

Putting AI to Work

# 15

## Human Oversight and Harm

# Learning Objectives

- Evaluate the importance of human supervision in AI decision-making, particularly in sensitive or high-stakes situations
- Demonstrate ethical judgment by applying personal accountability to the use and consequences of working with AI-generated content
- Identify the potential risks of AI misuse, including the generation of misinformation, violent content, or illegal materials
- Develop strategies to reduce the likelihood of the malicious, unethical, or irresponsible use of AI tools

# Module 15.1: Human Oversight

- AI systems can process information quickly but lack judgment, empathy, and context awareness.
- Without human oversight, AI may make decisions that are flawed, unfair, or harmful.
- Human supervision is critical in high-stakes environments such as healthcare, criminal justice, education, and finance.
- Humans must remain "in the loop" to review, interpret, and override AI decisions.
- AI tools should assist professionals and never replace them in making final decisions.

# Module 15.1: Examples of Human Oversight

- Hiring: AI screens résumés, but human recruiters review final candidates for fairness and diversity.
- Medical: Doctors use AI scan analyses as a second opinion but rely on clinical judgment for diagnosis.
- Loans: AI recommends applicants, but loan officers review for special circumstances not found in the data.
- Content moderation: AI flags potentially harmful content, but humans make the final decisions related to content removals.

# Module 15.1: Ethics in Action

- Blindly trusting AI can lead to unethical or discriminatory decisions.
- Human oversight ensures fairness, transparency, and accountability in decision making.
- Skipping human review in high-risk areas may violate laws or organizational policies.

# Module 15.1: Techie Dive

- Three categories of human oversight:
  - Human-in-the-loop: A person is actively involved in every decision (approving AI-suggested actions).
  - Human-on-the-loop: A person monitors AI activity and can intervene if needed.
  - Human-out-of-the-loop: AI acts autonomously without real-time human input.
- More sensitive tasks require staying closer "in the loop."

# Module 15.1: Business Lens

- Using AI without proper review exposes business to legal, financial, and reputational risk.
- Oversight ensures decisions meet regulatory standards and align with company values.
- Proper oversight boosts trust with customers, employees, and the public.

# Module 15.2: Individual Responsibility

- Users are responsible for AI outcomes since AI generates without understanding truth or harm.
- Individual responsibility means being aware of how your choices affect others.
- AI may assist you but does not replace your ethical judgment.
- "AI did it" is not a valid excuse for harmful outcomes.



# Module 15.2: Individual Responsibility (cont.)

Responsible use principles:

- Don't assume AI is right: Always double-check facts, especially for important decisions.
- Use AI with integrity: Avoid cheating, impersonation, or spreading misinformation.
- Take ownership of your actions: You're responsible for understanding and validating results.

# Module 15.2: Individual Responsibility (cont.)

Examples of responsibility:

- Academic: A student uses AI for brainstorming but rewrites the information in their own words with proper citations.
- Workplace: An employee uses AI to summarize emails but reviews the outputs carefully before sharing.
- Creative: An artist discloses AI assistance and respects copyright laws when sharing work.
- Ethical dilemma: A developer notices biased chatbot answers and reports the issue for correction.

## Module 15.2: Individual Responsibility (cont.)

- Ethical use means thinking beyond what's possible to asking what's right.
- Ask yourself: "Would someone affected by this still feel it's fair and honest?"
- Just because AI makes something easy doesn't make it ethical.

## Module 15.2: Techie Dive

- AI tools don't "know" what's ethical; they follow patterns from training data.
- Biased or deceptive data leads to misleading or harmful results.
- Developers work to build responsible models, but users must make thoughtful decisions.

## Module 15.2: Business Lens

- Companies must train employees to use AI ethically.
- A single AI-generated mistake can hurt a brand's reputation significantly.
- Organizations should clarify who is responsible for outputs and encourage an ethical review culture.

# Module 15.3: Harmful and Dangerous Content

- AI can be misused deliberately or can accidentally produce harmful content.
- Types of harmful content:
  - False information
  - Illegal instructions
  - Violence/hate promotion
  - Stereotypes
- Guardrails and filters help but aren't perfect.
- Human oversight and ethical awareness remain essential for preventing harm.

# Module 15.3: Harmful and Dangerous Content (cont.)

How AI content becomes harmful:

- Prompt misuse: Users can craft prompts to bypass filters or produce banned content.
- Hallucinations: AI may fabricate facts, dates, or sources that sound plausible but are false.
- Unintentional harm: Innocent prompts can lead to stereotypes or offensive phrasing.
- Amplification: Repeated exposure to harmful ideas can normalize them if unchecked.

# Module 15.3: Harmful and Dangerous Content (cont.)

Examples of harmful content:

- Medical misinformation: AI suggests incorrect medication dosage, risking user health.
- Violent instructions: AI complies with a request for dangerous how-to information.
- Offensive descriptions: AI generates characters based on racial or gender stereotypes.
- Fake news: AI creates a realistic article with false information that could spread online.
- Deepfake audio: A synthetic voice mimics a public figure making statements they never made.



## Module 15.3: Ethics in Action

- Generating dangerous content, even unintentionally, can lead to real-world harm.
- Users must think critically about prompts and review outputs before sharing them.
- Developers must build safety systems, and users must report misuse responsibly.

## Module 15.3: Techie Dive

- Most AI tools use moderation layers trained to recognize harmful outputs.
- Filters use classifiers and pattern matching, but users find ways to "jailbreak" them.
- Safety updates aim to close gaps, but bad actors often move faster than safeguards.
- RLHF helps models learn appropriate responses but isn't foolproof.

## Module 15.3: Business Lens

- A single inappropriate or false output could trigger legal, ethical, or reputational damage.
- Businesses must vet AI-generated content before publishing it.
- Strict employee guidelines are essential for risk management.
- Liability may fall on company even if AI was the source of harmful content.

# Module 15.4: Preventing Misuse

- Misuse includes:
  - Offensive content
  - Unethical automation
  - Fake profiles
  - Deepfakes
  - Overreliance
- The best defense combines strong policies, good judgment, and transparency.
- Prevention requires both individual responsibility and organizational systems.

# Module 15.4: Preventing Misuse (cont.)

## Prevention strategies:

- Set clear guidelines: Define acceptable and unacceptable AI use at school or work.
- Encourage human review: Make checking and confirming AI content a standard habit.
- Use trusted tools: Choose reputable platforms that are transparent about safety and moderation.
- Practice prompting with purpose: Design prompts thoughtfully with ethical outcomes in mind.
- Educate and report: Train users on risks and report harmful outputs to platforms.

# Module 15.4: Preventing Misuse (cont.)

Misuse examples and prevention:

- Fake reviews: A user writes false 5-star reviews.  
Prevention: Implement detection features and enforce terms of service.
- Inappropriate student content: A student creates offensive material. Prevention: Implement digital conduct rules and ethics education.
- Filter bypassing: Users reword prompts to trick AI.  
Prevention: Implement refined filters and warnings about rule breaking.
- Biased automation: A business autoapproves applicants without auditing. Prevention: Implement human review and fairness testing.

## Module 15.4: Ethics in Action

- Using AI to deceive, harm, or manipulate crosses ethical lines even if done as a joke.
- Misuse damages relationships, reputations, and trust in technology.
- Ethical AI use starts with integrity and a clear understanding of its impact.

## Module 15.4: Techie Dive

- Some AI systems log user behavior to detect misuse patterns.
- Repeated problematic prompts may flag accounts for review.
- RLHF helps models learn safe and appropriate responses but needs user support.



## Module 15.4: Business Lens

- Organizations face legal and reputational risks from AI tool misuse.
- Employee training, review protocols, and internal audits are essential safeguards.
- A single incident of misuse can erode customer trust and invite legal scrutiny.

# Key Takeaways

- Human oversight is essential because AI lacks judgment, empathy, and ethical reasoning.
- There are three oversight levels: in-the-loop, on-the-loop, and out-of-the-loop. Higher stakes require tighter oversight.
- Individual users are personally accountable for AI outcomes regardless of intent.
- Responsible use requires fact-checking, integrity, giving proper credit, and considering impact on others.
- AI can generate harmful content through misuse, hallucinations, bias, or dangerous amplification.
- Safety filters help but aren't perfect, so human review is critical before sharing outputs.
- Prevention requires clear guidelines, trusted tools, purposeful prompting, and reporting.
- Organizations must implement training, policies, and review processes for risk management.
- The best defense combines strong policies, good judgment, and organizational transparency.
- Ethical AI use asks: Is this fair, honest, and considerate of its impact on others?