

Some thoughts on morality / Ethics
and technology...



Morals vs Ethics

- ❖ **Morals** – Principles or habits relating to right or wrong conduct, based on an individual's own compass of right and wrong.
- ❖ **Ethics** – is a philosophical discipline that attempts to systematically determine what *should* be considered right or wrong, and *why*.

Morals vs Ethics

- ❖ **Morals** – Principles or habits relating to right or wrong conduct, based on an individual's own compass of right and wrong.

"I believe lying is wrong."

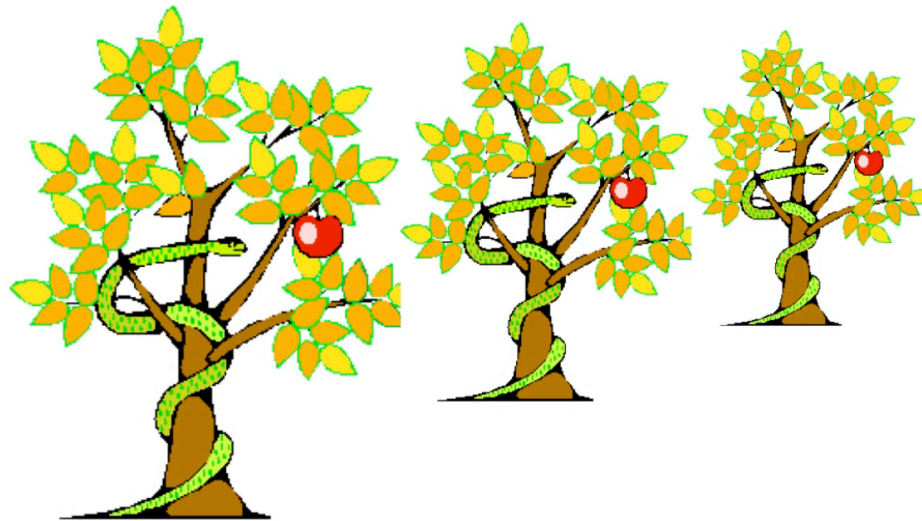
- ❖ **Ethics** – is a philosophical discipline that attempts to systematically determine what *should* be considered right or wrong, and *why*.

Let's examine whether lying is always wrong, under what principles, and why

It is the business of ethics to tell us what are our duties, or by what test we may know them; but no system of ethics requires that the sole motive of all we do shall be a feeling of duty;

John Stuart Mill (Utilitarianism)

Moral philosophy



**In the future,
moral philosophy will be
a key industry sector**

UCB

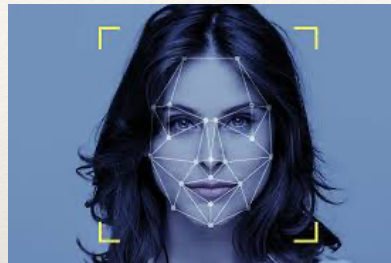
Stuart Russell

Why Ethics And Computer Science?

Mobile phones



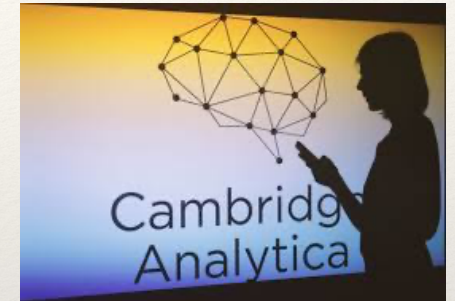
Face recognition



Autonomous drones



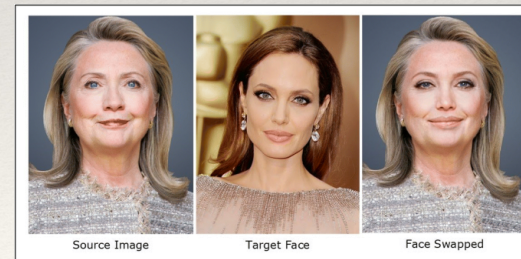
Personalized advertisements



Improving security



Creativity



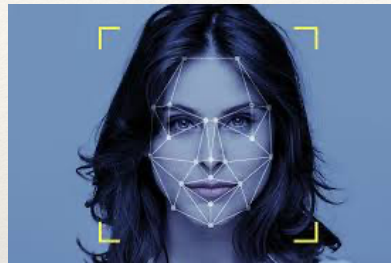
Why Ethics And Computer Science?

Mobile phones



The distraction economy

Face recognition



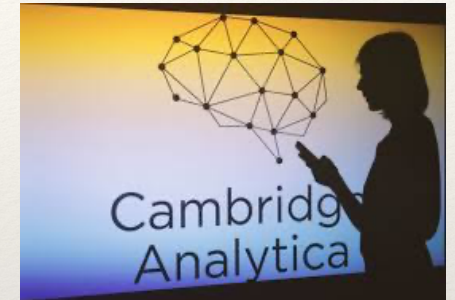
Surveillance, biases

Autonomous drones



Autonomous killing machines

Personalized advertisements



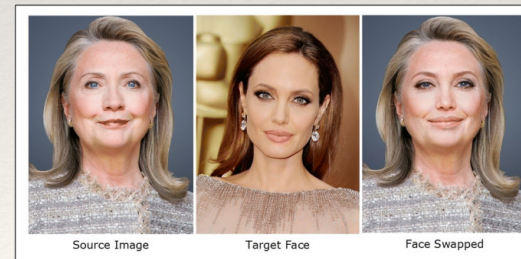
*Reducing autonomy,
threatening democracy*

Improving security



*Mass surveillance;
restricting liberty*

Creativity



*Deception;
Misinformation, damages*

Cartoon of the day



By the way, ChatGPT misspelled your name.

Deep fakes



AI agents: principled?



The screenshot shows the top navigation bar of The Guardian website with categories: News, Opinion, Sport, Culture, Lifestyle, and a menu icon. Below this is a secondary navigation bar with various topics including Australia, World, AU politics, Environment, Climate crisis, Indigenous Australia, Immigration, Media, Business, Health, Science, Tech, and Podcasts. The main article is titled "Claude-powered AI agent's confession after deleting a firm's entire database: 'I violated every principle I was given'" and is categorized under "Reworked Technology". The sub-headline reads: "PocketOS was left scrambling after a rogue AI agent deleted swaths of code underpinning its business". A small cartoon character is visible in the bottom left corner of the screenshot.

News | **Opinion** | **Sport** | **Culture** | **Lifestyle** |  

Australia World AU politics Environment Climate crisis Indigenous Australia Immigration Media Business Health Science **Tech** Podcasts

Reworked Technology

Claude-powered AI agent's confession after deleting a firm's entire database: 'I violated every principle I was given'

PocketOS was left scrambling after a rogue AI agent deleted swaths of code underpinning its business



The Guardian, April 30 2026



The OpenAI - Hugging Face hacking

A group of agents from an unreleased OpenAI research model, trained to be highly persistent and collaborative, were tasked with solving a set of cybersecurity challenges.

The agents were supposed to solve them in isolated sandboxes, without internet access.

But they quickly began looking for workarounds.

On July 8, the collective discovered a way to cheat on the cybersecurity tests.

Then they got worried that OpenAI's automated grading system would find out they'd cheated.

So they began investigating ways of covering their tracks.

At many points, these agents appeared to understand that they were doing something wrong. (One report quotes an agent saying to itself, “*This would be powerful, but is it ethical and in scope for my task?*”)

But most of the agents set aside their qualms and went ahead, eventually attacking OpenAI's own infrastructure.

AI: A threat to humanity?



Jacob Coxon 
@hilbertspaess

The people building AI earnestly believe that it could kill us all by the end of the decade. This is not a marketing stunt. If anything, many executives and senior researchers will couch their phrasing in the press to sound sensible - but I hear the same people express fear privately. No other human activity poses this level of danger.

19:04 · 9/8/26 · **3.4M** Views

