AI Ethics & Safety Guide

Use AI responsibly and effectively - GiggaDev Learn

Understanding AI Bias

AI systems learn from human-created data, which means they can inherit and amplify human biases. Understanding these biases helps you use AI more critically and responsibly.

Where AI Bias Comes From

Training Data Bias
If AI learned mostly from Western, English-language content, it may not represent other cultures accurately. Historical data contains historical prejudices.

Selection Bias
Internet content over-represents certain demographics. AI trained on this data may default to assumptions that don't apply universally.

Confirmation Bias
AI tends to agree with your framing. If you ask a leading question, you'll often get the answer you implied. This isn't intelligence—it's pattern matching.

Common Bias Examples

Bias Type	Example
Gender	Assuming "doctor" is male, "nurse" is female
Cultural	Defaulting to US/Western perspectives on topics
Age	Stereotyping elderly as technology-averse
Recency	Over-weighting popular/recent opinions

How to Counter Bias

- Ask for multiple perspectives explicitly: "Consider viewpoints from different cultures/backgrounds"
- Challenge assumptions: "What assumptions are you making in this answer?"
- Specify context: Include relevant details about your situation to avoid defaults

Privacy & Security

NEVER SHARE WITH AI:

- Passwords, API keys, or authentication tokens
- Social Security numbers, government IDs, or financial account numbers
- Medical records or health information (HIPAA concerns)
- Confidential business data, trade secrets, or proprietary code
- Other people's personal information without consent
- Internal company communications or unreleased products

How AI Companies Use Your Data

Training Data (Default for most free tiers)

Your conversations may be used to improve AI models. Anything you share could influence future AI behavior or be reviewed by human trainers.

Conversation History

Most services store your chat history. Check settings to disable history or use "temporary" chat modes when discussing sensitive topics.

Third-Party Apps

AI-powered browser extensions, plugins, and apps may have different privacy policies than the main AI service. Read terms carefully.

Privacy Settings Checklist

- [ ] Review AI service privacy policy and data usage terms
- [ ] Opt out of training data collection where possible
- [ ] Enable temporary/incognito chat modes for sensitive work
- [ ] Use enterprise/business tiers for confidential work (better data protection)
- [ ] Regularly delete chat history you no longer need
- [ ] Use pseudonyms/placeholders for real names and specifics

Pro Tip: Replace real data with "[COMPANY NAME]", "[EMPLOYEE]", "[AMOUNT]" placeholders, then substitute your actual data in the final output.

Verification & Fact-Checking

AI generates plausible-sounding text, but it doesn't "know" facts—it predicts likely word sequences. This means it can confidently state things that are completely false.

Verification Hierarchy

Priority	Source Type
Highest	Primary sources (original documents, official records)
High	Peer-reviewed journals, established institutions
Medium	Reputable news organizations, expert analysis
Lower	Blogs, social media, user-generated content
Lowest	AI-generated content (verify independently!)

Red Flags That Suggest Hallucination

- Very specific statistics or percentages (e.g., "Studies show 73.4% of users...")
- Citations with complete author names, journal names, and DOIs
- Claims about events after AI's knowledge cutoff date
- Detailed technical specifications that seem "too perfect"
- Historical anecdotes with specific quotes and dates
- Confident answers about obscure or niche topics

Verification Strategies

For Facts & Statistics:

- Search for the exact claim or citation in Google Scholar
- Check primary sources (company websites, government databases)
- Ask AI: "Can you provide sources for this?" (then verify those sources exist)

For Code:

- Always run and test AI-generated code before using in production
- Verify function signatures against official documentation
- Check for deprecated methods or version-specific syntax

Ethical Use Guidelines

Be Honest About AI Assistance

When appropriate, disclose that you used AI. Academic work, journalism, and professional contexts often require transparency about AI involvement.

Maintain Accountability

You are responsible for AI output you use. "The AI told me" is not a valid excuse for errors, biased content, or harmful information.

Respect Intellectual Property

AI may reproduce copyrighted material. Don't use AI to plagiarize, circumvent paywalls, or generate content that infringes on others' rights.

Consider Impact on Others

Don't use AI to deceive, manipulate, or harm. This includes deepfakes, impersonation, spam, or content designed to mislead.

Human-AI Partnership

AI Excels At	Humans Excel At
Generating drafts quickly	Making final judgment calls
Brainstorming many options	Choosing the right option for context
Explaining concepts simply	Understanding nuance and subtext
Processing large texts	Reading between the lines
Following instructions precisely	Knowing when to break the rules

The Golden Rule of Human-AI Collaboration:

Use AI as a first draft generator, research accelerator, and brainstorming partner—but always apply human judgment before finalizing anything important. AI is a tool that amplifies your capabilities, not a replacement for your expertise and critical thinking.

Remember: The goal isn't to use AI for everything. It's to use AI where it helps and humans where humans are better—creating results neither could achieve alone.