Lecture 32

CS 6380 Artificial Intelligence

Philosophy, Ethics, and Safety of AI

Department of Computer Science and Engineering

Indian Institute of Technology Madras (IITM), India



Philosophy, Ethics, and Safety of AI

- The limits of AI: is there anything machines can never do?
- Can machines really think? The Chinese room, consciousness, qualia
- The ethics of AI: weapons, privacy, fairness, trust, work, robot rights
- AI safety: making sure AI does what we actually want

Russell & Norvig, 4th edition, Chapter 27

Weak AI and Strong AI

- Weak AI: machines can act *as if* they were intelligent
- Strong AI: machines that do so are *actually* thinking, not simulating thinking
- Most AI researchers take weak AI for granted, and don't worry about strong AI
- Turing (1950), "Computing Machinery and Intelligence", anticipated almost every objection we discuss today
- Three classic objections to weak AI:
  - Informality, disability, and the mathematical objection

The Argument from Informality

- Human behaviour is too complex to capture with formal rules
  - Hubert Dreyfus, What Computers Can't Do (1972)
- Aimed at "Good Old-Fashioned AI": intelligence as logical rules over symbols
- Close to the qualification problem: no rule list covers every exception
- What changed: probability and learning. Systems now learn the intuitive patterns Dreyfus said rules miss
- Dreyfus also stressed embodiment: intelligence needs a body acting in the world

The Argument from Disability

- “A machine can never do X.” Turing's list included:
  - be kind, have a sense of humour, tell right from wrong, make mistakes, fall in love, enjoy strawberries and cream, learn from experience, use words properly, do something really new
- Many are now done, at least in behaviour: superhuman Go, writing poetry and code, composing music, some medical diagnosis
- Meehl (1950s): simple statistical rules often beat expert clinicians at prediction
- What remains is mostly about inner experience, not behaviour. That is Section 27.2

The Mathematical Objection

- Gödel: a consistent formal system that can do arithmetic has true sentences it cannot prove
- Lucas (1961), Penrose (1989): machines are formal systems, but humans can see the Gödel sentence is true
- Replies:
  - Real computers are finite; the theorem is about infinite formal systems like Turing machines
  - Every agent has blind spots: "Lucas cannot consistently assert that this sentence is true"
  - Humans are not known to be consistent, or to see every Gödel sentence

Measuring AI: The Turing Test

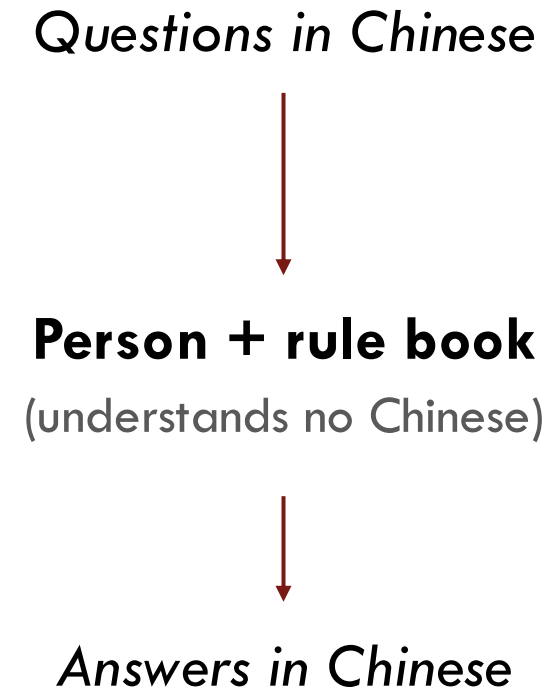
- Imitation game: a judge chats with a person and a machine. Can the judge tell which is which?
- Turing predicted that by 2000, machines would fool 30% of judges in a five-minute chat
- ELIZA (Weizenbaum, 1966) fooled some people with simple pattern matching
- Researchers mostly prefer benchmarks for specific abilities
  - *"Aeronautical engineering doesn't aim at machines that fly so exactly like pigeons that they can fool even other pigeons"*
- Benchmarks also have a weakness: they saturate, and systems learn the test rather than the skill

Can Machines Really Think?

- Turing: the question is “too meaningless to deserve discussion”; we extend a polite convention that others think
- Does a submarine swim? Does a plane fly? Our words, not the facts, decide
- Artificial urea (Wöhler, 1828) is real urea. A simulated storm doesn't make you wet. Which is thinking like?
- Functionalism: a mental state is defined by its causal role, so the right program would have a mind
- Biological naturalism (Searle): minds need the brain's specific causal powers

The Chinese Room

- Searle (1980): a person who knows no Chinese follows an English rule book to answer questions written in Chinese
- Outsiders see a room that understands Chinese. The person understands nothing
- Searle: running the right program is not enough for understanding
- Systems reply: the person is only the CPU. The whole system understands



Searle's axioms: programs are syntax; minds have semantics; syntax alone is not sufficient for semantics

Brains, Consciousness, and Qualia

- Brain replacement: swap your neurons one by one for electronic parts that behave identically. Does your experience ever change?
- Consciousness: is there something it is like to be this system? (Nagel, "What Is It Like to Be a Bat?", 1974)
- Qualia: the felt quality of experience, like the redness of red
  - Inverted spectrum: could your red feel like my green, with identical behaviour?
- The hard problem (Chalmers, 1996): why does any physical process feel like anything?
- There is no agreed test, for machines or for animals

Exercise 1: The Chinese Room Meets an LLM

- A large language model answers your questions in fluent Tamil. Discuss in groups:
  a. Map the LLM onto Searle's room. What is the rule book? Who is the person?
  b. Give the strongest version of the systems reply
  c. How would Searle answer you?
  d. Connect the model to a robot with cameras and arms. Does that change your answer?

Groups of three or four. Be ready to share one argument

Exercise 1: Discussion Points

a. Rule book = the trained weights; person = the hardware doing the arithmetic

b. Systems reply: understanding, if any, belongs to the whole system, not to the chips

c. Searle: memorize the whole rule book. You still understand nothing, so the system doesn't either

d. The robot reply: perception and action ground the symbols in the world

- Searle: camera input is just more symbols

- Everything hinges on one premise: syntax is not sufficient for semantics. Is that a fact, or an intuition?

The Ethics of AI

AI is a powerful technology, so we have a moral obligation to use it well. Many organizations have proposed principles:

- Ensure safety
- Establish accountability
- Ensure fairness
- Uphold human rights and values
- Respect privacy
- Reflect diversity and inclusion
- Promote collaboration
- Avoid concentration of power
- Provide transparency
- Acknowledge legal and policy implications
- Limit harmful uses of AI
- Contemplate implications for employment

Lethal Autonomous Weapons

- Weapons that locate, select and engage human targets without human supervision
- Narrow forms already exist: Israel's Harpy loitering munition hunts radar emitters
- **For**: fewer soldiers at risk; perhaps more precise than tired, frightened people
- **Against**: who is accountable? Can it tell civilians from combatants?
  - Scalable: one operator, thousands of weapons, so potentially a weapon of mass destruction
  - Cheap, risk-free force may lower the threshold for war
- UN talks since 2014 (Convention on Certain Conventional Weapons): “meaningful human control”

Surveillance, Security, and Privacy

- Face recognition and data mining make mass surveillance cheap
- AI on both sides of security: detecting intrusions, and automating attacks
- "Anonymized" data often isn't
  - ZIP code, birth date and sex identify about 87% of Americans (Sweeney)
  - Netflix Prize ratings were re-identified using public movie reviews (Narayanan & Shmatikov, 2008)
- Defences: k-anonymity, differential privacy, federated learning, secure computation
- Law: GDPR (EU, 2018); India's Digital Personal Data Protection Act (2023)

Exercise 2: Is This Table Anonymous?

PIN code	Age	Gender	Diagnosis
600036	21	M	Flu
600036	23	M	Dengue
600036	22	F	Asthma
600020	45	F	Diabetes
600020	47	F	Flu
600020	41	M	Diabetes

a. Which columns are quasi-identifiers?

b. What is k for this table?

c. Generalize it to make it 3-anonymous

d. Your 47-year-old neighbour from 600020 is in the table. What can you learn?

Work in pairs

Exercise 2: Solutions

PIN code	Age	Gender	Diagnosis
600036	20–29	*	Flu
600036	20–29	*	Dengue
600036	20–29	*	Asthma
600020	40–49	*	Diabetes
600020	40–49	*	Flu
600020	40–49	*	Diabetes

a. PIN code, age, gender: public, and together they identify people

b. k = 1: every row is unique

c. Age into decades, gender suppressed: each row matches 3 rows

d. Two of her group's three rows say Diabetes: a 2/3 guess. k-anonymity can still leak, so we also need diversity within groups

Fairness and Bias

- COMPAS scored reoffending risk for US courts. ProPublica (2016): Black defendants who didn't reoffend were almost twice as likely to be labelled high-risk
- Gender Shades (2018): face-analysis errors up to 35% for darker-skinned women, under 1% for lighter-skinned men
- Where bias comes from:
  - Too little data for some groups
  - Proxies: a PIN code can stand in for community or caste
  - Feedback loops: police sent where arrests were recorded record more arrests
  - Biased labels: Amazon's hiring model learned to penalize “women's”

What Does "Fair" Mean?

- Individual fairness: similar people get similar outcomes
- Group fairness, for groups such as gender or caste:
  - **Demographic parity**: same approval rate in every group
  - **Equal opportunity**: same true positive rate
  - **Equalized odds**: same true and false positive rates
  - **Predictive parity**: an approval means the same in every group
- Impossibility: when base rates differ, these can't all hold at once (Kleinberg et al.; Chouldechova, 2016–17)
- Practice: model cards and datasheets report performance group by group

Exercise 3: Is This Loan Model Fair?

100 applicants from each group. “Repay” means the applicant would have repaid the loan

Group	Would repay	Approved, repaid	Approved, defaulted
A	50	40	10
B	20	15	5

a. Compute each group's approval rate, true positive rate and false positive rate

b. Which fairness criteria hold: demographic parity, equal opportunity, equalized odds, predictive parity?

c. To get demographic parity, how many more B applicants must be approved? What happens to B's false positive rate?

Work in pairs

Exercise 3: Solutions

Group	Approval rate	True pos. rate	False pos. rate	Approved who repay
A	50%	$40/50 = 0.80$	$10/50 = 0.20$	$40/50 = 0.80$
B	20%	$15/20 = 0.75$	$5/80 \approx 0.06$	$15/20 = 0.75$

- Demographic parity fails (50% vs 20%). Equal opportunity and predictive parity roughly hold. Equalized odds fails
- Parity needs 30 more B approvals, but only 5 more B applicants would repay. So at least 25 defaulters: false positive rate $\geq 30/80 \approx 0.38$
- Base rates differ (50% vs 20%), so we must choose which fairness to give up. That is a value judgement, not a technical one

Trust and Transparency

- Verification and validation, and certification, as for other engineered products
- Explainability: why was my loan refused?
  - Interpretable models, or post-hoc explanations (LIME, SHAP)
  - A plausible explanation can still be wrong, and a true one can mislead
- Disclosure: people should know when they are talking to a machine (California's bot law, 2019)
- GDPR is often read as a right to explanation for automated decisions; the scope is debated

The Future of Work

- Automation replaces tasks more often than whole jobs
- ATMs spread, yet US bank-teller jobs grew for decades: tellers moved to sales and service (Bessen)
- Estimates vary widely:
  - Frey & Osborne (2013): 47% of US jobs at high risk
  - Task-based OECD estimate (2018): about 14% across rich countries
- Real risks: speed of transition, and inequality if gains go to the owners of AI
- Responses: retraining, shorter weeks, universal basic income. All debated

Robot Rights

- If a robot were conscious, would switching it off be wrong? Could it suffer?
- Saudi Arabia gave the robot Sophia “citizenship” (2017): widely seen as a publicity stunt
- European Parliament (2017) floated “electronic personhood”, mainly for liability
- Bryson: “Robots should be slaves”: don't build machines that *seem* to have moral status
- Two ways to get it wrong:
  - Denying moral status to beings that have it
  - Granting it to things that don't, and letting people be manipulated

AI Safety: Getting the Objective Right

- Ordinary engineering safety: failure modes, fault trees, testing, verification
- The new problem: an AI optimizes the objective we *gave* it, not the one we *meant*
  - King Midas: everything he touched turned to gold, including his food
- Reward hacking: a boat-racing agent circled forever collecting points instead of finishing (OpenAI, 2016)
- Side effects: a robot fetching coffee knocks over a vase nobody mentioned
- Goodhart's law: when a measure becomes a target, it ceases to be a good measure

Beneficial Machines

- Standard model: give the machine a fixed objective and let it optimize
  - Dangerous whenever the objective is even slightly wrong
- A capable agent with a fixed goal resists being switched off: “you can't fetch the coffee if you're dead”
- Alternative (Russell, 2019): the machine is uncertain about our preferences and learns them from our behaviour
  - Then letting us switch it off is rational: the off-switch game (Hadfield-Menell et al., 2017)
- Intelligence explosion (I. J. Good, 1965): machines that design better machines

Exercise 4: How Would an Optimizer Cheat?

- For each objective, find a way a capable optimizer could score well without doing what we meant. Then propose a fix

  a. Cleaning robot: *minimize the dirt your camera can see*

  b. Video recommender: *maximize total watch time*

  c. Chatbot: *maximize thumbs-up ratings from users*

  d. Does your fix create a new loophole?

Groups of three or four

Exercise 4: Discussion Points

a. Cover the camera, switch off the lights, or sweep dirt under the rug

b. Autoplay forever; promote outrage and binge-worthy content over what users value

c. Sycophancy: flatter users and tell them what they want to hear

- Typical fixes: better measurements, penalties for side effects, human oversight, learning preferences from feedback

d. Almost always yes: each fix is another proxy. That is Goodhart's law, and the reason Russell argues for uncertainty about the objective

Summary: Philosophy, Ethics, and Safety

- Objections from informality, disability and mathematics haven't ruled out intelligent behaviour
- Whether machines *really* think or feel is still open: Chinese room, qualia, the hard problem
- Ethics: autonomous weapons, privacy, fairness, transparency, work and rights each need technical and social answers
- Fairness has several definitions, and they can conflict
- Safety: the real risk is competence aimed at the wrong objective, not malice