Chapter 4: AI Ethics: Privacy and Responsibility

1.Setting the Stage

Activity 1: Quick Ethics Snapshots

Reflect on the following AI ethics mini-scenarios.
For each scenario:

1. Identify the ethical issue involved.
2. Determine who holds responsibility for addressing the issue.
3. Incorporate at least one key term in your discussion, such as *privacy*, *consent*, or *accountability*.

Mini-Scenarios:

1. An AI application tracks your location without requesting permission, claiming it's to enhance service quality.
2. A hospital's AI system inadvertently shares patient data. Who is responsible for rectifying the breach?
3. An autonomous vehicle controlled by AI is involved in a crash. Who should be held accountable?

Activity 2: AI Ethics Matching Exercise

Match each AI ethics term in Column A with its correct definition from Column B. Write the corresponding letter next to each term.

Column A: Terms	Column B: Definitions
1. Privacy	A. The capacity to make independent decisions without external control.
2. Consent	B. Responsibility for the outcomes produced by AI systems.

Column A: Terms	Column B: Definitions
3. Bias	C. Ensuring equitable treatment and outcomes across different groups.
4. Transparency	D. Openness about how AI systems operate and make decisions.
5. Accountability	E. Voluntary agreement to allow data collection or processing.
6. Fairness	F. Monitoring individuals or groups, often raising privacy concerns.
7. Surveillance	G. Systematic favoritism or prejudice in data or algorithms.
8. Autonomy	H. The right of individuals to control access to their personal information.
9. Explainability	I. The ability to understand and interpret how AI systems arrive at decisions.
10. Responsibility	J. The duty to act ethically and consider the impact of AI on society.

2.Take a Listen

Activity 1: Ethical AI Principles Matching

Match the following **scenarios** with the 5 key AI ethics principles from the video.

- **Beneficence** (AI should improve well-being)
- **Non-maleficence** (Do no harm)
- **Autonomy** (Preserve human freedom)
- **Justice** (Promote fairness/equity)
- **Explicability** (Explain how AI works)

Scenarios:

- A hospital uses an AI system to prioritize patients for organ transplants.

- A social media algorithm spreads fake news, polarizing society.
- A self-driving car's decision-making process is hidden from users.
- A company uses AI to automate jobs without retraining workers.

Activity 3: AI Ethics Debate: "Who Should Fix It?"

Discuss how the following **4 groups** address these simple AI issues:

- "A robot hiring tool is unfair to women."
- "A face-scanning app makes more mistakes on dark skin."
- "A social media app shows fake news."

1. **Tech Team** (people who build the AI)
2. **Government** (people who make rules)
3. **Users** (people who use the AI)
4. **Ethics Experts** (people who study right/wrong)

2. Read & Reflect

5 Pillars of Responsible Generative AI: A Code of Ethics for the Future

Why Generative AI Ethics Are Important—and Urgent

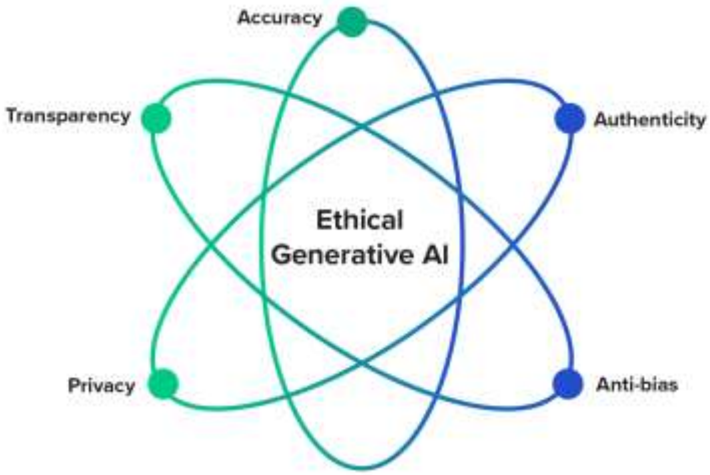
AI ethics and regulations have been debated among lawmakers, governments, and technologists around the globe for years. But recent generative AI increases the urgency of such mandates and heightens risks, while intensifying existing AI concerns around misinformation and biased training data. It also introduces new challenges, such as ensuring authenticity, transparency, and clear data ownership guidelines, says Toptal AI expert Heiko Hotz. With more than 20 years of experience in the technology sector, Hotz currently consults for global companies on generative AI topics as a senior solutions architect for AI and machine learning at AWS.

In particular, the potential inability to determine whether something is AI- or human-generated has far-reaching consequences. With deepfake videos, realistic

AI art, and conversational chatbots that can mimic empathy, humor, and other emotional responses, generative AI deception is a top concern, Hotz asserts. Also pertinent is the question of data ownership—and the corresponding legalities around intellectual property and data privacy. Large training data sets make it difficult to gain consent from, attribute, or credit the original sources, and advanced personalization abilities mimicking the work of specific musicians or artists create new copyright concerns. In addition, research has shown that LLMs can reveal sensitive or personal information from their training data, and an estimated 15% of employees are already putting business data at risk by regularly inputting company information into ChatGPT.

5 Pillars of Building Responsible Generative AI

To combat these wide-reaching risks, guidelines for developing responsible generative AI should be rapidly defined and implemented, says Toptal developer Ismail Karchi. He has worked on a variety of AI and data science projects—including systems for Jumia Group impacting millions of users. "Ethical generative AI is a shared responsibility that involves stakeholders at all levels. Everyone has a role to play in ensuring that AI is used in a way that respects human rights, promotes fairness, and benefits society as a whole," Karchi says. But he notes that developers are especially pertinent in creating ethical AI systems. They choose these systems' data, design their structure, and interpret their outputs, and their actions can have large ripple effects and affect society at large. Ethical engineering practices are foundational to the multidisciplinary and collaborative responsibility to build ethical generative AI.
Building responsible generative AI requires investment from many stakeholders.
To achieve responsible generative AI, Karchi recommends embedding ethics into the practice of engineering on both educational and organizational levels: "Much like medical professionals who are guided by a code of ethics from the very start of their education, the training of engineers should also incorporate fundamental principles of ethics."
Here, Gupta, Hotz, and Karchi propose just such a generative AI code of ethics for engineers, defining five ethical pillars to enforce while developing generative AI solutions. These pillars draw inspiration from other expert opinions, leading responsible AI guidelines, and additional generative-AI-focused guidance and are specifically geared toward engineers building generative AI.



5 Pillars of Ethical Generative AI

1. Accuracy

With the existing generative AI concerns around misinformation, engineers should prioritize accuracy and truthfulness when designing solutions. Methods like verifying data quality and remedying models after failure can help achieve accuracy. One of the most prominent methods for this is retrieval augmented generation (RAG), a leading technique to promote accuracy and truthfulness in LLMs, explains Hotz.

He has found these RAG methods particularly effective:

- Using high-quality data sets vetted for accuracy and lack of bias
- Filtering out data from low-credibility sources
- Implementing fact-checking APIs and classifiers to detect harmful inaccuracies
- Retraining models on new data that resolves knowledge gaps or biases after errors
- Building in safety measures such as avoiding text generation when text accuracy is low or adding a UI for user feedback

For applications like chatbots, developers might also build ways for users to access sources and double-check responses independently to help combat automation bias.

2. Authenticity

Generative AI has ushered in a new age of uncertainty regarding the authenticity of content like text, images, and videos, making it increasingly important to build solutions that can help determine whether or not content is human-generated and genuine. As mentioned previously, these fakes can amplify misinformation and deceive humans. For example, they might influence elections, enable identity theft or degrade digital security, and cause instances of harassment or defamation.

"Addressing these risks requires a multifaceted approach since they bring up legal and ethical concerns—but an urgent first step is to build technological solutions for deepfake detection," says Karchi. He points to various solutions:

- **Deepfake detection algorithms**: "Deepfake detection algorithms can spot subtle differences that may not be noticeable to the human eye," Karchi says. For example, certain algorithms may catch inconsistent behavior in videos (e.g., abnormal blinking or unusual movements) or check for the plausibility of biological signals (e.g., vocal tract values or blood flow indicators).
- **Blockchain technology**: Blockchain's immutability strengthens the power of cryptographic and hashing algorithms; in other words, "it can provide a means of verifying the authenticity of a digital asset and tracking changes to the original file," says Karchi. Showing an asset's time of origin or verifying that it hasn't been changed over time can help expose deepfakes.
- **Digital watermarking**: Visible, metadata, or pixel-level stamps may help label audio and visual content created by AI, and many digital text watermarking techniques are under development too. However, digital watermarking isn't a blanket fix: Malicious hackers could still use open-source solutions to create fakes, and there are ways to remove many watermarks.

It is important to note that generative AI fakes are improving rapidly—and detection methods must catch up. "This is a continuously evolving field where detection and generation technologies are often stuck in a cat-and-mouse game," says Karchi.

3. Anti-bias

Biased systems can compromise fairness, accuracy, trustworthiness, and human rights—and have serious legal ramifications. Generative AI projects should be engineered to mitigate bias from the start of their design, says Karchi.

He has found two techniques especially helpful while working on data science and software projects:

- **Diverse data collection**: "The data used to train AI models should be representative of the diverse scenarios and populations that these models will encounter in the real world," Karchi says. Promoting diverse data reduces the likelihood of biased results and improves model accuracy for various populations (for example, certain trained LLMs can better respond to different accents and dialects).
- **Bias detection and mitigation algorithms**: Data should undergo bias mitigation techniques both before and during training (e.g., adversarial debiasing has a model learn parameters that don't infer sensitive features). Later, algorithms like fairness through awareness can be used to evaluate model performance with fairness metrics and adjust the model accordingly.

He also notes the importance of incorporating user feedback into the product development cycle, which can provide valuable insights into perceived biases and unfair outcomes. Finally, hiring a diverse technical workforce will ensure different perspectives are considered and help curb algorithmic and AI bias.

4. Privacy

Though there are many generative AI concerns about privacy regarding data consent and copyrights, here we focus on preserving user data privacy since this can be achieved during the software development life cycle. Generative AI makes data vulnerable in multiple ways: It can leak sensitive user information used as training data and reveal user-inputted information to third-party providers, which happened when Samsung company secrets were exposed. Hotz has worked with clients wanting to access sensitive or proprietary information from a document chatbot and has protected user-inputted data with a standard template that uses a few key components:

- An open-source LLM hosted either on premises or in a private cloud account (i.e., a VPC)
- A document upload mechanism or store with the private information in the same location (e.g., the same VPC)
- A chatbot interface that implements a memory component (e.g., via LangChain)

"This method makes it possible to achieve a ChatGPT-like user experience in a private manner," says Hotz. Engineers might apply similar approaches and employ creative problem-solving tactics to design generative AI solutions with privacy as a top priority—though generative AI training data still poses significant privacy challenges since it is sourced from internet crawling.

5. Transparency

Transparency means making generative AI results as understandable and explainable as possible. Without it, users can't fact-check and evaluate AI-produced content effectively. While we may not be able to solve AI's black box problem anytime soon, developers can take a few measures to boost transparency in generative AI solutions.

Gupta promoted transparency in a range of features while working on 1nb.ai, a data meta-analysis platform that is helping to bridge the gap between data scientists and business leaders. Using automatic code interpretation, 1nb.ai creates documentation and provides data insights through a chat interface that team members can query.

"For our generative AI feature allowing users to get answers to natural language questions, we provided them with the original reference from which the answer was retrieved (e.g., a data science notebook from their repository)." 1nb.ai also clearly specifies which features on the platform use generative AI, so users have agency and are aware of the risks.

Developers working on chatbots can make similar efforts to reveal sources and indicate when and how AI is used in applications—if they can convince stakeholders to agree to these terms.

Retrieved from: https://www.toptal.com/artificial-intelligence/future-of-generative-ai-ethics, April 12,2025

Activity 4: Pillar Match-Up Challenge

Match each **ethical pillar** of generative AI with its **corresponding practice/example** from the article.

Ethical Pillar	Practice / Example
A. Accuracy	1. Hosting open-source LLMs in private cloud accounts
B. Authenticity	2. Providing source documents in chatbot responses
C. Anti-bias	3. Using adversarial debiasing during model training
D. Privacy	4. Deepfake detection using abnormal blinking detection
E. Transparency	5. Fact-checking APIs and classifiers for LLMs

Activity 5: Spot the Risk

Read the scenarios below and choose the **main risk** that each highlights from the options provided.

Options: A. Misinformation
B. Privacy violation
C. Lack of authenticity
D. Algorithmic bias
E. Lack of transparency

Scenarios:

1. A chatbot shares incorrect medical advice based on outdated sources.
2. An AI image circulates online showing a politician in a fake scandal.
3. A developer uses training data that favors only Western dialects.
4. An employee pastes confidential company data into ChatGPT.
5. A user gets a generative AI result without any reference to the source.

Activity 6: Ethical Engineer's Toolkit

Fill in the blanks using appropriate terms from the list below. Not all words will be used.

Word Bank:
deepfake detection – diverse data – transparency – consent – retrieval augmented generation – blockchain – adversarial debiasing – hallucination – open-source – watermarking – misinformation – empathy

Paragraph:
To ensure **(1)**__________, engineers can use tools like **(2)**__________ or systems that cite sources. Preventing bias starts with collecting **(3)**__________ and using techniques like **(4)**__________. For privacy, some teams rely on **(5)**__________ LLMs and secure environments. To spot manipulated content, techniques like **(6)**__________ and **(7)**__________ are increasingly being adopted.

4. Grammar in Action

HELP BOX: Modal Verbs – Obligation, Advice, Possibility & Permission

Discussions around ethical AI involve a lot of expressions of **obligation** (what we *must* do), **recommendation** (what we *should* do), and **possibility** (what *might* happen if we're not careful). This aligns perfectly with modals.

Function	Modal Verbs	Explanation	Example (related to AI use)
Obligation / Necessity	**must, have to, need to**	Expresses strong necessity, duty, or requirement (legal, moral, or practical)	Developers **must** prevent misuse of generative AI.
			Organizations **have to** follow ethical guidelines.
			Users **need to** evaluate AI outputs critically.
Advice / Recommendation	**should, ought to**	Expresses advice, best practices, or moral suggestions	AI systems **should** be trained on diverse datasets.
			Companies **ought to** promote transparency in AI tools.
Possibility / Risk	**might, could, may**	Indicates something that is possible but not certain	Generative AI **might** produce biased or false content.
			Misuse of AI **could** lead to misinformation.
			Users **may** not fully understand AI-generated data.
Permission / Ability	**Can**	Expresses permission or general ability	Developers **can** set limitations on AI-generated outputs.

Activity 7: Fill in the blanks

Fill in the blanks with the correct modal verb

Options: must, have to, should, can, might, could, need to

1. AI developers ______ check their models for bias before deployment.
2. Tools like ChatGPT ______ be used for educational purposes.
3. If not used carefully, AI ______ spread misinformation.
4. We ______ protect user data from unauthorized access.
5. You ______ always double-check AI-generated content.
6. Engineers ______ disclose when content is AI-generated.
7. Some users ______ misunderstand the capabilities of AI.

Activity 8: Match the Modals to Their Use

Match the modals with their use

A. Modals	B. Uses
1. must	a. advice or best practice
2. should	b. possibility
3. can	c. obligation
4. might	d. permission or ability

Activity 9: Modal Verbs in AI Applications

Read each sentence carefully and choose the correct modal verb to complete the sentence.

1. Developers ____________________ (must / may / can) ensure AI systems are transparent in their decision-making processes.
2. Organizations ____________________ (might / should / ought to) prioritize data protection and security when implementing AI technologies.
3. AI models ____________________(must /can / may) be trained on diverse datasets to improve their accuracy.
4. Users ____________________ (should / might / need to) evaluate AI-generated content critically to detect potential biases.
5. Companies ____________________ (can / must / ought to) provide clear explanations of AI decision-making processes.
6. AI systems ____________________ (could / may / must) produce biased results if trained on unbalanced datasets.
7. Developers ____________________ (should / ought to / might) implement measures to prevent AI-generated content from spreading misinformation.

8. Users ____________________ (can / may / need to) use AI-generated data for decision-making, but must also verify its accuracy.
9. AI applications ____________________ (might / could / must) be used to improve customer service, but only if designed and implemented responsibly.
10. AI systems ____________________ (can / should / must) be designed to explain their decisions and provide transparent results.

5. Speak & Debate

Activity 10: The Line Between Help and Harm

Answer questions on the following scenarios:

- Is this ethical or unethical?
- Does it respect or violate privacy?
- Is the company acting responsibly?

Scenarios:

1. An AI app tracks users' moods by analyzing voice tone and sells emotional data to marketers.
2. A school uses AI to monitor students' eye movement during online exams.
3. A healthcare chatbot shares symptom data with insurance companies.
4. A city uses AI to predict crime in "high-risk" neighborhoods and sends more police.

Activity 11: Ethical Compass – What Would You Do?

Scenario:

A tech company develops an AI tool that can detect signs of mental health issues (such as anxiety, depression, or burnout) by analyzing students' emails. The university is considering deploying it to support student well-being—but without informing students that their emails are being analyzed.

Discussion Questions:

1. **What would you do if you were:**
   - A university decision-maker?
   - A student rights advocate?

- A mental health professional?

2. **What ethical values are at stake?**
   - Consider **privacy**, **consent**, **beneficence**, **autonomy**, **transparency**, and **data protection**.

5. Writing for Impact

Essay Question:

What are the ethical implications of Artificial Intelligence (AI) on individual privacy, and why is responsible development and deployment of AI technologies essential?