SOCIETAL IMPLICATIONS OF AI

VECTOR INSTITUTE

INSTITUT VECTEUR

MAY 13, 2019

Introduction

Recent advances in AI largely due to growth and analysis of large data sets enabled by the internet, advances in sensory technologies and applications of deep learning

In coming years, as public encounters new AI applications in domains such as transportation and healthcare, they must be introduced in ways that build trust and understanding, and respect human and civil rights.

Not threat to our existence or plot to take over the world, in coming decades we can expect AI systems to increasingly being applied in transportation; service robots; healthcare; education; low-resource communities; public safety and security; employment and workplace; and entertainment.

Consider implications of advances in AI for jobs, social interactions, privacy and war

Brief introduction – big picture view of topic

Sources: AI Now, Ada Lovelace Foundation reports

Participatory: discuss personal experience, reaction -- ammifairnessandprivacy@gmail.com

Social Issues & Challenges

1. Labor and Automation: understanding AI and automation's impacts on labor. How it will be used as a tool of employee hiring, firing and management, including surveillance.

2. Bias and Inclusion: a growing concern wrt the design and social implications of AI decision-making systems.

3. Rights and Liberties: role of AI systems in supporting or eroding citizens' rights and liberties in areas like criminal justice, law enforcement, housing, hiring, lending

4. Ethics and Governance: history of AI research and development -- whose concerns ultimately reflected in ethics of AI, and how ethical codes and strategies could be developed

1. Impact of AI on Labor

Which sectors will AI impact most?

- those involving 'predictable physical' activities; data collection/processing

Need research on: Are robots, automated systems replacing or complementing human workers?

- some occupations displaced (taxi and truck drivers)
- others AI transforms nature of work itself
- Example: algorithmic form of management used by Uber -- digital matching market, but info asymmetric

New: AI-driven management relies on data collection, beyond workplace - fitness trackers, productivity apps, smartphones w/ monitoring

- exploit insight to increase profits, manipulate behavior
- Uber - behavioral economics – drivers set earnings targets; msg drivers about targets to keep cars on road in quiet times

1. Impact of AI on Labor

Other AI management systems provide new, invasive methods

- capture information from all employee computer activities
- aggregates info, uses AI to detect anomalous behaviors
- analyze emails for sentiment

Predictive algorithms involved in shift to 'gig economy' - more on-call work, when stores busy, predicted with AI

- new laws – require 7 days notice of upcoming schedule
- other response - Universal Basic Income

Other professions impacted include medical diagnosticians

- dermatologist sees about 200,000 cases during her lifetime
- Stanford ML algorithm ingested nearly a 130,000 cases in about three months
- each new dermatology resident starts from scratch; ML algorithm keeps ingesting, growing, and learning

But key insight not just diagnosis, understanding why (causes)

- will 'automation bias' reduce focus, learning (arithmetic, spelling, driving)?
- maintain chain of discovery that often begins in clinic?

2. Bias & Inclusion

Different meanings of terms important, e.g., bias

- popular, colloquial:

  -- judgement based on preconceived notions or prejudices, as opposed to impartial evaluation of facts

  -- different groups should be treated equally

- legal: impartiality is core value, ex/ juror selection
- ML/statistics:

  -- difference between *estimator's* expected value and true value of parameter being estimated

  -- selection bias: errors in estimation that result when some members of a population are more likely to be sampled than others

Biases in AI-driven systems:

- racial disparities in disciplinary actions and recommendations for advanced coursework
- predictive policing reinforces racially-biased police practices by recommending increased deployment in neighborhoods of color
- datasets used to train health-related AI often rely on clinical trial data, historically skewed toward white men, even when health conditions primarily affect people of color or women.

3. Rights & Liberties

Population registries

- Book of Life registry project in South Africa (1967-1983): IBM assisted South Africa in classifying population by racial descent
- National Security Entry-Exit Registration System (9/11/01): centralized documentation for non-citizens in the United States from a list of 25 predominantly Muslim countries that Bush administration deemed dangerous.

Law enforcement agencies currently integrating AI and related algorithmic decision-support systems from the private sector, e.g., police body cameras

- Possibly highly biased dataset, predictive policing
- could enhance transparency, accountability -- Voigt et al (2017), Language from police body camera footage shows racial disparities in officer respect. *PNAS*

3. Rights & Liberties

Legal system: risk assessment algorithms

- evidence-based ML can improve accuracy of statistical, actuarial methods for risk forecasting -- reduce number of defendants held in jail as they await trial
- But limited data, features, one-sided, unwarranted trust

Privacy – new frontier

- asymmetries between institutions that accumulate data, people who generate it
- shift in quality of data used for AI (DeepMind partnership with UK's National Health Service)
- expansion of AI into areas like urban planning -- deployment of IoT devices and sensors, throughout daily lives, tracking human movements, preferences and environments
- predictive privacy harms (detecting substance abusers from Facebook posts)

4. Ethics and Governance

Heated debates about whether AI systems should be used in sensitive or high-stakes contexts, who can decide, and proper degree of human involvement

Decision-making, by AI systems and people who build them, often obscured from public view and accountability

- Facebook mining user data to reveal teenagers' emotional state for advertisers; Cambridge Analytica individual profiles, fake news reportedly shifting election results
- user consent, privacy and transparency often overlooked

Education: Blue Sky Agenda for AI Education: democratization of AI education, emphasizes inclusiveness

Ethical codes (Future of Life Institute, IEEE, ACM, …): set moral precedents and start conversations

But effective invisibility of many of these systems, inescapable pervasiveness make public discourse difficult and opting-out impossible

Algorithms, Data, & AI

Much of discussion centers on impact of AI, but actually involves issues around topics beyond AI

More general theme: how to develop ethically and societally relevant technologies based on algorithms, data, and AI (ADA)

- Algorithm: unambiguous procedure for solving a given class of problems, often used to mean an automated decision-making process

- Data: encoded information about one or more target phenomena (objects, events, processes, people), relevant to ethics & society:

1. process of collecting, organizing data requires assumptions about what is significant, useful
2. digitally encoded data allows information to be duplicated, transferred, and transformed much more efficiently than ever before
3. new forms of analysis allows those possessing large amounts of data to acquire novel insights

- AI: any technology that performs tasks that might be considered intelligent
  - AI can be used to optimize processes and may be developed to operate autonomously, creating complex behaviors that go beyond what is explicitly programmed

State of Field: AI & Ethics

- Shared set of key concepts and concerns emerging, with widespread agreement on some core issues (such as bias) and values (such as fairness) that an ethics of ADA should focus on

- Agreeing on these issues & values -- an important step for ensuring that ADA-based technologies developed and used for benefit of society

- Over last two years, these have begun to be codified in various sets of 'principles', but:

1. no clarity, consensus around the meaning of central ethical concepts (privacy, bias, and explainability) and how they apply in specific situations. disciplines, cultures

2. insufficient attention given to tensions between the ways technology may both threaten and support different values

3. insufficient evidence on (a) key technological capabilities and impacts, and (b) the perspectives of different groups

Tensions Between Values

Identifying and resolving tensions between the ways technology may both threaten and support different values

Examples of tensions between values central to current applications of AI:

- Quality of services vs. privacy: using personal data may improve public services by tailoring them based on personal characteristics, but compromise privacy
- Personalization vs. solidarity: personalized services, information may bring economic & individual benefits, but risks furthering divisions, undermining community solidarity [ex/ personalized insurance ML system forecasts future medical, educational need; advantaged no longer see reasons to support those with greater needs]
- Privacy vs. transparency: need to respect privacy or intellectual property may make it difficult to provide satisfying information about algorithm, training data
- Convenience vs. dignity: increasing automation could make lives more convenient, but risks undermining unquantifiable values and skills that constitute human dignity and individuality

More Current Tensions

More examples of tensions between values

- Accuracy vs. explainability: accurate algorithms may be based on complex methods (e.g., deep learning), the internal logic of which is not fully understood
- Accuracy vs. fairness: an algorithm which is most accurate on average may systematically discriminate against a specific minority
- Preferences vs. equality: automation and AI could invigorate industries, spearhead new technologies, but also exacerbate exclusion and poverty
- Efficiency vs. safety: pursuing technological progress as quickly as possible may not leave enough time to ensure that developments are safe, robust and reliable

Sources

AI Now Reports (2017; 2018). Crawford et al., AI Now

Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research (2019). Whittlestone et al., Leverhulme Centre for the Future of Intelligence, University of Cambridge; Nuffield Foundation

A.I. Versus M.D. (2017). Siddhartha Mukherjee, The New Yorker

Course: The Ethics and Governance of Artificial Intelligence (2018). https://www.media.mit.edu/courses/the-ethics-and-governance-of-artificial-intelligence/