Ethics and Software Engineering research

ARC Laureate Professor John Grundy

September 2023

https://www.monash.edu/it/humanise-lab

MONASH University

MONASH INFORMATION TECHNOLOGY

Australian Government
Australian Research Council

HUMANISE

MONASH University
HumaniSE Lab
Human-centric Software Engineering Lab
CREATING SOFTWARE FOR THE DIVERSE HUMAN COMMUNITY

Acknowledgement of Country

As we gather for this meeting physically dispersed and virtually constructed let us take a moment to reflect the meaning of place and doing so recognise the various traditional lands on which we do our business today.

We acknowledge the Elders – past, present and emerging of all the land we work and live on and their Ancestral Spirits with gratitude and respect.

I acknowledge the people of the Kulin nations, the traditional owners of the land on which I am meeting with you from today.

Outline

What SE research has ethical considerations?

Key ethical dilemmas / breech of ethics possibilities in SE research

How do we address them (properly)?

SE Research & Ethical Considerations

- User studies
  - Evaluating software prototypes with end users
  - Survey, interview, observe developers, end users
- Data analysis
  - Sensitive data from software e.g. health data
  - Development data e.g. github, Stack Overflow, user reviews
  - Harvested data e.g. from company web sites
- Software development
  - Extending existing software
  - Reusing libraries, APIs, data
- Machine learning & SE
  - Training ML models
  - Using Large Language Models and related

User Studies

Must have Ethics approval before begin

Recruitment – what if recruiting industry practitioners to ask about company sensitive data – get company/manager approval?

Paying participants – need Ethics approval, some +ve e.g. their valuable time being used - shouldn't they be paid for???

Sensitive data - what if ask developers about their bosses as part of management of SE teams study??

Reporting – must aggregate data, make sure identity of participants can't be determined e.g. what if 1 woman on team and ask about gender-related issues on working in team?

Follow ethics approval, dataset rules e.g. don't harvest emails from Bugzilla issues and send emails to…

SE Data analysis

Often have access to highly sensitive data e.g. health – must be very careful about rules around using, informed consent, anonymity etc

Harvesting app reviews, Stack Overflow posts, gihhub discussions – needs Ethics approval (apparently) ; be careful with quoting/link to usernames etc

Balanced dataset to avoid unbalance data / over-fitting / inaccurate conclusions

Have permission to scrape, use the data (see terms and conditions)

Making data available for replication studies etc

Follow Ethics approval around data harvesting, analysis, storage etc

Software Development

Licenses allow modification, distribution

Ask authors for update versions, permissions

Making data (code etc) available for replication studies

Proper attribution of source of APIs, libraries, prototype etc

Criticising others code (or data) – be transparent, fair, provide yours so you can be criticised! Give chance to respond if possible.

Machine Learning & SE Data

Unbalanced data sets – balance, carefully caveat conclusions etc

Wrong ML methods, not optimised hyperparameters

Not providing data for others replication studies

Picking and choosing data

Picking and choosing measurements, reported measurements (p-hacking etc)

LLMs – is it ethical to use these when no idea what data trained on, has appropriate permission been given etc???

Summary

Some key gotchas and mitigations

- Many Software Engineering studies need Ethics approval in advance
- Pay participants in studies for their valuable time
- Make sure have permission to use code, data (just because on web…)
- Use care around choice of data, ML, tuning, etc
- Provide datasets (if allowed to - usually can't provide e.g. raw interview data as Ethics Committee usually doesn't allow) for replication, checking
- Be conservative about conclusions from studies
- Don't plagiarise – includes data, code etc as well as paper content!