Social and Ethical Responsibilities of Computing (SERC):

New Activities at MIT

David Kaiser

November 2021
AI in the World

SERC at MIT

New Teams, New Tools
Challenges

The interface of artificial intelligence (AI) + machine-learning (ML) techniques with people — both individuals and groups — presents special challenges.

These challenges are exacerbated when AI + ML techniques move beyond research settings into real-world situations.
Challenges

The interface of artificial intelligence (AI) + machine-learning (ML) techniques with people — both individuals and groups — presents special challenges. These are often described as “algorithmic bias,” though there exists a whole range of challenges, across the entire ML pipeline.


These challenges are exacerbated when AI + ML techniques move beyond research settings into real-world situations.
Challenges

The interface of artificial intelligence (AI) + machine-learning (ML) techniques with people — both individuals and groups — presents special challenges. These are often described as “algorithmic bias,” though there exists a whole range of challenges, across the entire ML pipeline.

These challenges are exacerbated when AI + ML techniques move beyond research settings into real-world situations.
Facial Recognition Technologies

The 2019 NIST report (third in a series) analyzed 189 algorithms from 99 (mostly commercial) developers. They assessed performance both for 1:1 matching (for verification of a known individual) and for 1:n matching (for attempted matches within a whole dataset). [Main report: 75pp, plus 1200pp technical appendices.]
Facial Recognition Technologies

The 2019 NIST report (third in a series) analyzed 189 algorithms from 99 (mostly commercial) developers. They assessed performance both for 1:1 matching (for verification of a known individual) and for 1:n matching (for attempted matches within a whole dataset). [Main report: 75pp, plus 1200pp technical appendices.]
Facial Recognition Technologies

The 2019 NIST report (third in a series) analyzed 189 algorithms from 99 (mostly commercial) developers. They assessed performance both for 1:1 matching (for verification of a known individual) and for 1:n matching (for attempted matches within a whole dataset). [Main report: 75pp, plus 1200pp technical appendices.]

Summary: False positive rates highest for people from Africa and East Asia, lowest for people from Eastern Europe (large effect, $\sim 10^2 x$). Across geographical sets, false positives higher for women than men ($\sim 2 - 5x$).
Facial Recognition Technologies

The 2019 NIST report (third in a series) analyzed 189 algorithms from 99 (mostly commercial) developers. They assessed performance both for 1:1 matching (for verification of a known individual) and for 1:n matching (for attempted matches within a whole dataset). [Main report: 75pp, plus 1200pp technical appendices.]

Summary: False positive rates highest for people from Africa and East Asia, lowest for people from Eastern Europe (large effect, $\sim 10^2 \times$). Across geographical sets, false positives higher for women than men ($\sim 2 - 5 \times$).

Lots of research activity to develop efficient mitigation strategies in laboratory settings...
Facial Recognition Technologies

The US has *thousands* of distinct law-enforcement jurisdictions. Commercial facial-recognition technologies are *already being used* across the country, subject to *no regulation, standardization, or oversight*.
Facial Recognition Technologies

The US has thousands of distinct law-enforcement jurisdictions. Commercial facial-recognition technologies are already being used across the country, subject to no regulation, standardization, or oversight.

Several reported cases of Black men being wrongfully arrested due to incorrect FRT matches indicate a combination of failures: inadequate technical calibrations plus human failures to follow appropriate procedures.

More generally: new technologies are deployed within existing institutional frameworks.
Facial Recognition Technologies

The US has *thousands* of distinct law-enforcement jurisdictions. Commercial facial-recognition technologies are *already being used* across the country, subject to *no regulation, standardization, or oversight*.

Several reported cases of Black men being wrongfully arrested due to incorrect FRT matches indicate a *combination of failures*: inadequate technical calibrations *plus* human failures to follow appropriate procedures.

Contrast this with other examples of *forensic science* within the US, such as *fingerprinting*: far from perfect, but subject to *expert review, training, and standardization*. 
Facial Recognition Technologies

The US has *thousands* of distinct law-enforcement jurisdictions. Commercial facial-recognition technologies are *already being used* across the country, subject to *no regulation, standardization, or oversight*.

Several reported cases of Black men being wrongfully arrested due to incorrect FRT matches indicate a *combination of failures*: inadequate technical calibrations *plus* human failures to follow appropriate procedures.

One billion surveillance cameras are now in place within 50 countries. Within the US alone, facial images of *half the adult US population* are already included in databases accessible to law enforcement.
Natural-Language Processing

To date, one of the largest and most-used corpora of email — critical to algorithmic natural-language processing — comes from employees of Enron, the large Texas-based energy company that declared bankruptcy in December 2001.

Ultimately more than 20 executives pleaded guilty or were convicted, including on multiple felony fraud charges.
To date, one of the largest and most-used corpora of email — critical to algorithmic natural-language processing — comes from employees of Enron, the large Texas-based energy company that declared bankruptcy in December 2001.

Ultimately more than 20 executives pleaded guilty or were convicted, including on multiple felony fraud charges.

In 2003, the U.S. Federal Energy Regulatory Commission released 1.6 million emails sent to or from 158 Enron senior executives between 2000 – 2002. After minimal processing, the emails were simply made publicly available on a website. Employees were given a limited opt-out period.

(Recall 2003: Mark Zuckerberg was still an undergraduate ...)

Natural-Language Processing
Natural-Language Processing

More than 20k academic studies have been published that make use of the Enron email corpus (~1k per year).
Natural-Language Processing

More than 20k academic studies have been published that make use of the Enron email corpus (~1k per year).

Although no legal actions were ever brought against 99.83% of Enron email users, researchers are still finding $O(10^4)$ examples of sensitive personally identifiable information within the publicly available corpus, including Social Security Numbers, credit card numbers, birthdates, bank account numbers... Not to mention multiple extramarital affairs and other embarrassing episodes.

To probe the largest public-domain email database for indicators of fraud, we apply machine learning and accomplish four investigative tasks. First, we identify persons of interest (POI), using financial records and email, and report a peak accuracy of 95.7%. Secondly, we find any publicly exposed personally identifiable information (PII) and discover 50,000 previously unreported instances. Thirdly, we automatically flag legally responsive emails as scored by human experts in the California electricity blackout lawsuit, and find a peak 99% accuracy. Finally, we track three years of primary topics and sentiment across over 10,000 unique people before, during and after the onset of the corporate crisis. Where possible, we compare accuracy against execution times for 51 algorithms and report human-interpretable business rules that can scale to vast datasets.
More than 20k academic studies have been published that make use of the Enron email corpus (~1k per year).

Beyond the privacy concerns, the ubiquitous corpus continues to be used as training data for various natural-language-processing algorithms.

“If you think there might be significant biases embedded in emails sent among employees of a Texas oil-and-gas company that collapsed under federal investigation for fraud stemming from systemic, institutionalized unethical culture, you would be right. The Enron emails are simply not representative—not geographically, not socioeconomically, not even in terms of race or gender. Indeed, researchers have used the Enron emails specifically to analyze gender bias and power dynamics. And yet the Enron emails remain a go-to dataset for training AI systems.”

The Enron email corpus is a canonical example of “biased, low-friction data.”

Beyond “Algorithmic Bias”

Understanding Potential Sources of Harm throughout the Machine Learning Life Cycle

As machine learning (ML) increasingly affects people and society, awareness of its potential unwanted consequences has also grown. To anticipate, prevent, and mitigate undesirable downstream consequences, it is critical that we understand when and how harm might be introduced...

1. **Historical bias** arises when there is a misalignment between world as it is and the values or objectives to be encoded and propagated in a model. It is a normative concern with the state of the world, and exists even given perfect sampling and feature selection.

2. **Representation bias** arises while defining and sampling a development population. It occurs when the development population under-represents, and subsequently fails to generalize well, for some part of the use population.

3. **Measurement Bias** arises when choosing and measuring features and labels to use; these are often proxies for the desired quantities. The chosen set of features and labels may leave out important factors or introduce group- or input-dependent noise that leads to differential performance.

4. **Aggregation bias** arises during model construction, when distinct populations are inappropriately combined. In many applications, the population of interest is heterogeneous and a single model is unlikely to suit all subgroups.

5. **Evaluation bias** occurs during model iteration and evaluation. It can arise when the testing or external benchmark populations do not equally represent the various parts of the use population. Evaluation bias can also arise from the use of performance metrics that are not appropriate for the way in which the model will be used.

6. **Deployment Bias** occurs after model deployment, when a system is used or interpreted in inappropriate ways.

policing and sentencing; healthcare; real estate and finance; hiring …
Social and Ethical Responsibilities of Computing at MIT

**SERC Research**
- Developing responsible research and development practices
  - **Ethical Computing Platform**
    - Philosophy, Quest, Industrial
    - Performance Center, Cambridge City
    - Government + Corporate Research
    - Partners
  - Catalyzing new research and collaborations
    - Computing, Data and Racial
    - Justice Action Group
    - After-incident Report Group

**SERC Teaching**
- Developing original pedagogical materials
  - **Ethical Computing Protocol**
    - Taught in 6.031, 6.033, 6.034, 6.170, 6.5057, 24.131, 24.133
  - **MIT SERC Case Studies Series**
  - **Active Learning Projects**

**Broader Engagements**
- Catalyzing partnerships with external stakeholders
  - **Computing & Policy Task Forces**
    - AI & Finance (F’19)
    - AI & Mobility (S’21)
  - **International Student & Research Exchange - Europe**
    - Politecnico de Milano
    - TU Delft, Chalmers, ETH Zurich, RWTH
  - **Leading Responsible AI & Education**
    - Exchange with Harvard, Toronto, Stanford
SERC Leadership Team

**Julie Shah**
Professor, Department of Aeronautics and Astronautics
Director, Interactive Robotics Group, CSAIL
Associate Dean, SERC

**David Kaiser**
Germeshausen Professor of the History of Science
Professor of Physics
Associate Dean, SERC
Since Spring 2020:
70+ faculty and PIs
50+ undergraduates
35+ graduate students
7+ postdocs

Participants in SERC Action Groups represent all 5 Schools at MIT plus the new MIT Schwarzman College of Computing. They have engaged in a sustained fashion within multidisciplinary groups for one or more semesters during the past two years.
SERC Teaching

**Vision:**

Fractal model for embedding SERC material throughout the curriculum, making it *inescapable*.

Develop original pedagogical materials by *multidisciplinary teams* with members from across computing, data sciences, humanities, arts, and social sciences—for use in *each* of these types of classes.
Vision:

Fractal model for embedding SERC material throughout the curriculum, making it inescapable.

Develop original pedagogical materials by multidisciplinary teams with members from across computing, data sciences, humanities, arts, and social sciences—for use in each of these types of classes.

Commission and publish a new series of peer-reviewed case studies. The cases are brief (4k words), based on original research, and appropriate for use in undergraduate instruction across a range of existing courses and fields of study.

https://mit-serc.pubpub.org
SERC Teaching

Vision:

**Fractal model** for embedding SERC material throughout the curriculum, making it *inescapable*.

Develop original pedagogical materials by **multidisciplinary teams** with members from across computing, data sciences, humanities, arts, and social sciences—for use in *each* of these types of classes.

Commission and publish a new series of **peer-reviewed case studies**. The cases are **brief** (4k words), based on **original research**, and appropriate for use in **undergraduate instruction** across a range of existing courses and fields of study.

Each submission is **reviewed by 4-6 senior researchers at MIT**, drawn equally from computing and data sciences and from arts, humanities, and social sciences. Each submission is **also reviewed by MIT undergraduate volunteers** for balance and accessibility.

Cases are written by **subject-area experts**, not limited to MIT.

[https://mit-serc.pubpub.org](https://mit-serc.pubpub.org)
Winter 2021

The Case of the Nosy Neighbors
by Johanna Gunawan and Woodrow Hartzog

Inspired by companies like Clearview AI, Nextdoor, and Amazon, this case study asks students to assume the role of a high-ranking ethics-focused employee at a (fictional) neighborhood-focused social media company. It

Who Collects the Data? A Tale of Three Maps
by Catherine D'Ignazio and Lauren Klein

Who makes maps and who gets mapped? Using a comparative reading of three maps, this case study introduces the idea that data may be useful, but they are not neutral. Rather, they

The Bias in the Machine: Facial Recognition Technology and Racial Disparities
by Sidney Perkowitz

Facial recognition technology (FRT) appears in uses from providing secure access to smartphones to identifying potential suspects

The Dangers of Risk Prediction in the Criminal Justice System
by Julia Dressel and Hany Farid

Courts across the United States are using computer software to predict whether a person will commit a crime, the results of which are

https://mit-serc.pubpub.org
Summer 2021

Hacking Technology, Hacking Communities: Codes of Conduct and Community Standards in Open Source
by Christina Dunbar-Hester
Published: Aug 10, 2021
In recent years, the freeflare and open source software (FOSS) and hacking communities have engaged in lively debates about diversity and inclusion in their ranks.

Understanding Potential Sources of Harm throughout the Machine Learning Life Cycle
by Harsh Suresh and John Guttag
Published: Aug 10, 2021
As machine learning (ML) increasingly affects people and society, awareness of its potential unintended consequences has also grown. To anticipate, prevent, and mitigate undesirable downstream consequences, it is critical that we understand when and how harm might be introduced.

Identity, Advertising, and Algorithmic Targeting: Or How (Not) to Target Your “Ideal User”
by Tanya Kant
Published: Aug 10, 2021
Targeted or “personalized” marketing is an everyday part of most web users’ experience. But how do companies “personalize” commercial web content in the context of mass data aggregation?

Wrestling with Killer Robots: The Benefits and Challenges of Artificial Intelligence for National Security
by Erik Lin-Greenberg
Published: Aug 10, 2021
Countries around the world are increasingly turning to artificial intelligence (AI) for national security tasks ranging from intelligence analysis to identifying and attacking targets without human input. This proliferation of AI-enabled military technologies has significant...

Public Debate on Facial Recognition Technologies in China
by Tristan O. Brown, Alexander Staitman, and Celine Sui
Published: Aug 10, 2021
China's ascent on the global stage in the fields of artificial intelligence and facial recognition has been widely noted in Western-language scholarship and media. Much of the attention has focused on the applications of these technologies in government security systems and...
Summer 2021

Hacking Technology, Hacking Communities: Codes of Conduct and Community Standards in Open Source
by Christina Dunbar-Hester
Published: Aug 10, 2021
In recent years, the free, flexible and open source software (FOSS) and hacking communities have engaged in lively debates about diversity and inclusion in their ranks.

Understanding Potential Sources of Harm throughout the Machine Learning Life Cycle
by Harsh Suresh and John Guttag
Published: Aug 10, 2021
As machine learning (ML) increasingly affects people and society, awareness of its potential unwanted consequences has also grown. To anticipate, prevent, and mitigate undesirable downstream consequences, it is critical that we understand when and how harm might be introduced.

Identity, Advertising, and Algorithmic Targeting; Or How (Not) to Target Your "ideal User"
by Tanya Kant
Published: Aug 10, 2021
Targeted or “personalized” marketing is an everyday part of most web users’ experience. But how do companies “personalize” commercial web content in the context of massive data aggregation?

Wrestling with Killer Robots: The Benefits and Challenges of Artificial Intelligence for National Security
by Erik Lin-Greenberg
Published: Aug 10, 2021
Countries around the world are increasingly turning to artificial intelligence (AI) for national security tasks ranging from intelligence analysis to identifying and attacking targets without human input. This proliferation of AI-enabled military technologies has significant...

Public Debate on Facial Recognition Technologies in China
by Tristan G. Brown, Alexander Statman, and Celine Sui
Published: Aug 10, 2021
China's ascent on the global stage in the fields of artificial intelligence and facial recognition has been widely noted in Western-language scholarship and media. Much of the attention has focused on the applications of these technologies in government security systems and...

computer science, law, urban studies, literature, anthropology, political science, media studies, history

https://mit-serc.pubpub.org
Summer 2021

Computer science, law, urban studies, literature, anthropology, political science, media studies, history

Companion site coming soon: original homework problems, in-class demos, and active learning projects.

https://mit-serc.pubpub.org
SERC Research

Vision:

Resources and tools used by researchers who are engaged in large-scale applied research partnerships.

Research programs conducted with external partners that inform the design and use of the tools for better outcomes.

Faculty stewards and champions trained to work in multidisciplinary teams to assess potential harms as well as benefits from computing research.
SERC

SERC Research

Vision:

Resources and tools used by researchers who are engaged in large-scale applied research partnerships.

Research programs conducted with external partners that inform the design and use of the tools for better outcomes.

Faculty stewards and champions trained to work in multidisciplinary teams to assess potential harms as well as benefits from computing research.

Informed by safety-engineering practices in other fields, develop after-incident reports on unintended consequences originating from computing research, development, or implementation.
Informed by safety-engineering practices in other fields, develop after-incident reports on unintended consequences originating from computing research, development, or implementation.

Vision:

Resources and tools used by researchers who are engaged in large-scale applied research partnerships.

Research programs conducted with external partners that inform the design and use of the tools for better outcomes.

Faculty stewards and champions trained to work in multidisciplinary teams to assess potential harms as well as benefits from computing research.
SERC Research

Vision:

Resources and tools used by researchers who are engaged in large-scale applied research partnerships.

Research programs conducted with external partners that inform the design and use of the tools for better outcomes.

Faculty stewards and champions trained to work in multidisciplinary teams to assess potential harms as well as benefits from computing research.

Informed by safety-engineering practices in other fields, develop after-incident reports on unintended consequences originating from computing research, development, or implementation.

Our team: computer science, robotics and automation, law and public policy, urban studies and planning, architecture, philosophy, history of technology.

Train crash at Montparnasse station, Paris, 1895
Informed by safety-engineering practices in other fields, develop after-incident reports on unintended consequences originating from computing research, development, or implementation.

**Vision:**

- **Resources** and **tools** used by researchers who are engaged in large-scale applied research partnerships.

Research programs conducted with **external partners** that inform the design and use of the tools for **better outcomes**.

Faculty stewards and champions trained to work in **multidisciplinary teams** to assess potential harms as well as benefits from computing research.

**Our team:** computer science, robotics and automation, law and public policy, urban studies and planning, architecture, philosophy, history of technology.

**First study:** dataset combinations (including location / mobility data) and the reidentification of sensitive, personal healthcare information.

Train crash at Montparnasse station, Paris, 1895
New Tools

*Ethical Computing Platform:* a free and open online platform that research groups can use *iteratively* to address a series of questions about benefits and burdens of their proposed research design, data collection and protection protocols, and potential unintended consequences of a given project.
New Tools

**Ethical Computing Platform**: a free and open online platform that research groups can use iteratively to address a series of questions about benefits and burdens of their proposed research design, data collection and protection protocols, and potential unintended consequences of a given project.

coming soon!
New Tools

**Ethical Computing Platform**: a free and open online platform that research groups can use iteratively to address a series of questions about benefits and burdens of their proposed research design, data collection and protection protocols, and potential unintended consequences of a given project.
Computing and the data sciences are enabling fantastic progress and benefits for society…
Computing and the data sciences are enabling **fantastic progress** and **benefits for society**…

… but like **any new technology**, they carry risks of **unintended consequences** and **real harms** for individuals and groups.
Computing and the data sciences are enabling fantastic progress and benefits for society... 

... but like any new technology, they carry risks of unintended consequences and real harms for individuals and groups.

The most significant challenges arise when techniques leave the lab and impact real-world situations.
Identifying potential consequences and rectifying harms requires **input** and **expertise** from **across the fields** of science, engineering, humanities, and the social sciences …

… as well as working with **partners beyond academia** — including people who are **impacted** by the latest technologies as well as the people who **implement** them.
Identifying potential consequences and rectifying harms requires **input** and **expertise** from across the **fields** of science, engineering, humanities, and the social sciences …

Our aim with **SERC** is to **freely share** what works and what doesn’t, so others can **adopt** and **adapt** what we do. This requires building **new connections** among researchers and students, across the MIT community and beyond.

... as well as working with **partners beyond academia** — including people who are **impacted** by the latest technologies as well as the people who **implement** them.