

Data ethics

Summer 2023

Felix Chopra

Plan today

1. Ethics & research
 - What and why
 - Ethics in social sciences
2. Ethics & social data science
 - Privacy
 - GDPR
3. Ethics & scraping
 - Practical advice

Ethics

What is ethics?

- Every society has its own moral compass, shaped by culture, history, and experience.
- Examples:
 - Killing someone for pleasure: Universally seen as immoral due to respect for human life.
 - Helping a stranger without expecting anything in return: Generally seen as moral, valuing altruism and community.
- Ethics: A *systematic* approach to moral judgments based on reason, analysis, synthesis and reflection.
- Law: often the institutional embodiment of ethics

What is ethics?

- Challenges
 - No universally agreed-upon general theory of ethical behavior
 - Disagreements are common (among philosophers/between cultures)
- Distinguish
 - **Descriptive ethics:** Observing and detailing how people actually make moral choices and reason about them (e.g. “white” lies are acceptable)
 - **Normative ethics:** Determining how people *ought* to act, establishing standards or norms for right and wrong behaviors.

Influential ethical theories

- **Utilitarianism** (e.g. Bentham, Mill, Hume)
 - Consequentialist, outcome-based theory of ethical behavior
 - Benefits of an act should outweigh the harm
 - Highly influential in econ & law
- **Deontology** (e.g. Kantianism)
 - Rule-based theories of right and wrong (“Don’t steal”)
 - Focus on universal principles instead of situation-specific circumstances

Research & ethics

What's worth knowing?

- Research programs can involve ethical dilemmas and challenges.
- Should we pursue research even if...?
 - ...it can be abused by others?
 - Natural sciences → development of nuclear, biological or chemical weapons
 - Mathematics & computer sciences → decryption and surveillance
 - Behavioral sciences → manipulation, propaganda
 - ...it causes harm to humans or animals?
 - Experimental treatment for a disease
 - Psychological stress in experiments
 - ...it requires deceiving/manipulating participants?

What's worth knowing?

- Today, researchers must consider ethical dilemmas connected to their research.
- However, this has not always been the case...

History of ethically problematic research

– Medical sciences

- J. Marion Sims' (1813-1883) founding research for modern gynecology
- Tuskegee Study of Untreated Syphilis in the Negro Male (1972, USA)

– Social sciences

- Monster Study (Johnson, 1922)
- Robbers Cave Experiment (Sherif, 1954)
- Hawlow's (1958) mother-infant separation study
- Milgram Experiment (1961)
- Stanford prison experiment (Zimbardo, 1971)

Institutional Review Boards (IRB)

- Recognition of need for guidelines for human subject research
 - Nuremberg Code (1954)
 - Declaration of Helsinki (1964)
 - Belmont Report (1979)
- **IRB:** Prior review of all research involving human subjects to protect the welfare, rights, privacy, and dignity of human subjects.
- Criteria
 - Beneficence: Max. benefits for science/humanity/subjects and avoid/min. harm
 - Respect: Protect the autonomy & privacy rights
 - Justice: Ensure fair distribution of costs and benefits of research

Ethics approval today

- You need ex-ante IRB approval to conduct research
 - (Good) journals will refuse to publish your study without IRB!
- You “apply for ethics” at your local IRB board (questionnaire):
 - What do you want to do?
 - What ethical dilemmas do you expect? How do you address them?
 - Do the benefits justify the costs/risks?
- IRB reviews application, asks for clarifications, may ask for changes to the study; sometimes multiple rounds of revisions
- KU econ: https://www.economics.ku.dk/research/ethic-committee_econ/

IRB review is not easy

Let's play IRB

- I will present the design of research studies
- In your groups (2 min):
 - Make an approval decision
 - If you don't approve: Why not? What would change your mind?
- Then: we vote & discuss in class

Case I: Morals and markets

- Do market interaction promote immoral behavior compared to non-market institutions?
- Experiment with students, randomized to condition A/B.
 - **A:** Student receives a real mouse. Measure WTP for giving up the mouse. If students sells the mouse, it is killed by the experimenter.
 - **B:** Students are anonymously grouped into sellers and buyers. Sellers have mice, buyers have money. Double auction clears the market.
 - Again, every transaction involves the death of the mouse
 - Subjects know the rules, give consent, there is no deception
 - All mice are so-called “surplus mice” that would have died anyways
- Is the market price of the life of a mouse lower in condition B?

Case II: Social status & violence

- Does perceived social status affect our willingness to behave violently towards others?
- Laboratory experiment:
 - Participants submit photos where they look good before the study, and are then paired into groups (anonymous)
 - Outcome: Decision to inflight a painful electric shock on the other participant in return for money
 - Treatment: Three very attractive members of the opposite sex say whether they would prefer you or the other participant in your group as a romantic partner

Case III: Is media censorship effective?

- Design:
 - Field experiment at a Chinese university.
 - Survey students on political knowledge, beliefs, and exposure/access to politically sensitive information before and after the interventions
 - Treatment: Provide random subset with access to technology to circumvent the “Great Firewall”, encourage them to obtain acquire politically sensitive info from Western news websites.

Some guiding principles

- Principle of informed consent
 - Exceptions only for field experiments in rare cases
 - However: Our standards of consent may not be feasible everywhere (e.g., reluctance to sign legal forms in developing countries)
- Principle of anonymity
- Principle of relative improvement
 - Interventions are not allowed to make people worse off compared to the status quo
 - Example: Job search. Intervention can assign extra incentives for job search, but we cannot punish people for not searching.
- Principle of no deception (only econ)
 - Journals will not publish studies involving deception

Ethics & Social Data Science

Ethics of Big Data

“In the analog age, most social research had a relatively limited scale and operated within a set of reasonably clear rules. Social research in the digital age is different. Researchers—often in collaboration with companies and governments—have more power over participants than in the past, and the rules about how that power should be used are not yet clear”

(Bit by Bit, Ch. 6: Ethics)

Challenges

- A lot easier to observe (certain) behaviors w/o awareness or consent
- Unanticipated secondary use of data
- Easier to directly intervene in ‘participants’ life
- Inconsistent & overlapping rules, laws and norms on data use

Ethics of Big Data

- **Privacy and anonymity:**
 - How to ensure anonymity if data is big?
 - If photographing people in public places is ok, is recording what they say on Facebook also ok?
- **Informed consent:** Is it necessary? Is it consent "informed" if study details are shrouded in 80 pages of legal click-through?
- **Ownership of data**
- **(Algorithmic fairness)**

Anonymity in big data is an illusion



Phone locations 0500h Monday morning → can predict where people at given time with 85% accuracy

Individual data and privacy

- Statistics Denmark:
 - Data users cannot present data at the individual level
- Absolute no-go's:
 - Max of the income distribution
 - Median of income distribution
 - Max income in parish
- Things are different if you can anonymize data (e.g. a man in his 20s in Copenhagen)
- But! Well-known examples of re-identification from public data
 - Often in combination with auxiliary data
 - An [overview](#)
 - An [example based on credit card data](#)

Why privacy?

Why privacy?

Individual demand for privacy might reflect

1. Intrinsic value of privacy – a principle of privacy
 - Preference-based explanation
2. Privacy to preserve informational rents
 - Consumers vs firms
3. Privacy and politics

I. Privacy for its own sake

- People may simply value privacy in itself (preference), even if privacy provides no instrumental benefits!
- But: Public goods problems
 - Example: medical research.
 - Share existing info on medical history, no cost to individuals. Some will not contribute, citing privacy concerns – but benefits of research accrue to everybody
 - DK: No consent necessary for register studies or re-use of data
 - Similar: Privacy rules for social science research, or monitoring in public places
- Visit <https://teol.ku.dk/privacy/> on how the concept originally evolved

II. Preserve informational rents

- **Consumers:** Willingness to pay (WTP), characteristics, and behavior often private information
 - Willingness to pay: 1st class vs. 2nd class train ticket
 - Characteristics: Taste, Genetics, Personality
 - Behavior: driving and insurance, [physical activity](#)
 - Value of time / search costs
 - Example: [Internet steering](#)
- **Firms:** Intellectual Property Rights, corporate strategy
 - Industrial espionage major problem
 - LinkedIn-story;
 - Firms where data is only asset

Digression: Necessary secrecy

- You cannot be told how your bank constructs your credit scores. Why?
 - Goodhart's law: people will attempt to outmaneuver measurements
 - Thought experiment: If spending on shoes good indicator of account overdraft → shoe lovers will have others buy for them, ceases to be a good measure

Case of Google Flu

- Google Flu: web searches for flu symptoms predicted actual regional flu cases
- By-product of Google's main service
- But from 2010, not so well: overestimated actual flu cases, partly as result of autosuggest feature, partly because model was overfitted (will return to that under machine learning)
- Best predictor: number of cases **past** week

III. Privacy and politics

- Authorities are not allowed to register party identification
 - Originally for freedom of political expression but also: majority in city council could pay out cash assistance / kontanthjælp based on, say, union membership
- These days: Privacy as a political platform

Privacy and law

Legal framework for personal data

- Before 2018:

Persondataloven

- After 2018:

General Data Protection
Regulation (GDPR)

GDPR

- Link: <https://gdpr-info.eu>
- "The objective of this new set of rules is to give citizens back control over of their personal data, and to simplify the regulatory environment for business."
- **Individual consent** plays a much larger role (but special rules for DK)
- **Revocation of access - right to deletion** ("right to be forgotten")
- Some types of personal data are considered **sensitive** (health, political views, social problems)

GDPR

- Very different rules for
 - Research
 - Public administration
 - Private firms / organizations
- Potentially large penalties for non-compliance or misuse
- New job: DPO – Data Protection Officer
- Fair to say: Interpretation of GDPR work-in-progress

Research exemption

- § 10. Information as mentioned in the data protection regulation, article 9, subsection 1, and Article 10 may be processed if this is done solely for the purpose of carrying out statistical or scientific studies of significant societal importance, and if the processing is necessary for the purpose of carrying out the studies.
- The information covered by subsection 1, may not later be processed for other than scientific or statistical purposes. The same applies to the processing of other information which is only carried out for statistical or scientific purposes in accordance with Article 6 of the Data Protection Regulation.

DK exception: Re-use data collected for other purposes for research

- (50) Processing of personal data for purposes other than the purposes for which the personal data were originally collected should only be permitted if the processing is compatible with the purposes for which the personal data were originally collected. In this case, no other legal basis is required than that which justified the collection of the personal data.

If processing is necessary to carry out a task in the public interest or belongs to the exercise of public authority, which the data controller has been assigned, EU law or the national law of the Member States can determine and clarify the tasks and purposes for which it should be compatible and lawful to carry out further processing. Further processing for archival purposes in the public interest, for scientific or historical research purposes or for statistical purposes should be considered to be compatible with lawful processing activities.

Trade-offs

- Sacrifice accuracy/quality for privacy?
- In some cases: no trade-off in analysis, only in presentation
- Danish firm data: Stat Denmark does not report figures for industries with very few firms
- New approaches: Analysts don't see data, but can make calculations on it
 - May limit *fee*/for data; but some data is better than no data at all
- More general problem: How much info do we get from data under constraint of “no (re)identifiability”?
 - Active research area in computer science


Scraping

Ethics considerations

- Is it ethical to scrape competitors' *likes* on Facebook? Is it illegal?
 - Ethics (and law) sometimes used as arguments to stifle competition.
 - See [LinkedIn case](#) - reviewed by SCOTUS, and reaffirmed by Appeal courts
- Can you scrape data and resell? Or repackage?
- Does data collection cause significant costs (time or money) to firms and/or individuals?
- Typically, more welcoming towards students, but be careful – and if in doubt, ask us!

LinkedIn Data Scraping Ruled Legal



Emma Woollacott Senior Contributor 
Cybersecurity

f

t

in



Photocredit: Getty GETTY

A court has ruled that it's legal to scrape publicly available data from LinkedIn, despite the company's claims that this violates user privacy.

San Francisco-based start-up hiQ Labs harvests user profiles from LinkedIn and uses them to analyze workforce data, for example by predicting when employees are likely to leave their jobs, or where skills shortages may emerge.

After LinkedIn took steps to block hiQ from doing this, hiQ won an injunction two years ago forcing the Microsoft-owned company to remove the block. That injunction has now been upheld by the 9th US Circuit Court of Appeals in a 3-0 decision.

Permissions

- Go to the *Robots Exclusion Protocol* of a website by adding `/robots.txt` to its URL:
 - **User-agent**: the type of user to which the rules apply
 - **Disallow**: Directories and subdomains of the website not allowed to be scraped
 - **Allow**: What you are allowed to scrape

```
User-agent: *
Disallow: /*/about-us/contact/contact-spotify-password/
Disallow: /*/about-us/contact/contact-spotify-account/
Disallow: /*/get-spotify/*
Disallow: /*/xhr/*
Disallow: /*/external/*
Disallow: /*/legal/advertiser-terms-and-conditions/
Disallow: /*/account/cls/*
Disallow: /*/starbuckspartners
Disallow: /starbuckspartners
Sitemap: https://www.spotify.com/sitemap.xml
```

During scraping...

- Limit your request speed to reasonable levels to avoid harm to the website owner with `time.sleep(delay)`
- Identify yourself: Use the `User-Agent` attribute to provide contact information

```
1  import requests
2
3  # identify yourself
4  headers = {"User-Agent": "Felix Chopra (felix.chopra@econ.ku.dk).  
Researcher at University of Copenhagen. Collecting data for research  
purposes. Contact me if you have any questions."}
5
6  # request
7  url = "https://www.wikipedia.org/"
8  response = requests.get(url, headers=headers)
9
10
11
```

What is public?

- This course: try to limit yourself to publicly available information.
- What not to do
 - Download content behind paywalls (e.g. NYT)
 - Scrape posts from non-public discussion boards or non-public content from social media where sensitive information could be disclosed
 - Do not access documents/data you should not have access to (e.g., hacked data you find in the dark web)
- If in doubt, err on the side of not collecting data.

Preserving anonymity

- This course: don't collect personally identifiable information.
- Examples:
 - User names, addresses, location data
 - Social media posts: Is there a risk of identification?

Questions for your projects

- Do you respect privacy?
- Can individuals be identified in your data?
- What are potential consequences of (re)identification?
- What are the terms and conditions for scraping and using data?
- Are there ethical considerations
 - With respect to individuals?
 - With respect to firms or organizations?

The end

+ bonus material

Ethics of algorithmic decisions

Challenges

- Is it unethical find correlation btw smoking and lung cancer, even if insurance companies use this to increase premiums for smokers?
 - What about correlation between genetic markers and, say, chronic diseases, increased mortality risk?
- Ethics is not about preventing stuff from being done
 - but reasonable balance between costs and benefits (ex: hidden camera/mic : not ok for mundane things, but maybe ok if benefits are huge; random drug screening of employees may violate privacy, but ok if job involves public safety)

Ethical considerations for big data

- What about business ethics?
 - Example: Google Location. Show where friends/family are in real time – but requires consent
 - Should you see data from friends of friends?
- Algorithms as "Weapons of Math Destruction"
 - Insurance based on where you live, your name/ethnicity
 - Entry into university based on prediction of completion?
 - Loan interest rates based on past behavior?
 - FAT ML: Fair, Accountable, and Transparent Machine Learning

Example – biased technology

- Buolamwini and Gebru found that face recognition was biased against people of color and women
 - huge implications for consequences of other tech that depends on it: being found by the police, whether phone can unlock
- Potential biases in predictive algorithms:
 - recidivism risk (used in relation to criminal cases)
 - study completion (admission, use of resources)
 - fraud detection (credit card, social benefits, tax)

The role of social science in tech

- What are the societal/economic consequences of adopting certain algorithms?
 - Can algorithms debias human biases, e.g. in police inspections? Preliminary answer – yes and simultaneously raise efficiency (e.g. jail likely re-offenders).
 - Can algorithms be used for inclusive policies, yes.

The end