



AI Ethics & Safety

Kevin Crook

[KevinCrook.com](https://kevincrook.com)



Privacy

If you are not buying something,
then you are the product.

Check your AI Chatbot's privacy policies
and settings (often varies by paid tier).

Avoid pasting sensitive information into an
AI Chatbot.



Environmental Impacts

Training and running AIs uses a lot of electricity and water.

The AI race has worldwide utilities recommissioning and extending coal plants and building new fossil and nuclear plants.



Bias

Garbage In / Garbage Out (GIGO).

Biased training data leads to
biased answers.



Copyright Violation

Als are sometimes trained on copyrighted material (books, articles, art, music), usually without permission or royalties.

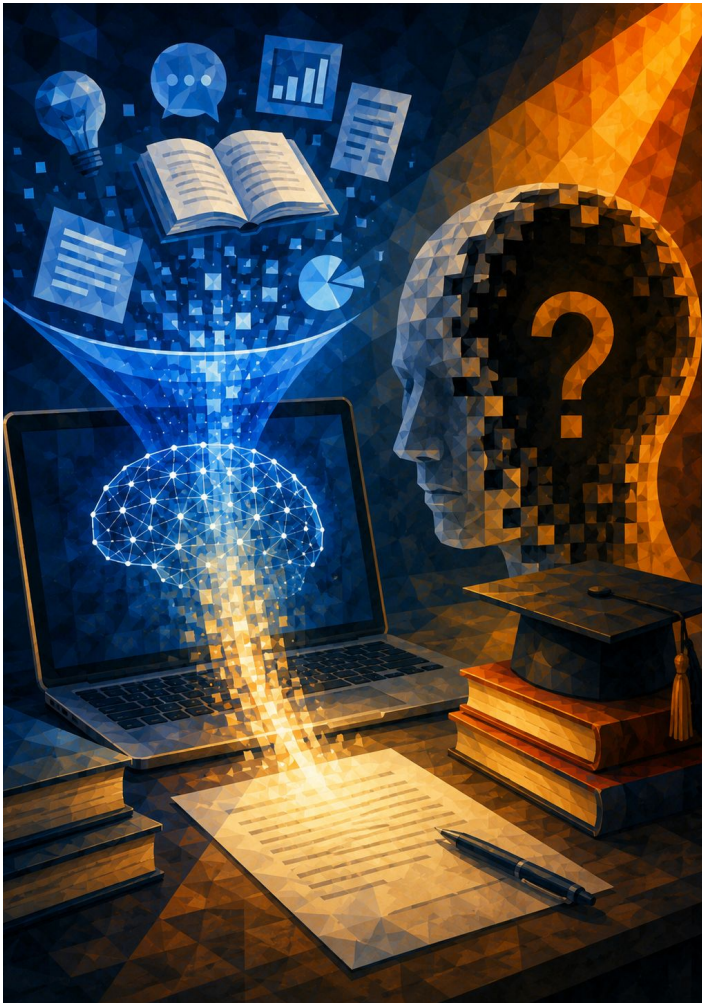
Als can sometimes reproduce copyrighted content in their responses.

Ongoing lawsuits and settlements are deciding what's legal, with the rules still being written.



Echo Chambers / Filter Bubbles

Als can reinforce your existing beliefs by agreeing too easily or feeding you similar perspectives.



Academic Cheating

When prohibited, using AIs to do school work.

Important question:
Are students learning?

Note that AIs can help learning if used properly.



Self-Harm and Harm to Others

LLMs should have guardrails that refuse to help with self-harm, suicide, harming others, dangerous acts, etc.



Workforce Realignment & Job Displacement

Basic AI skills are increasingly required for many roles at all levels.

AI is automating tasks and making some jobs obsolete, especially entry-level.



Autonomous Weapons

Weapons that can select and attack targets with little or no human control.

Big debate: should a machine ever decide to take a human life?

AI arms race.



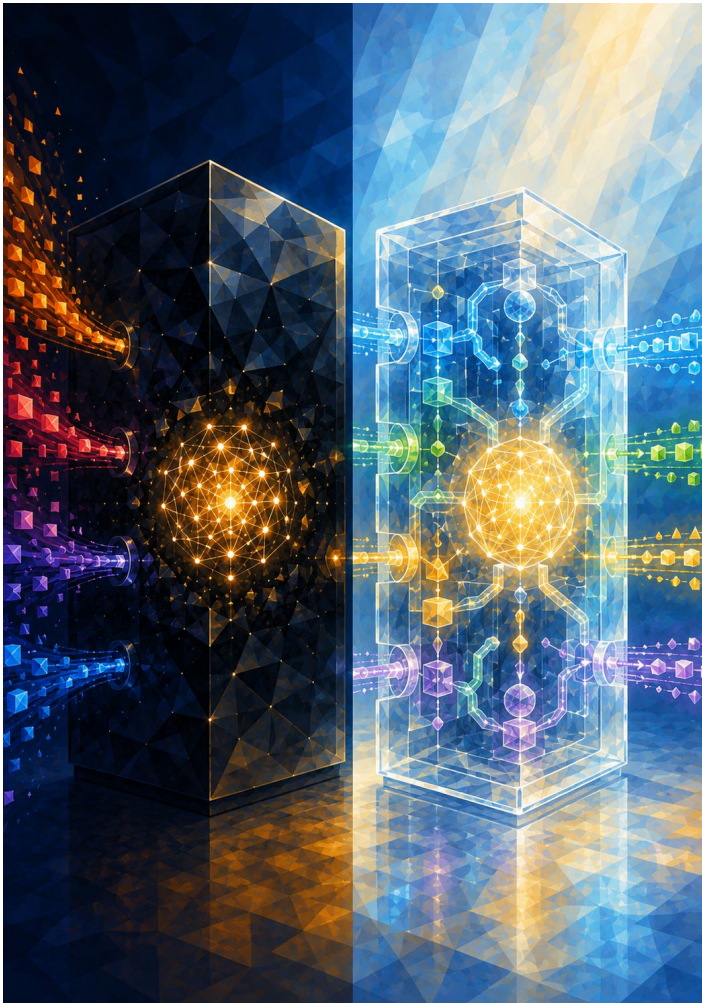
Cybersecurity

AI Helps Attackers

More convincing phishing emails, voice cloning scams, malware assistance, finding software vulnerabilities.

AI Helps Defenders

Detecting attacks, spotting fraud, finding and fixing vulnerabilities faster.



Black Box AI

The norm, no transparency,
We don't see how it reaches answers.

White Box AI

Rare, transparent.
We see how it reaches answers.

Explainable AI (XAI)

XAI is white box AI that clearly shows how it reaches answers.



False Accusations

Claims such as:

- Credit / Loan / Insurance / Redlining Decision
- Detecting AI Use in Writing
- Detecting Lies
- Guiding Prison Sentencing / Probation

Black box, no transparency.

Accuses but can't explain why to allow humans to defend themselves.

Vendors refuse to submit their AIs to formal, independent, scientific scrutiny.



Intelligence Explosion

Humans create an AI smarter than themselves.

That AI creates a new AI smarter than itself.

Process repeats, with each AI generation creating a new AI smarter than itself.

Each AI generation arrives faster than the last AI generation.



Asimov's Three Laws of Robotics

- 1) A robot may not injure a human being or, through inaction, allow a human being to come to harm.
- 2) A robot must obey orders given by human beings except where such orders would conflict with the First Law.
- 3) A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.



AI Dystopia



AI Dystopia

Rossum's Universal Robots (R.U.R.):

Coined the term robot and features a robot uprising against humans.



AI Dystopia

Paperclip Maximizer:

Harmless goal of making paperclips accidentally destroys the world.



AI Dystopia

Tale of the Omega Team:

AI that takes over the world's computing resources with no physical robots involved.



AI Ethics & Safety

Kevin Crook

KevinCrook.com

The End