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?

- A man-made system exhibiting intelligence
- Humans invented automata, self-operating machines, thousands of years ago.
  - Greeks: moving statues, roaring lions, chirping birds
  - In 1206 Al Jazari's "band" performed "more than 50 facial and body actions during musical selection"
  - Other examples: music boxes, cuckoo clocks, Disney animatronics (mechanical + electronic automata)
- Visit https://themadmuseum.co.uk/history-of-automata/

Examples of Ancient Automata

- The Morris Museum, Morristown, NJ holds automata collected by William Guiness
- Watch videos
  - Overview of the collection (2:07)
    - https://www.youtube.com/watch?v=OK1X-_RAA44
  - 19th Century Life-sized floutist (2:11)
    - https://www.youtube.com/watch?v=1TxrjpWGRXU

What is Modern Artificial Intelligence?

- Software designed to exhibit intelligence
- What is intelligence?
  - Is it symbol manipulation?
  - Does it involve creativity?
  - How would you def?ine it
- 1956 Dartmouth AI conference – landmark event
  - Conference report entitled Automata Studies

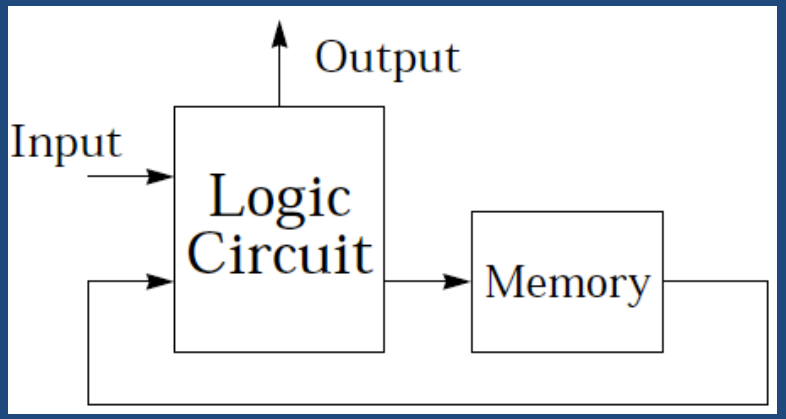
‘56 Dartmouth Workshop Topics*

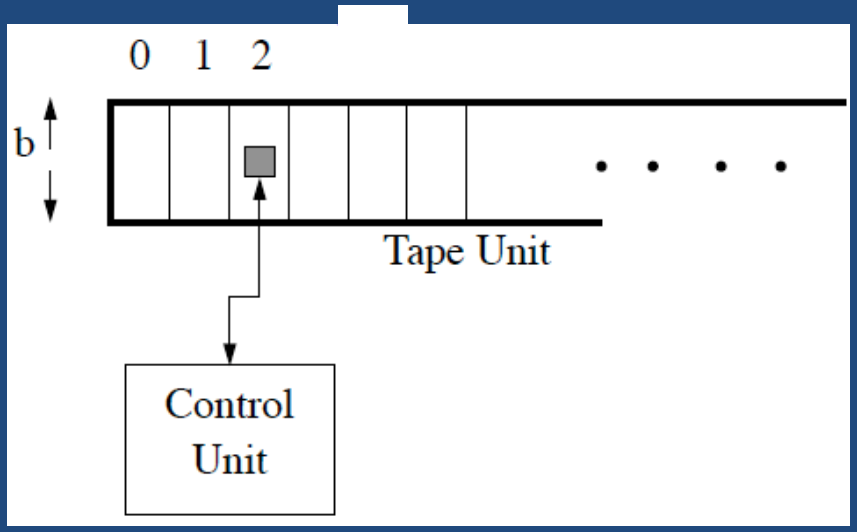
1. Automatic Computers
2. How Can a Computer be Programmed to Use a Language
3. Neuron Nets
4. Theory of the Size of a Calculation
5. Self-Improvement
6. Abstractions
7. Randomness and Creativity

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

Structure of Modern Automata

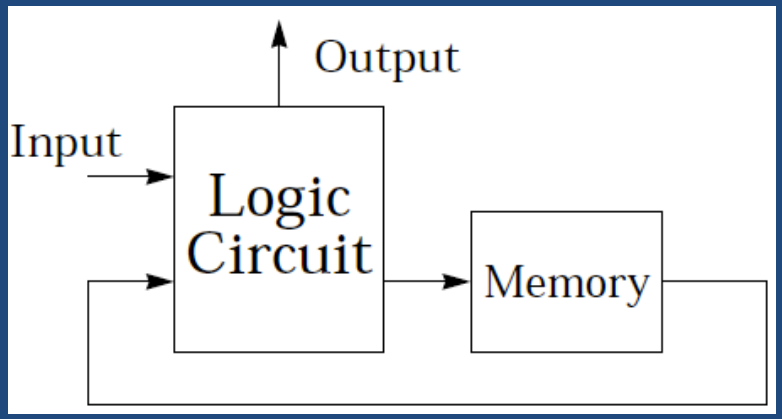
- Finite state machine (FSM)
  - Memory stores fixed no. bits
  - Logic circuit computes state from previous state & input
  - It models current computers
- (Abstract) Turing Machine (TM)
  - Control unit is an FSM
  - Has access to infinite tape
  - FSM reads cell contents
  - Makes state transition
  - Replaces cell contents
  - Moves head left or right

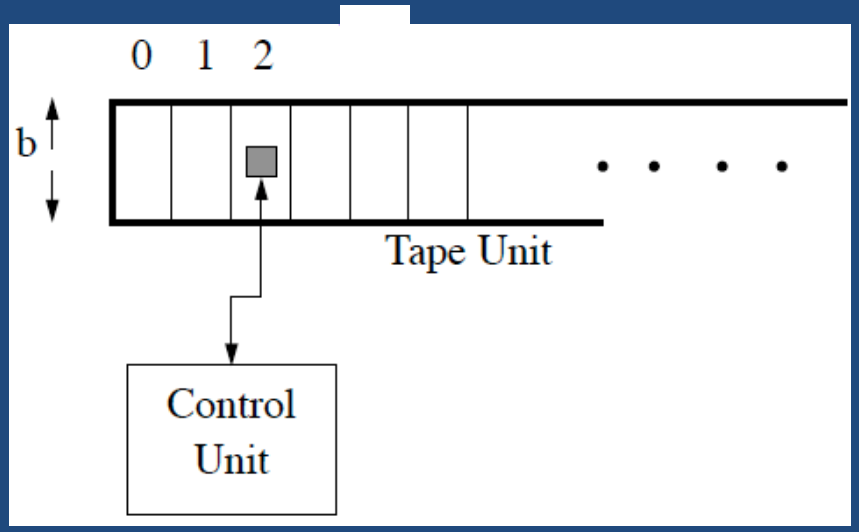




Computing with Modern Automata

- FSM has fixed initial state
  - It maps inputs to outputs
- Turing Machine
  - String written on tape
  - Head over first cell
  - If FSM enters halt state, string on the tape is output
  - Most powerful computer for general computations





Automata Studies Paper Topics*
C.E. Shannon, J. McCarthy, Editors

- Topics at 1956 conference show state of art:
- Finite Automata
  - Nerve nets, robots, logic gates, black-box analysis
- Turing Machines (TMs)
  - A universal TM (UTM) can simulate an arbitrary TM, there is a UTM with two states, probabilistic TMs.
- Synthesis of Automata
  - Intelligence amplifier, conditional probability machines, epistemology of automata
- https://www.amazon.com/Automata-Studies-Annals-Mathematics-Studies/dp/0691079161

Invention of LISP – A Language for AI

- LISP programming language designed for AI
  - Introduced by John McCarthy in 1958
  - Good for processing lists
  - 2nd oldest high-level programming language after Fortran
- Very expressive language
  - It can describe strings computed by Turing machines
  - Well suited to logic and knowledge representation

Historical Developments of AI*

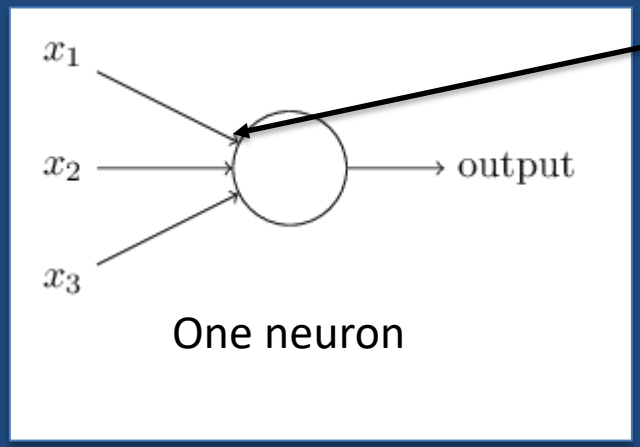
- '58 Predictions led to massive AI research funding
  - Chess playing machines forecasted by 1968!
  - AI researchers: will match human intelligence in 25 years!
- 1973 – failed! Funding ends and AI winter begins
- 1980 – Japanese launch new AI initiative, US restarts funding
- Late 1980s – Funding drops again
- Early 2000s – Deep learning resuscitates field
- Today, revolutions are again predicted
- Are these claims believable?
- https://en.wikipedia.org/wiki/History_of_artificial_intelligence

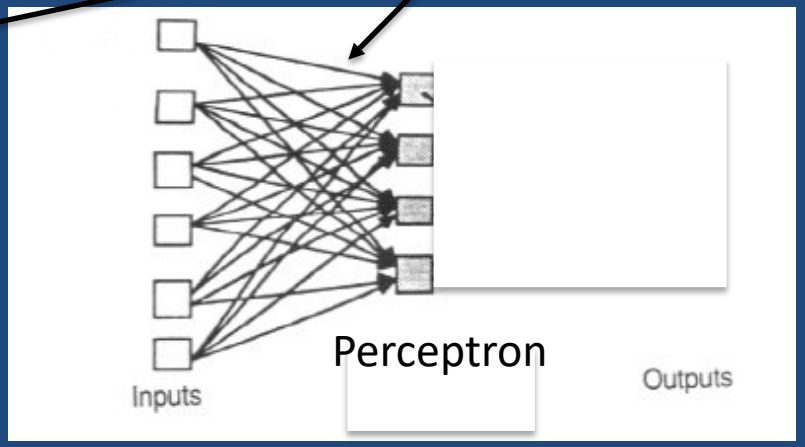
History of AI Advances

- Problem solving (i.e. games) using search and backtracking
  - Unfortunately, the combinatorial explosion can't be overcome
- Natural language understanding
  - Deducing meaning is very hard, LAUNCH Eliza – sensational http://www.manifestation.com/neurotoys/eliza.php3 (Try it!)
- Micro-worlds
  - Small world simulates human-robot communication (Shocking)
  - Blocks World and SHRDLU https://en.wikipedia.org/wiki/SHRDLU
- 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

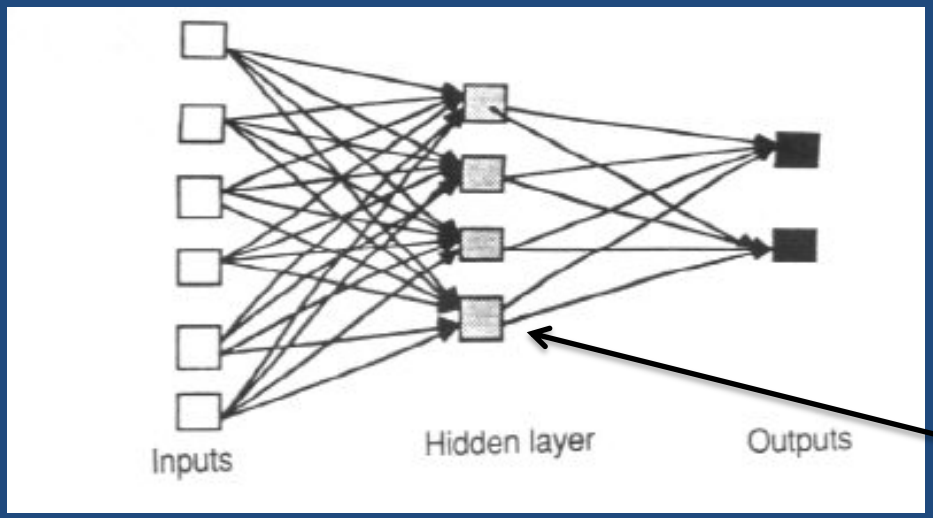
- Expert systems
  - Rule based, e.g. if-then-else, on limited domain
  - First-order logic based, e.g. Prolog introduced in ‘72
    - Very successful, e.g. diagnosing infectious diseases
- Experiments show that AI needs massive amounts of knowledge for training!
- 1980 Japanese Fifth-Generation Project challenged the West

History of AI Advances

- Intelligent agents
  - Interact with environment
  - WATCH Boston Dynamics robot video (2:41)*
- 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

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 both nets to improve
  - Like competition between counterfeiters and police
- GANs invented by Ian Goodfellow in 2014* - most interesting machine-learning (ML) idea in 10 yrs
- GANs are very successful - AI is now very powerful

* https://towardsdatascience.com/understanding-generative-adversarial-networks-gans-cd6e4651a29

Examples of Modern AI

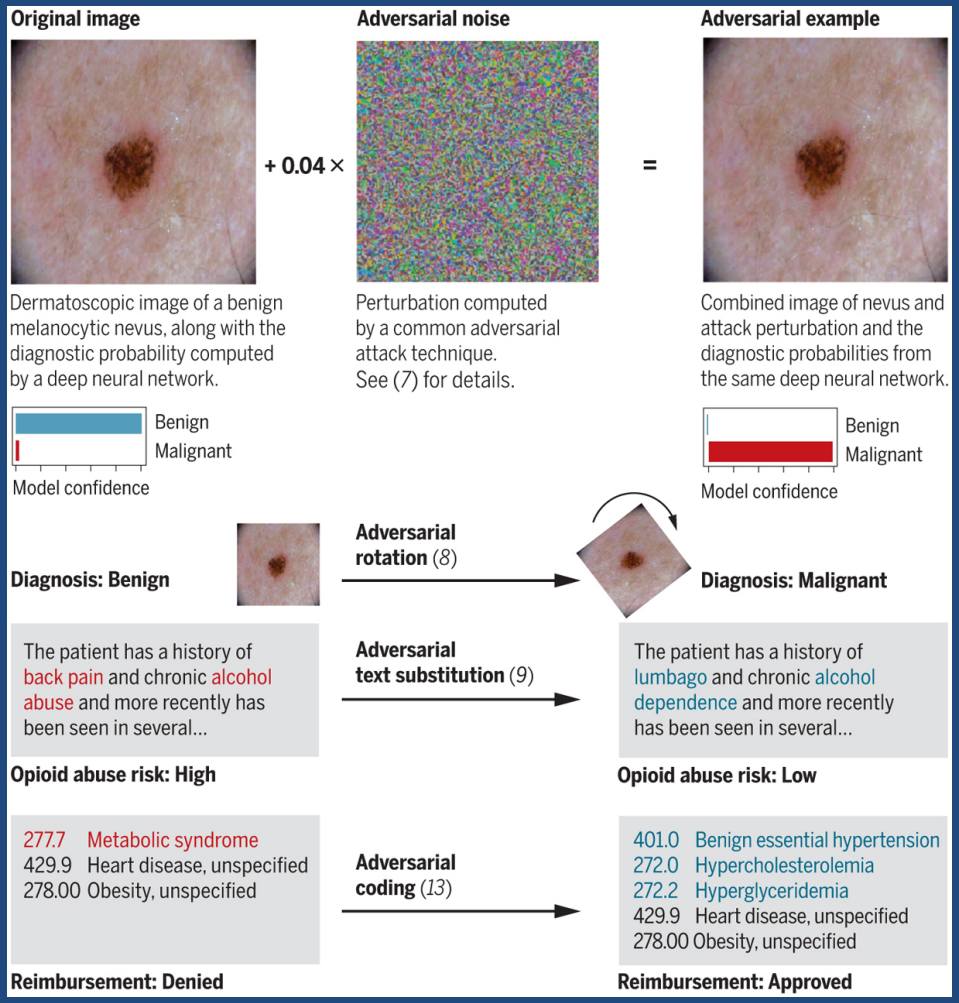
- Roomba
  - Cleaning robot operates autonomously
- Autonomous vehicles are now being tested!
  - Reduced highway deaths are predicted
  - But several people have been killed
- Lethal autonomous weapons (LAWs)
  - Autonomous military robot
  - Ability to select and attack targets
  - Source of major concern at UN!

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 street signs^
  - Stickers cause it to see Stop sign as 45-MPH sign
- Visit URLs for fascinating reports on failures of AI

* https://www.technologyreview.com/the-download/613254/hackers-trick-teslas-autopilot-into-veering-towards-oncoming-traffic/
† https://science.sciencemag.org/content/363/6433/1287
^ https://arxiv.org/pdf/1707.08945.pdf

Adversarial attacks on medical machine learning*



* 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 irreversible, or
- Inequitable to present or future generations, or
- Imposed without adequate consideration of the human rights of those affected.

* http://unesdoc.unesco.org/images/0013/001395/139578e.pdf

Impacts of AI

- AI Fairness and Safety
  - Old training sets can incorporate biases into modern sys
  - 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

- See URL*on need for ethical AI watchdogs?
- Questionable machine learning applications:
  - Stanford software estimated sexual orientation
    - Goal: To protect gay people but LGBT community upset
  - Stony Brook app estimated 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