Lecture 6: Sociological foundations (contd.)

Sruti S Ragavan

So far…

- HAX guidelines & their underlying principles
- Cognitive psychology & Sociology
- Cognitive psychology → how humans perceive, pay attention, remember, make decisions, solve problems, experience emotions
    - Foundations for humans interacting with the world (also interfaces)
- Sociology → study of groups of people (culture, society, interactions within groups)
    - Computers are social actors (even when they were "mechanistic")
    - With computers being closer to humans → expectations evolve
    - Important for Human-AI interaction than in classical HCI

So far…

- Symbolic interactionism → agent’s cues, behaviors, symbols, etc.
    - Make clear what the system does & how well it can do so.
- Role → behaviours, rights, obligations attached to a social position.
    - What role it plays, what about multiple roles with conflicts?
- Socialization & Social interactions
    - Over time, norms change around AI/humans (e.g., use of driverless cars)
    - Social interactions → appropriateness, trust, fairness, bias, etc.

Deviance and Social Control

- Accepted behaviours → conformity → social order
- Deviation from this behavior → penalty (social and legal)
  - Deterrent!
- In general, people confirming to social norms keeps order.
  - Stop a red, wait for pedestrian, board a bus after alighters get off.
  - We <u>trust</u> others to respect the same norms!
- What if someone deviates?
  - Driver injures a pedestrian at a red → fines; what if driverless car?
  - Tutor gives wrong advice → fired from job; what about AI tutors?
- Important for developers & long-term policy / regulations

Society is imperfect!

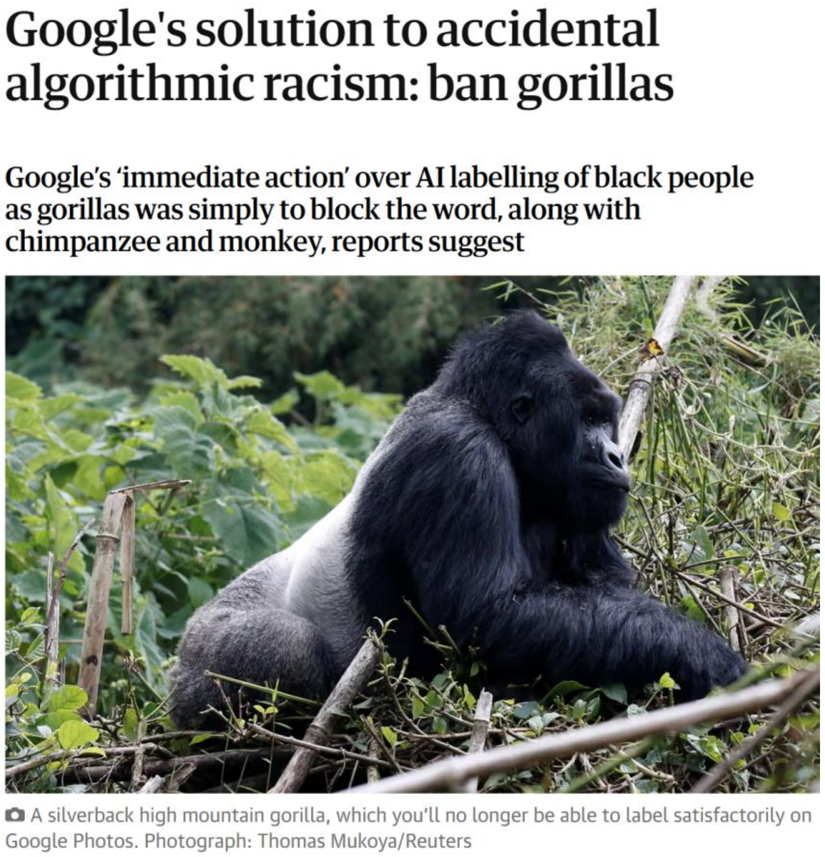
- Diverse and unequal
  - All members of a group are not equal
  - All groups are not equal
  - Some have greater power, control over what happens in the group/society.
  - Some have better representation, greater access to resources, etc.
  - People have favorites / preferences / alignments
  - People are also biased against some groups / individiuals
    - Treat them "unfairly" (not as equal with others)
    - Discrimination – gender, age, race, caste, skin color, academic disciplines, ...
- Can AI act the same way, since, it is just like people?
- People are sensitive to these issues → should AI be too?
- [We don't know answers, yet!]

Fairness

- In some contexts, expectation that no discrimination be made
  - Between members of a group / between groups
  - Jobs, criminal justice, franchise, …
- With AI used in areas such as law, loans, etc.
  - Justice → offer what each individual actually deserves
- AI unfortunately is not there yet!
  - Models discriminate – especially against some groups (than in favor of others)
  - In other words, models are biased → lean away from one side!

Biases: Representativeness

- In society, some groups are invisible
  - Not invite women/grad/HSS students to an event
  - Less visible groups → exclusion → even less visible
- Happens in AI too!
  - Selection / sampling biases in training / test data
  - During collection, labeling
  - Easily avoided!
- Consciously look for diversity in data & labels



Google's solution to accidental algorithmic racism: ban gorillas

Google's 'immediate action' over AI labelling of black people as gorillas was simply to block the word, along with chimpanzee and monkey, reports suggest

A silverback high mountain gorilla, which you'll no longer be able to label satisfactorily on Google Photos. Photograph: Thomas Mukoya/Reuters

Biases: Stereotypes

- Data reflects society’s stereotypes
- Generative AI is trained on content that reflects biases
  - “Girls like pink”, “Men are seldom teachers and nurses”, “chairman / watchman/postman”
- Black men sentenced unfairly, men treated unfairly in gender violence cases, women not hired historically in various roles
- Data from biased system → training data → biased models
- Hard to fix, but needs other forms of labelling / roles
  - Not as decision makers, but make decisions based on quality clusters

Transparency, trustworthiness, accountability

- Models be transparent → tell me why you made this decision
  - Decisions made by courts of law owe explanations to citizens and the conflicted parties
  - RTI, appeals, audits happen!

- Models should be subject to similar standards
  - Transparent in their data, decision-making; black boxes are not an option!
  - Make sources, explanations, inconsistencies explicit
  - Frame the AI in appropriate roles that clarify these
  - When in doubt, models must downgrade themselves

Accountability

- Accountability is the obligation of individuals, institutions, or systems to be answerable for their actions, decisions, and outcomes
- In the case of AI:
  - It should be answerable for its decisions.
  - When wrong → who goes to court?
  - Developer? Commissioning organization? User?
- Developer → debias AI, make it transparent
- Organizations → Make roles of everyone clear (person, AI, etc.), audit AI
- Regulation needs to catch up (still hasn’t!).

Ethical AI

- AI consistent with human values and social ethics
- Fairness
- Privacy and security [Guest lecture next Friday!]
- Reliability
- Appropriate usage (doesn’t harm others) / break laws
    - Plagiarizing AI?
- Inclusive

TRUSTWORTHY AI

AI that is accurate, fair, accountable, transparent and ethical

Trust

- Belief that another party act in your own interest / respect a “contract”
- Willingness of one party to be “vulnerable” to another
- Makes social life predictable, make decisions under uncertainty
- There’s always a risk:
    - Trust willingness to take the risk (based / not on prior data)
    - Distrust is mitigating the risk
    - Underlies a lot of decisions we make on a daily basis!

Human trust vs. trustworthy AI

- Trust → <u>attitude</u> of a person (towards AI)
  - Informed/Calibrated or Unwarranted

- Unwarranted trust + AI failure = “betrayal”
- Calibrated trust + AI failure ≠ betrayal

- Trustworthiness → <u>attribute</u> of a model / AI
  - Ability to actually stick to a “contract”
  - Helps humans feel less betrayed by AI (even when it sometimes goes wrong)!

What makes trustworthy AI?

- Lawful, Ethical, Robust
- Seven guidelines:
    - Human agency and oversight: do not strip humans of their right / power to exercise judgment and make decisions
    - Technical robustness & safety
    - Privacy and data governance
    - Transparency
    - Diversity, non-discrimination, fairness
    - Societal and environmental wellbeing
        - Million$: Is the carbon burn worth it?
    - Accountability

Summary and looking ahead..

- Guidelines for designing human-centred AI
- The principles behind them
- Next week:
  - Guest lecture on privacy (Dr. Sharma, 29 August, tentatively)
- From Monday:
  - Actually building human-centred AI systems
  - Starting with debiasing data
  - Carry computers!

Quiz tomorrow

- Material: All, minus today’s readings
- Open notes
  - You many carry one A4-sheet worth of hand-written notes.
  - Carry stationery (incl. sketching supplies!).
  - Sheets will be provided
  - Will be for 45 mins, and then we can disperse!

Readings

- Skim:
  - https://developers.google.com/machine-learning/crash-course/fairness
  - Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies

Questions + attendance