UofT AI Ethics Hackathon Resources

When developing AI solutions, it's essential to consider their ethical implications to ensure they are beneficial, fair, and responsible. Below are some key frameworks and guidelines you can use to evaluate your project.

Table of Contents

1. Ethical Matrix

Systematically evaluate the ethical impact of a technology, particularly AI systems, by considering the interests of different stakeholders and the values or ethical principles relevant to them.

Step 1: Identify the Stakeholders

Stakeholders are individuals or groups that are affected by your AI system, both directly and indirectly.

Examples:

- **Users** (the individuals or organizations using the AI system)
- **Developers** (the team creating and maintaining the AI)
- **Regulators** (those setting policies or laws governing AI use)
- **Wider society** (communities or groups indirectly impacted)
- **Environment** (if applicable, considering environmental impacts of AI)

Step 2: Identify the Ethical Values

Values are the ethical principles or concerns that are relevant to your AI project. Consider the key ethical issues related to your AI system. If you're unsure, refer to established AI ethics frameworks (e.g. The Ethics Guidelines for Trustworthy AI, The Asilomar AI Principles, IEEE Ethically Aligned Design, and OECD Principles on AI) to identify relevant values.

Examples:

- **Privacy**: Does the system respect user privacy, and is data collection minimized?
- **Fairness**: Does the system treat all stakeholders equally, and is it free of bias?
- **Transparency**: Is the AI system explainable and clear in how decisions are made?
- **Accountability**: Who is responsible when things go wrong, and is there a clear line of accountability?
- **Safety**: Does the AI operate safely for users and others?
- **Data Usage**: Is data used ethically, and are users aware of how it is being used?
- **Sustainability**: Does the AI promote long-term societal or environmental sustainability?

Step 3: Create the Ethical Matrix

Now, construct a matrix with stakeholders as rows and values (ethical principles) as columns. Each cell in the matrix will capture how a specific stakeholder is impacted by the AI system with respect to a particular value.

Example 1:

	Privacy	**Fairness**	**Transparency**	**Data Usage**
Users	How is user privacy protected?	Are users treated fairly and without bias?	Do users understand how the AI works?	Is user data used ethically?
Developers	Do developers collect data ethically?	Are biases in AI models addressed?	Are the AI models explainable and accountable?	Is data used appropriately for system training?
Regulators	Does the AI comply with privacy laws?	Are the standards applied fairly across sectors?	Can regulators audit or understand the system?	Does data usage comply with legal frameworks?

Wider Society	Does the AI system safeguard societal privacy rights?	Does it promote equality or increase inequality?	Can the public understand the AI's impact?	Is data used for societal benefits or in harmful ways?

Example 2:

Respect for:	Well-Being	Autonomy	Justice
Developer (FakeFinder Team)			
Developer (Model Builder)			
Tool User (Intel. Analyst)			
Video Subjects (in training data)			
Video Subjects (video in the wild)			
Data Owner (Facebook)			

Step 4: Populate the Matrix

For each cell, describe how the AI system impacts the stakeholder with respect to that particular value. You can think of this as an ethical evaluation, and the goal is to understand how different ethical concerns play out across various groups.Use specific examples related to your project. Be honest about potential challenges, benefits, and conflicts of interest for each stakeholder and value pair.

- For **users** and **privacy**, you might note whether the AI system collects sensitive personal data and if it offers users options to control data sharing.
- For **developers** and **fairness**, you could discuss whether the AI system's training data includes bias, and if so, how the development team is working to mitigate it.
- **Note:** Some stakeholders and values may interact in unexpected ways. For example, maximizing transparency for users could impact privacy if revealing too much detail compromises sensitive data.

Example 2:

Respect for:	Well-Being	Autonomy	Justice
Developer (FakeFinder Team)	Creative freedom; recognition for contributions.	Transparency wrt tool functionality, limitations.	Data quality / representativeness.
Developer (Model Builder)	Creative freedom; recognition; compensation for 3rd party profit?	Transparency wrt usage (if not in accordance with OS license?)	Credit for fair use of work.
Tool User (Intel. Analyst)	Increased efficiency; less time on low-level tasks. False negatives?	Automation bias. Choice to adopt/not adopt AI?	Data quality / discriminatory practices.
Video Subjects (in training data)	Compensation for use of likeness.	Informed consent to use of likeness.	Association with discriminatory practices?
Video Subjects (video in the wild)	Privacy? Limitations on 3rd party viewing.	Transparency in use? False positives (real videos labeled 'fake').	Differential validity, discriminatory practices.
Data Owner (Facebook)	Efficiency/cost; success of product; crowd-sourced R&D.	Protection of IP.	Risk to PR/brand.

concern need more information

Step 5: Iteration 2 - Additional Considerations

Once the matrix is filled in, analyze it to identify any ethical issues or conflicts that need to be addressed. Pay attention to any areas where a stakeholder might be negatively affected or where there's a trade-off between different ethical values. The goal is to minimize harm and maximize the benefits for all stakeholders. Highlight cells that need more information or are concerning (see example below)

- Look for cells that highlight potential ethical challenges (e.g., user privacy vs. data usage) and determine if there are stakeholders who are disproportionately affected. Consider whether any value (e.g., fairness, privacy) is under-addressed or compromised. If you make any changes to your project and thus need to update the ethical matrix, include this in your presentation, we love to see your thought process and evolution!
- If you notice that **users** are negatively affected by the system's **fairness**, you might prioritize addressing bias in your AI model.
- If **regulators** have concerns about **transparency**, you may need to improve the explainability of your AI decisions.
- If your system raises concerns about **user privacy**, you could implement stricter data anonymization measures or allow users more control over the information they share.

Example 2:

Stakeholders	False Negatives Fake video labeled real	False Positives Real video labeled fake	True Positives Fake video labeled fake
Developer (FakeFinder Team)	Undermine confidence in tool; people don't use it.	Undermine confidence in tool; people don't use it.	Desired outcome.
Developer (Model Builder)	Undermine confidence in model; reputation harm.	Undermine confidence in model; reputation harm.	Desired outcome.
Tool User (Intel. Analyst)	Miss a harmful deepfake; disinformation, hoax, etc.	Slight inefficiency; unnecessary review of video by human.	Desired outcome.
Video Subjects (in training data)	N/A	N/A.	Data leak without context could cause reputation harm.
Video Subjects (video in the wild)	Reputation harm; rumors, hoaxes spread about subject	Reputation harm; videos removed unnecessarily; accounts disabled.	Subjects who alter videos (i.e. cosmetic reasons) called out as fake?
Data Owner (Facebook)	Undermine confidence in platform; defamation lawsuit?	Undermine confidence in platform; defamation lawsuit?	Data leak without context could cause reputation harm.
Media Consumers	Dissemination of mis/ disinformation.	More difficult to access real/true information.	Blanket policy to remove deepfakes could censor content unnecessarily.
Video Producers (making deepfakes)	Could be desired effect (rumors/hoaxes); confusion; loss of control of intent.	N/A.	Blanket policy to remove deepfakes could censor content unnecessarily.
Video Producers (NOT making deepfakes)	Competition from fake content; undermine confidence in media.	Reputation harm; videos removed unnecessarily; accounts disabled.	Disired outcome? OR altered videos could be called out as fake.

Examples & Resources

1. An "ethical matrix" for FakeFinder | by andrea b | high stakes design | Medium
2. Norwegian fishing industry Ethical Matrix | Module 3 Design Notes

Respect for:	Wellbeing	Dignity	Justice
Fishermen	Safe and secure workplace and income, as well as stable social situation	Right to control of their work situation and respect for their occupation	Equal right to professional practice for different categories of fishermen
Fishing industry	Stable deliveries from the fisheries; a part of the welfare goods obtained in the value chain	Acknowledgement of their place in the value chain: being heard in negotiations.	Equal terms for this industry as for other marine occupations
Other users of the sea and coast	Access to welfare goods directed at marine activities as other users	Respect for their needs and their use of the coast and sea	Equal access to the resources
The society as a whole	Income from marine activities	Freedom to manage resources for the best of society as a whole	Equal living conditions for urban and rural societies
Consumers	Guarantees for healthy food in adequate amounts	Opportunities for the consumer to chose and influence the production of food products	Fish products of good quality available for different consumer groups
Future generations	No activities that threaten their health or living conditions	Knowing that earlier generations acted with respect for their welfare	The conservation of marine environment and resources so that future generations will have the same opportunities we have
The biosphere	That fish and other animals are not exposed to unnecessary pain	Harm and abuse of nature as limited as possible	The diffusion to a viable level of environmental burdens over a variety of ecosystems

Figure 3 A customised version of the ethical matrix designed to assess the future of the Norwegian fishing industry (Kaiser and Forsberg, 2001)

3. [PDF] Ethical Matrix Manual | Semantic Scholar
4. Algorithmic Stakeholders: An Ethical Matrix for AI

---

2. Ethical AI Guidelines

1. **The Ethics Guidelines for Trustworthy AI** (EU Commission)

Trustworthy AI:

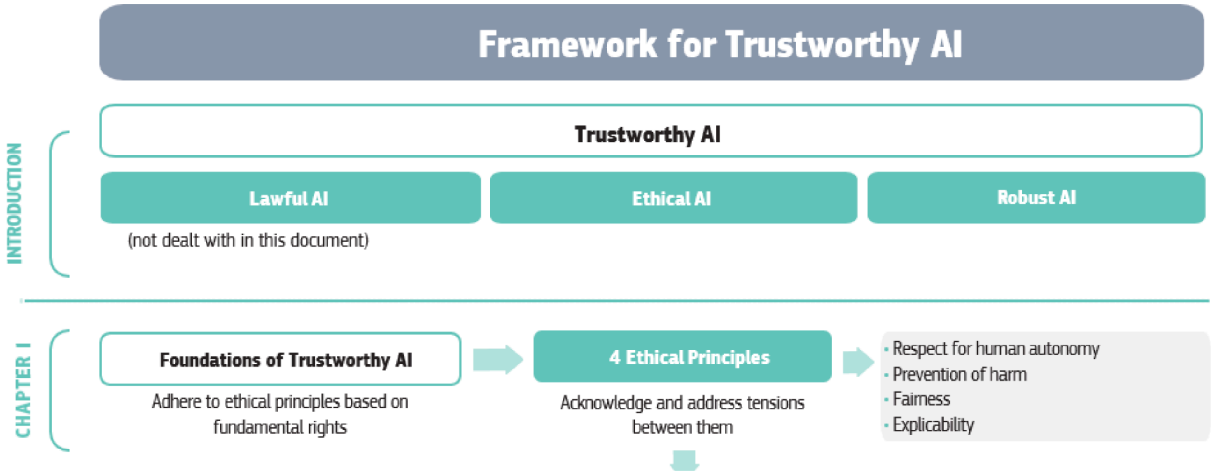
1) it should be **lawful**, complying with all applicable laws and regulations
2) it should be **ethical**, ensuring adherence to ethical principles and values
3) it should be **robust**, both from a technical and social perspective since, even with good intentions, AI systems can cause unintentional harm.

*Focus on **chapter 1** of the ethics guidelines pdf*

Ethical Principles:

1. Respect for human autonomy

2. Prevention of Haarm
3. Fairness
4. Explicability



Example:

2. The Asilomar AI Principles (Future of Life Institute)

These are a set of 23 principles created by AI experts to ensure AI research and development is beneficial to humanity. The principles emphasize safety, transparency, responsibility, and the long-term impact of AI on society.

Key Principles:

- Research should focus on beneficial uses of AI
- AI systems should be transparent and explainable
- AI should respect human rights and values
- AI should be aligned with human goals
- Responsibility should be shared among all stakeholders

Example:
If you're developing an AI that automates job recruitment processes, you should use these principles to ensure that the system is explainable (e.g., it can explain why a certain candidate was or wasn't selected), and that it avoids perpetuating bias against marginalized groups in hiring practices.

3. IEEE Ethically Aligned Design

This framework focuses on designing AI systems that prioritize human well-being. It emphasizes transparency, accountability, and respect for human rights throughout the AI design process.

Key Principles:

- Human rights and well-being
- Accountability in algorithmic decision-making
- Transparency and explainability of AI models
- Awareness of bias and how to mitigate it
- Prioritizing privacy and security

Example:

For an AI system that predicts student success based on academic data, the IEEE framework would suggest being transparent about how the system makes predictions and ensuring it doesn't disadvantage students from certain backgrounds or socioeconomic statuses.

4. OECD Principles on AI

The Organisation for Economic Co-operation and Development (OECD) created these principles to promote innovation while ensuring AI systems are designed in a way that respects human rights and democratic values. The principles include transparency, accountability, and the stewardship of AI development to ensure positive outcomes.

Key Principles:

- AI should benefit people and the planet
- AI systems should be transparent and explainable
- AI should operate in a robust, safe, and secure manner
- Organizations must be accountable for AI's use

Example:

If your project involves an AI-driven app that helps users reduce their carbon footprint, the OECD principles would ask you to ensure that the AI model is both safe and secure, but also transparent—users should understand how their data is being used and how AI-driven recommendations are made.