CURATED STREAM

Ethical Considerations in AI

Hallucinations, bias, toxicity, regulation — and what to do about them.

PART 01 — INTRODUCTION

Why ethics in AI matters

AI is now everywhere — so the ethics matter more, not less

Artificial intelligence sits inside everyday tools — search, writing, coding, design, healthcare, hiring, and government services. As the technology has scaled, so have the ethical questions about how it is built and where it is deployed.

This stream walks through four interconnected concerns — hallucinations, bias, toxicity, and the collective impact — and the principles and rules now in force to manage them.

1 Aug 2026
EU AI Act fully applicable for high-risk systems

+32%
rise in malicious prompt-injection attempts (Nov 2025 - Feb 2026)

97%
of breached organisations lacked proper AI access controls (IBM 2025)

What we'll cover

01

General ethics

Fairness, accountability, transparency, responsibility.

02

Hallucinations

False or fabricated content — and why models still produce it.

03

Bias

How data and design can entrench discrimination.

04

Toxicity

Hate speech, harassment, and recommender harms.

05

Collective impact

How these issues compound — and the regulatory response.

PART 02 — PRINCIPLES

General ethics

Four principles for responsible AI

Ethics is built in at every stage of the AI lifecycle, from design to deployment and beyond.

Fairness

Outcomes that do not systematically disadvantage groups by race, gender, age, or socioeconomic status.

Accountability

Clear lines of responsibility for what an AI system decides, recommends, or generates.

Transparency

Users understand when AI is being used, on what data, and with what limits.

Responsibility

Designers, deployers, and users each have duties — not just the model.

Principles in practice

Healthcare AI

FDA-authorised AI medical devices have passed 1,250 listings (Jul 2025). Radiology dominates with around 956 devices.

Patient privacy: data handling, retention, consent.

Equitable access: models that perform across demographics, not only the training majority.

Clinician oversight: AI assists diagnosis; the clinician remains accountable.

Autonomous vehicles

Safety, liability and the protection of human life — engineering choices with direct ethical weight.

Safety: how the system behaves under edge-case sensor failures.

Liability: who answers when a fully autonomous vehicle is at fault.

Transparency: auditable decision logs for regulators after an incident.

PART 03 — FABRICATED CONTENT

Hallucinations in AI

What we mean by 'hallucinations'

A hallucination is content generated by AI that sounds authoritative but is false, fabricated, or manipulated.

Two overlapping families:

Confabulation: an LLM invents facts, citations, or quotes that do not exist.

Synthetic manipulation: deepfake video and audio, fabricated news articles, manipulated images of public figures.

Both blur the line between fact and fiction, and both are getting cheaper to produce. The AI Act Omnibus (political agreement July 2026) adds new rules specifically on AI-generated intimate content.

Deepfake-style portrait collage

Why LLMs still make things up

Sequence completion

LLMs predict the next plausible token, not the most truthful one. Fluency is rewarded; accuracy is a side-effect.

Training-data gaps

Models fill silence with patterns from elsewhere. Where coverage is thin, confabulation is the default.

No grounding by default

Without tools like web search, retrieval, or citation, a model has nothing to check itself against.

What's changed in 2026: Frontier models (Claude Fable 5 / Opus 4.8, GPT-5.6, Gemini 3.5 Flash) reduce hallucination rates by combining longer context windows, retrieval, and tool use. They are better — not perfect. Always verify load-bearing facts.

PART 04 — FAIRNESS & EQUITY

Bias in AI

Where bias creeps in

Bias in AI is not a single problem. Each stage of the pipeline can encode assumptions that produce discriminatory outcomes.

The harms are concrete:

- Criminal justice — risk-assessment tools shown to produce harsher recommendations for minority defendants.
- Hiring — recruitment screeners that downweight female candidates because their training data favoured male hires.
- Healthcare — diagnostic models trained on majority populations that under-detect conditions in others.
- Credit and insurance — proxy variables (postcode, schooling) standing in for protected attributes.

Three places bias enters

Data bias

Training data over- or under-represents groups.

Algorithmic bias

Model architecture or loss function amplifies skew.

Deployment bias

A model used outside the population it was built for.

PART 05 — HARMFUL CONTENT AT SCALE

Toxicity in AI

Toxic content at platform scale

Illustration: 'toxic' content streams

The problem

Hate speech, harassment, and misinformation now spread at platform speed. Recommender systems can amplify the most engaging — not the most accurate — content, producing radicalisation pathways and polarisation.

Where AI sits in this

Generation: GenAI lowers the cost of producing toxic content at scale.

Recommendation: ML ranking decides what each user sees next.

Moderation: AI classifiers do most of the filtering — and inherit their own biases.

PART 06 — HOW HARMS COMPOUND

The collective impact

Compounding harms — and the response

Hallucinations

Fabricated content erodes trust in any information source.

Bias

Skewed outputs harden stereotypes and discrimination.

Toxicity

Harmful content at scale damages mental health and discourse.

Eroded public trust

When harms reinforce each other they undermine media, institutions, and the legitimacy of AI itself — making it harder for safe systems to earn trust.

The 2026 response

EU AI Act: fully applicable 2 Aug 2026 — conformity assessments, technical documentation, CE marking, and EU database registration for high-risk systems.

NZ: Responsible AI Guidance for the Public Service (GenAI Foundations) is the official skills and capability framework.

Closing thought

Ethics is not a constraint on AI — it is the condition for AI being worth using.

Addressing hallucinations, bias, and toxicity is no longer optional in 2026 — the regulation has arrived, the threat landscape has matured, and the public's tolerance for unsafe systems is lower than it has ever been.

Design with the four principles

Verify load-bearing AI outputs

Track regulation actively

Useful reading

Aotearoa New Zealand

Responsible AI Guidance for the Public Service: GenAI Foundations
NZ Digital Government — official capability framework.

NZ AI Strategy — Investing with Confidence
MBIE — identifies education as a key barrier; projects AI could lift NZ GDP by up to $108 billion by 2038.

Microsoft National AI Skilling Partnership
Government, industry, tertiary and community providers.

International

EU AI Act
Entered into force 1 Aug 2024; fully applicable for high-risk systems from 2 Aug 2026.

OWASP Top 10 for LLM Applications (2025)
Prompt injection ranks #1.

Australian Government — Voluntary AI Safety Standard
Ten guardrails for organisations deploying AI.

Curated Stream refreshed July 2026