Data Ethics and Professionalism

ISEA Session 15

Patrick C. Kennedy
University of Oregon
May 10, 2024

Overview

1. Data ethics
2. Reproducibility
3. Communicating results

Data ethics


IES
Institute of
Education Sciences


AmplifyLearn.AI


UNIVERSITY of WASHINGTON
eScience Institute
ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS

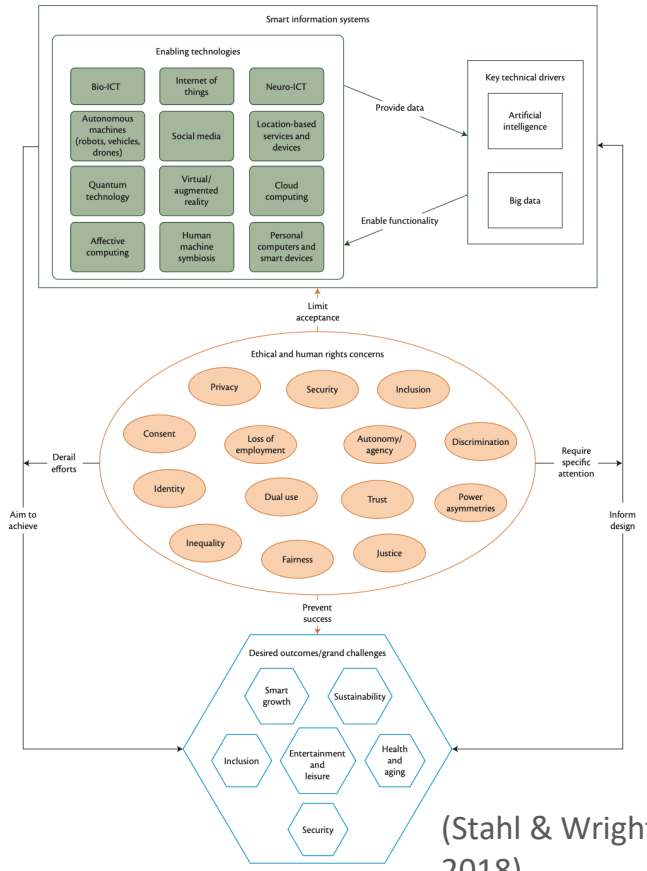

UNIVERSITY OF
OREGON

Motivation

- **Potential and peril**
  - **Massively accelerated rate at which data are produced**
  - **Ways in which data are analyzed and understood are changing rapidly**
- **Artificial intelligence (AI)**
  - **The use of hardware, software, and applications to perform analytics and mimic human cognitive capabilities** (Stahl & Wright, 2018)

Smart information systems

- Shorthand for technologies that use AI, machine learning, & big data (Stahl & Wright, 2018)
- Main takeaway:
  - Wide range of tech
  - Driven by AI & big data
  - Influenced by many concerns
  - That affect desired outcomes



(Stahl & Wright, 2018)

Smart information systems (SIS)

- Modern SIS are ubiquitous
  - Amazon recommendations/Alexa
  - Google search/Google Translate
  - Facebook/Instagram
- Other (education) examples?

Edtech SIS

- Personalized learning systems (PLS) based on individual students' needs & skills (Regan & Jesse, 2019)
- Both exciting and controversial
  - Facilitates collection of more, and more granular, information about students, teachers, and families
- Widely used and promoted in and beyond US
  - Gates Foundation, Chan Zuckerberg Foundation
  - Ubiquitous: used in nearly all US school districts

IES Institute of Education Sciences

AmplifyLearn.AI

UNIVERSITY of WASHINGTON eScience Institute
ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS

UNIVERSITY OF OREGON

Just privacy?

- Privacy as multi-faceted concept incorporating multiple distinct ethical concerns (Westin, 1967; Solove, 2008)
- US discussions largely focus on privacy (Regan, 1995)
- Can lead to a myopic focus on protecting student information from inappropriate access
- Reality is more nuanced

IES Institute of Education Sciences

AmplifyLearn.AI

UNIVERSITY of WASHINGTON eScience Institute
ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS

UNIVERSITY OF OREGON

Information privacy

- Minimize the amount of information collected to that required for a specific purpose
- Addressed by fair information practice principles (FIPP): notice, consent, choice, and transparency
  - Family Educational Rights and Privacy Act (FERPA, 1974)
  - Children's Online Privacy Protection Act (COPPA, 1998)
- Widespread agreement that focusing on info privacy is ineffective in a big data context

Anonymity

- The idea that individuals should be able to remain anonymous if they so choose
- Nearly impossible to anonymize modern data sets, as a handful of characteristics are likely sufficient to identify individuals, particularly when combined with other data

Surveillance

- Monitoring of activities, extraction of data about those activities, and analysis of resulting data
  - Time to answer a question or read a page
  - Keystrokes or other patterns of responding
  - Location, time of day, other students doing same task
  - Cross-matched with other data (e.g., how much a student moves, or time spent on social networking)
  - Predictive analytics to determine patterns, strengths and weaknesses, and advice about how to personalize

Autonomy

- Big data algorithms jeopardize autonomy by leading or nudging people to make specific decisions
- Although PLS may seem to be in students' best interests, they can influence e.g., instruction without allowing students a choice

Discrimination & Bias

- SIS can
  - Perpetuate prejudices and accentuate social inequities in subtle ways
  - Create new forms of inequality
  - Introduce potential for bias from:
    - Those who design the systems
    - The algorithms themselves
- Algorithmic complexity can make identification of bias and discrimination difficult

“Average person” in US states

Institute of Education Sciences

AmplifyLearn.AI

UNIVERSITY of WASHINGTON eScience Institute
ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS

UNIVERSITY OF OREGON

Ownership

- Who owns data produced by PLS?
- To what extent does data generated about and by students as they use PLS belong to the school or district vs. companies?
- Should there be limits on how companies use this data (e.g., to improve their offerings)?

What can be done?

- Responsible Research and Innovation (RRI)
  - Efforts to ensure that SIS processes & outcomes are acceptable, desirable, and sustainable
  - Built on principles of technology ethics, technology assessment, science and technology studies, and philosophy of technology

RRI

- Consensus is elusive
- Common foci include:
  - Stakeholder engagement
  - Openness and transparency
  - Willingness to be flexible and responsive
  - Integration into projects, funding, & support environments

IES Institute of Education Sciences

AmplifyLearn.AI

UNIVERSITY of WASHINGTON eScience Institute
ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS

UNIVERSITY OF OREGON

Discussion

- Join the breakout room most relevant to you
  - 1) Info privacy, 2) anonymity, 3) surveillance
  - 4) autonomy, 5) discrimination/bias, 6) ownership
- Discuss implications of that concern in Edtech
  - What can we, as educators with data science training, do to bring awareness to or help mitigate the issues?
  - How can we help develop a culture of responsibility among stakeholders for the processes and outcomes they develop and implement in schools?

IES Institute of Education Sciences

AmplifyLearn.AI

UNIVERSITY of WASHINGTON eScience Institute
ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS

UNIVERSITY OF OREGON

BREAK

> 5-minute break

Reproducibility

Reproducibility

- **What** does it mean to be reproducible?
- **Why** does reproducibility matter?
- **How** is reproducibility achieved?

IES Institute of Education Sciences

AmplifyLearn.AI

UNIVERSITY of WASHINGTON eScience Institute
ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS

UNIVERSITY OF OREGON

Reproducibility: What

- Conceptual
  - Replicating a study with new, independent data
    - Expensive
    - Hard(er) to get funded and published
    - Methodological challenges
- Computational
  - When others reproduce study results given only the original data, code, & documentation
    - Retains many advantages while minimizing the largest barrier (i.e., costs)

Reproducibility: Why

- **Benefits those who do it**
  - **Encourages robust documentation**
  - **Makes revisions easier**
  - **Promotes modularity and reuse of code**
  - **Provides an indication of rigor, trustworthiness, & transparency**
  - **Increases citation rates**

Reproducibility: Why

- **Benefits the larger community**
  - **Makes findings more accessible**
  - **Allows others to learn from your work**
  - **Facilitates follow-up studies**
  - **Leads to faster progress**
  - **Provides protection when mistakes occur**

Reproducibility: How

- **Plan**
  - Develop a well-defined question
  - Write and register study protocol
  - Justify the proposed sample size
  - Construct a data management plan
  - Proactively address sources of bias
- **Execute**
  - Avoid questionable practices
  - Interpret significance carefully
  - Make research open
- **Report**
  - Report all findings
  - Follow relevant reporting guidelines

IES Institute of Education Sciences

AmplifyLearn.AI

UNIVERSITY of WASHINGTON eScience Institute
ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS

UNIVERSITY OF OREGON

Reproducibility: Analysis

- Before: Plan data storage & organization
  - Location & format, data structure, metadata
- During: Use coding best practices
  - Clean, well commented code; code review; document environment & parameters
- After: Finalize & share results
  - Include input data, scripts, program versions, parameters, and important intermediate results
  - Choice of repository, prioritize DOI for citations

IES Institute of Education Sciences

AmplifyLearn.AI

UNIVERSITY of WASHINGTON eScience Institute
ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS

UNIVERSITY OF OREGON

Reproducibility: Resources

- UW eScience Reproducible and Open Research
  - http://uwescience.github.io/reproducible/
- Coursera Reproducible Research course
  - https://www.coursera.org/learn/reproducible-research
- Reproducibility and Replicability in Science
  - https://www.ncbi.nlm.nih.gov/books/NBK547537/

IES Institute of Education Sciences

AmplifyLearn.AI

UNIVERSITY of WASHINGTON eScience Institute
ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS

UNIVERSITY OF OREGON

Communicating results

Clear communication

- > Motivate the contribution
- > Provide context (what's known)
- > Balance interpretability and accuracy

A Checklist for Communicating Science and Health Research to the Public: https://www.nih.gov/about-nih/what-we-do/science-health-public-trust/checklist-communicating-science-health-research-public

IES Institute of Education Sciences

AmplifyLearn.AI

UNIVERSITY of WASHINGTON eScience Institute
ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS

UNIVERSITY OF OREGON

Clear communication

> **Use visuals that can be easily understood**

A Checklist for Communicating Science and Health Research to the Public: https://www.nih.gov/about-nih/what-we-do/science-health-public-trust/checklist-communicating-science-health-research-public

IES Institute of Education Sciences

AmplifyLearn.AI

UNIVERSITY of WASHINGTON eScience Institute
ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS

UNIVERSITY OF OREGON

Clear communication

- > Be respectful
- > Avoid potentially offensive terms
- > Include other relevant resources
- > Provide citations and sources

A Checklist for Communicating Science and Health Research to the Public: https://www.nih.gov/about-nih/what-we-do/science-health-public-trust/checklist-communicating-science-health-research-public

IES Institute of Education Sciences

AmplifyLearn.AI

UNIVERSITY of WASHINGTON eScience Institute
ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS

UNIVERSITY OF OREGON

Innovative communication

- Get the basics right
  - Define objectives, specify your audience, frame your message, and develop a dissemination plan
- Use websites, social media, and unique identifiers to make your work visible
- Foster participation and collaboration

(Ross-Hellauer et al., 2020)

IES Institute of Education Sciences

AmplifyLearn.AI

UNIVERSITY of WASHINGTON eScience Institute
ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS

UNIVERSITY OF OREGON

Innovative communication

- Embrace open science
  - Principles of equitable participation and transparency that enable others to collaborate in, contribute to, scrutinize and reuse research, and spread knowledge as widely as possible
- Think beyond traditional research outputs

(Ross-Hellauer et al., 2020)

IES Institute of Education Sciences

AmplifyLearn.AI

UNIVERSITY of WASHINGTON eScience Institute
ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS

UNIVERSITY OF OREGON

Innovative communication

- > Engage stakeholders
- > Think data visualization
- > Reflect and respect diversity
- > Find and use the right tools
- > Evaluate, evaluate, evaluate

(Ross-Hellauer et al., 2020)

IES Institute of Education Sciences

AmplifyLearn.AI

UNIVERSITY of WASHINGTON eScience Institute
ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS

UNIVERSITY OF OREGON

Assignment

- Write a brief reflection (300-500 words) on the current or anticipated role of ethics & reproducibility in your own work.
  - What areas are most pertinent to you? How might principles of reproducibility help address (some of) these issues?

References

Alston, J. M., & Rick, J. A. (2020). A beginner's guide to conducting reproducible research. *Bulletin of the Ecological Society of America,* 102(2), e01801. https://doi.org/10.1002/bes2.1801

Regan, P. M. (1995). Legislating privacy: Technology, social values, and public policy. Chapel Hill: University of North Carolina Press.

Regan, P.M., & Jesse, J. (2019). Ethical challenges of edtech, big data and personalized learning: twenty-first century student sorting and tracking. *Ethics and Information Technology*, 21, 167–179. https://doi.org/10.1007/s10676-018-9492-2

Ross-Hellauer, T., Tennant, J.P., Banelyte˙, V., Gorogh, E., Luzi, D., Kraker. P., et al. (2020) Ten simple rules for innovative dissemination of research. PLoS Comput Biol 16(4): e1007704. https://doi.org/10.1371/journal.pcbi.1007704

Schwab, S., Janiaud, P., Dayan, M., Amrhein, V., Panczak, R., Palagi, P. M., Hemkens, L. G., Ramon, M., Rothen, N., Senn, S., Furrer, E., & Held, L. (2022). Ten simple rules for good research practice. *PLoS Computational Biology*, 18(6), e1010139

Solove, D. (2008). Understanding privacy. Cambridge: Harvard University Press.

Stahl, B.C. & Wright, D. (2018). Ethics and privacy in AI and big data: Implementing Responsible Research and Innovation. *IEEE Security & Privacy*, 16 (3), 26-33. https://doi.org/10.1109/MSP.2018.2701164

Westin, A. (1967). Privacy and freedom. New York: Atheneum.

IES Institute of Education Sciences

AmplifyLearn.AI

UNIVERSITY of WASHINGTON eScience Institute
ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS

UNIVERSITY OF OREGON