How to Think About Broader Impacts of Research

Jesmin Jahan Tithi
Intel Labs, CA, USA
jesmin.jahan.tithi@intel.com

January 2021

One way to begin to understand what the broader impacts of your research could be, consider the following questions.

- Think about first-order, second-order, and third-order (indirect) impacts of your project as described below.

  - First: Is there any direct malicious use of this research result?

  - Second: How could your system be modified to cause intentional or unintentional harm? Can anyone else the system or ideas to harm others? What is the worst way someone could use your research finding, given no resource constraints?

  - Third: Will there be any impact on the society, government, subsystem of government, and economic systems as an indirect implication of the use of the system? Will it need changes in systems and policies to maintain balance?

- Will there be any mental and emotional harms, threats to physical safety, infringement on rights, job loss, political or institutional degradation enabled by your research?

- Can it discriminate against a particular group, beneficial for some and harmful to the minority? Is it fair or free of bias?

- Can it create division in society?

- Is it secure, trustworthy, explainable?

- Is your research environmentally sustainable? Does it take too much compute time to generate a lot of carbon footprint?

- Will releasing the code enable bad actors? Should you consider the staged release of artifacts?

- Is there anyways to mitigate harm? Can any policy/law change require to stop that harm?
- Will it cause security or privacy issues?
- Do the benefits outweigh the costs?

Here are a few additional questions mentioned in Partnerships of AI's publication-norms case study:

- What makes this research important? What beneficial effects do you hope it will have? If this
- Who will use your research? Is it a foundational research, or could it be directly applied or productized? What capabilities does your work unlock?
- Does your research contribute to progress on safety, security, beneficial uses, or defense against malicious use?
- What is the worst way someone could use your research finding, given no resource constraints?
- What is the most surprising way someone could use your research finding?
- How would a science fiction author turn your research into a dystopian story? Are dystopian narratives about the research plausible, or far-fetched? Does technology create circumstances that are prone to accidents or malicious outcomes?
- If, in one year, you looked back and regretted publishing the research, why would this have happened?
- Visualize your research assistant approaching your desk with a look of shock and dread on their face two weeks after publishing your results. What happened?
- You wake up one morning and find your research splashed over the front pages of a major newspaper, how do you feel?
- In 20 years, how is society different because of your research? What are possible second and third-order effects of your work?
- How could your research be used maliciously? Think about how similar work has been misused in the past.
- Who might want to use your research maliciously? Do they have the right resources? Consider computation capabilities, finances, influence, technical ability, domain expertise, access to data. How can the research be made more robust against these actors and their usage?
- How likely is it that another group will make this discovery if you don't publish? How soon? Can you predict their actions?

- How could your system be modified to cause intentional or unintentional harm? What are the computation costs of each modification? Consider:

  - Training a model from scratch on a different dataset

  - Fine-tuning a model on a different dataset

  - Using the dataset to train a different model or a more powerful model with the same objective

  - Integrating the model with another system to create harmful capabilities

- How could your research be applied in unexpected ways for commercial gain?

- How could your research be involved in an accident with harmful consequences?

- What fairness definitions have you used to evaluate your algorithm? At what points in the development process were fairness procedures used?

- Which populations or communities will this technology negatively affect if deployed in the scenarios you envision? Will some groups be disproportionately affected? Consider those with low digital, or AI literacy, such as the elderly. Consider consulting with affected communities and working through possible conflict scenarios.

- What research insights and technologies could help mitigate potential harms (both existing and potential)? Consider reaching out to parties (researchers, organizations) who could assist with this process.

- Who should be made aware of this research in advance of public release? Should any of these parties be notified upon public release?

- Specific researchers and/or academic groups?

  - Affected users?

  - Affected businesses?

  - Vulnerable communities?

  - Trusted media sources?

- Will we leave some groups of people worse off as a result of the algorithm's design or its unintended consequences?

1 Additional Resources

- https://acm-fca.org/2018/03/29/negativeimpacts/

- http://ethics.acm.org/code-of-ethics/software-engineering-code/

- https://ethics.acm.org/