted to improve recognition
  - Weight changes made via backpropagation of errors
Generative Adversarial Networks (GANs)

- GANs are pairs of competing neural nets*
  - One net generates examples
  - Second net evaluates the examples
- Competition drives nets to improve
  - E.g. counterfeiters vs police
- GANs were invented by Ian Goodfellow in 2014 to make machine-learning systems smarter
  - GANs are very successful - AI is very powerful
- [https://deeplearning4j.org/generative-adversarial-network](https://deeplearning4j.org/generative-adversarial-network)
The Power of AI

• Roomba
  – Cleaning robot operates autonomously

• Autonomous vehicles are now being tested!
  – Reduced highway deaths are predicted
  – But several accidents resulted in death of driver

• Lethal autonomous weapons (LAWs) studied
  – Autonomous military robot
  – Ability to select and attack targets
The Weakness of AI*

- Adversarial attacks
  - Manipulating ML system with specially crafted inputs
- Small stickers trick Tesla – veers into wrong lane
  - Published by Tencent Keen Security Lab, March 2019
- UCB prof trains vision system on stickers\(^{^\wedge}\)
  - Stickers cause it to see Stop sign as 45-MPH sign
- Adversarial attack on medical machine learning\(^{\dagger}\)
  - See next slide

\(^{\dagger}\) [https://science.sciencemag.org/content/363/6433/1287](https://science.sciencemag.org/content/363/6433/1287)
Adversarial attacks on medical machine learning*

Original image: Dermatoscopic image of a benign melanocytic nevus, along with the diagnostic probability computed by a deep neural network.

Adversarial noise: Perturbation computed by a common adversarial attack technique. See (7) for details.

Adversarial example: Combined image of nevus and attack perturbation and the diagnostic probabilities from the same deep neural network.

Diagnosis: Benign

The patient has a history of back pain and chronic alcohol abuse and more recently has been seen in several...

Opioid abuse risk: High

277.7 Metabolic syndrome
429.9 Heart disease, unspecified
278.00 Obesity, unspecified

Adversarial coding (13)

Reimbursement: Denied

Diagnosis: Malignant

The patient has a history of lumbago and chronic alcohol dependence and more recently has been seen in several...

Opioid abuse risk: Low

401.0 Benign essential hypertension
272.0 Hypercholesterolemia
272.2 Hyperglyceridemia
429.9 Heart disease, unspecified
278.00 Obesity, unspecified

Adversarial text substitution (9)

Reimbursement: Approved

* [https://science.sciencemag.org/content/363/6433/1287](https://science.sciencemag.org/content/363/6433/1287)
Isaac Asimov’s Rules for Robots

• A robot may not injure a human being or, through inaction, allow a human being to come to harm.

• A robot must obey orders given it by human beings except where such orders would conflict with the First Law.

• A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.
UNESCO Precautionary Principle

When human activities may lead to morally unacceptable harm that is scientifically plausible but uncertain, actions shall be taken to avoid or diminish that harm. Morally unacceptable harm refers to harm to humans or the environment that is

– Threatening to human life or health, or
– Serious and effectively irreversibly, or
– Inequitable to present or future generations, or
– Imposed without adequate consideration of the human rights of those affected.

Impacts of AI

• AI Fairness and Safety
  – Old training sets can build in biases
  – Humanity can be endangered if robots not constrained
  – What restrictions should be imposed on designers?

• Employment*
  – Pessimists:
    • McKinsey: Half of today’s jobs will be automated by 2055
  – Optimists:
    • Gartner: AI will create > 500,000 jobs by 2020

* https://www.forbes.com/sites/danielmarlin/2018/01/16/millennials-this-is-how-artificial-intelligence-will-impact-your-job-for-better-and-worse/
AI Ethics

• Use of biometric data is growing
  – Facial recognition widely used in China

• China is assigning a social score to each citizen
  – Points won for aiding elders, biking to work
  – Points lost for violations, e.g. jay-walking
  – High scores benefit citizens
  – Low scores penalize them

• Robotic surgery
  – Mistakes on humans can be very costly
AI Ethics*

• Do technologists need ethical watchdogs?
• Questionable machine learning applications:
  – Stanford software estimates sexual orientation
    • Goal: To protect gay people but LGBT community upset
  – Stony Brook app estimates ethnicity from photos
• Ethical guidelines need to be updated for AI
  – Universities have institutional review board (IRBs)
  – But criteria don’t include big data or social impacts

* https://www.wired.com/story/ai-research-is-in-desperate-need-of-an-ethical-watchdog/
AI Ethics News*

• At 2017 Neural Information Processing Systems
• Kate Crawford’s keynote cited photo recognizers
  – Google service labeled some black people as gorillas
  – UVA software associated kitchen photos w. women
• Victoria Krakovna of Future of Life Institute
  – Assembled master list of unintended AI behaviors†

* https://www.wired.com/story/artificial-intelligence-seeks-an-ethical-conscience/
† https://vkrakovna.wordpress.com/author/vkrakovna/
Ethics in the News

• Volkswagen Official Gets 7-Year Term in Diesel-Emissions Cheating, NYT, 12/6/17
• IEEE has a Global Initiative on Ethics of Autonomous and Intelligent Systems
• ACM drafting Code of Ethics & Professional Conduct
• France investigates printer companies for planned obsolescence, NETWORKWORLD, 1/5/18
• Google Employees Protest Work for the Pentagon, NYT, 4/4/18
Ethics in the News

• US vs Microsoft (2001)
  – Justice Department sued MSFT for monopoly & antitrust practices, bundling of Explorer in its OS
  – Case settled; MSFT agreed to share its APIs

• Facebook-Cambridge Analytica (CA) data scandal
  – A. Kogan, Cambridge U., provided app that collected Facebook data on ≥ 87 million users for CA in 2014
  – Data used for electoral campaigns by Senator Cruz in 2015 and Trump in 2016 and Brexit 2016 campaign
Review

- Birth of Artificial Intelligence
- Development of AI technologies
- AI Winter and Renewal
- Challenges of autonomous machines
- Social impacts of AI
- Ethical dimensions introduced by AI
- Ethics in the news