SAINTS

UKRI AI Centre for Doctoral Training in Safe AI Systems

UNIVERSITY of York

AI & Ethics

Permissibility, Sub-optimality, and Personal Responsibility

Prof Tom Stoneham
Ethics Lead for Safe AI Systems

Predicting Humans with AI

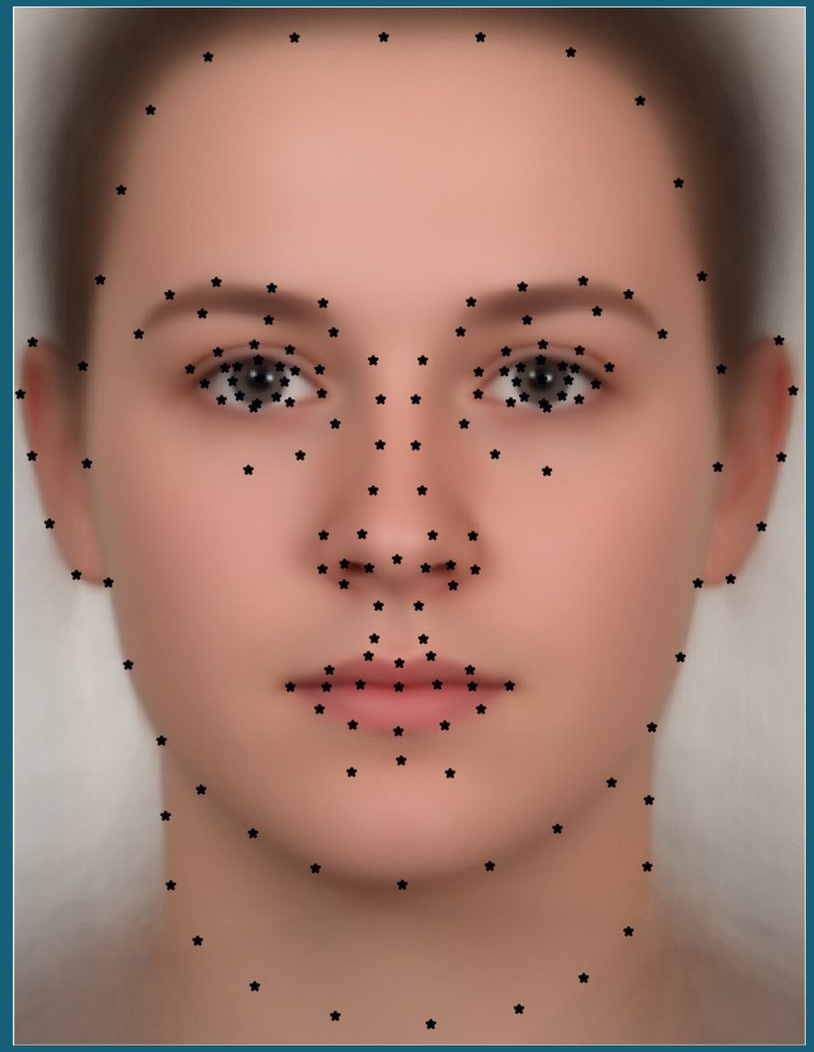
1. Predicting <u>population behaviours</u>
   E.g. traffic flow, consumer choices, voting

2. Predicting <u>individual behaviours</u>
   E.g. targeted advertising, loan applications, recidivism

3. Predicting <u>responses to questions or instructions</u>
   E.g. chatbots, text or search summaries, transcripts

In all cases, prediction is of what has highest probability given training data and input

2 (and sometimes 3) can be biased and reduce diversity to the average

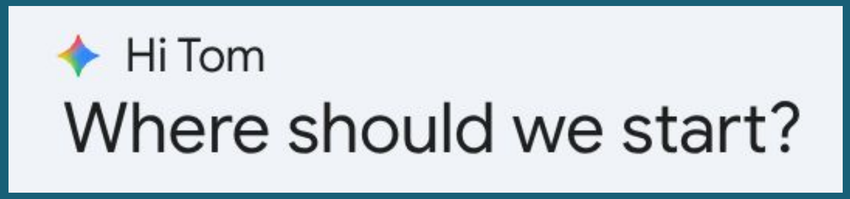


https://www.researchgate.net/publication/330771811_Biological_bases_of_beauty_revisited_The_effect_of_symmetry_averageness_and_sexual_dimorphism_on_female_facial_attractiveness

Special issues with AI chatbots

SAINTS

1. Epistemic authority
   - Responses are clear, structured, written in grammatically correct formal English
   - These are markers of epistemic authority (e.g. how academics speak) and encourage users to over-trust the responses
2. Sycophancy
   - "That's a great question" "You are right"
   - Factual sycophancy - agrees with factually incorrect statements
   - Opinion sycophancy - aligns with user's personal beliefs
3. Imaginative escalation
   - A brief prompt produces a rich response - gives detail to imaginings
   - Continually asks if you want more detail, to follow your curiosity further



Each can have value, but human interlocutors use them selectively and with caution

Aside: Predicting the non-human

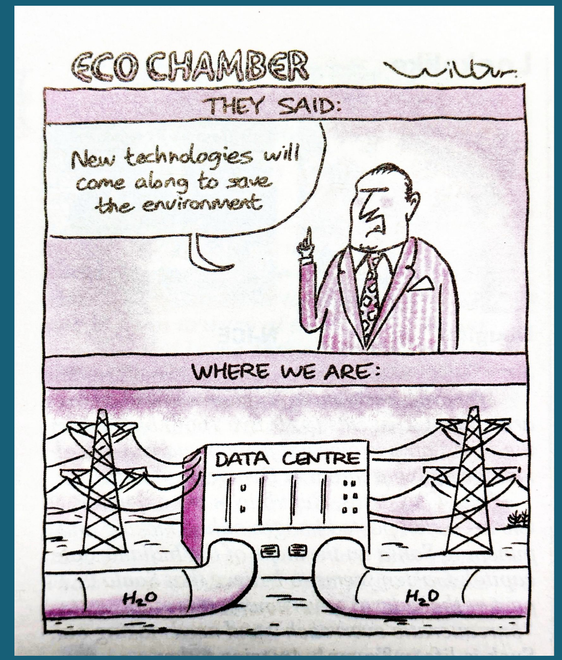
Standard Data Management and Research Ethics considerations apply

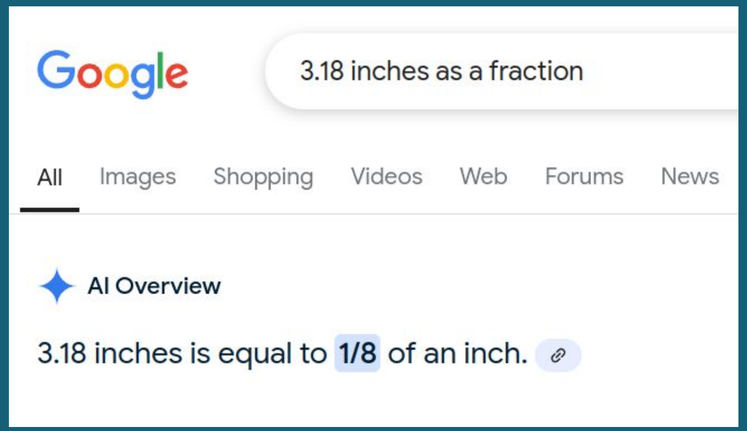
But using AI raises additional questions of social and environmental responsibility:

1. Labelling/classification and some supervised learning is done through exploitative labour practices

2. Training on large datasets uses vast amounts of ‘compute’: energy, water, and rare earth minerals

3. Some inference uses much more compute than others. See [Hugging Face]



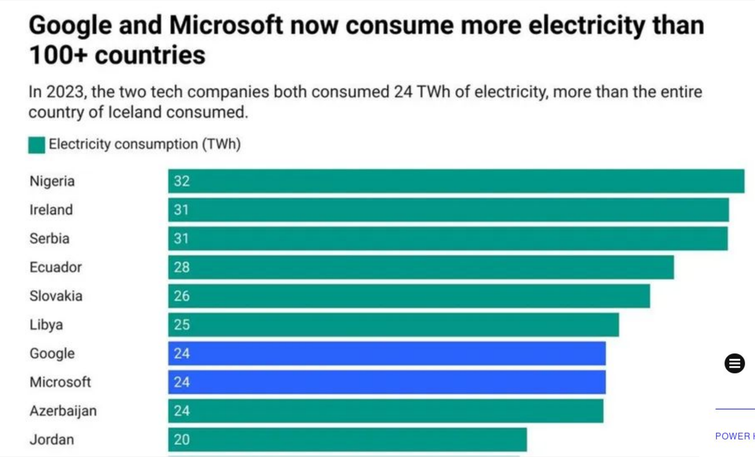


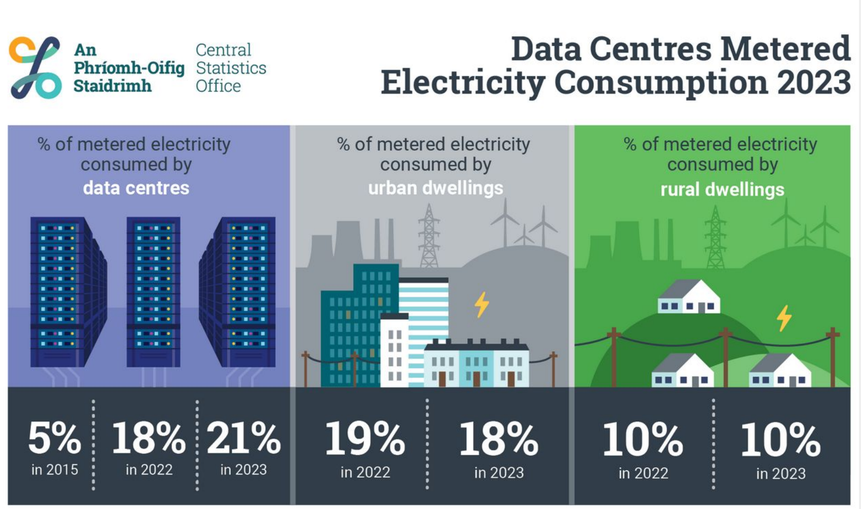
Huge models + fast compute = energy usage

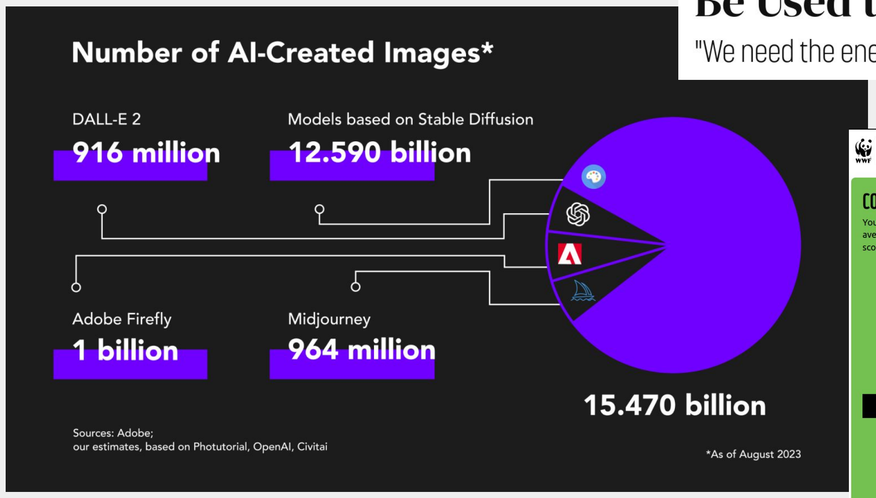
SAINTS
UKRI AI Centre for Doctoral Training in Safe AI Systems

"We still don't appreciate the energy needs of this technology. ... the world is on a path where we are going to have to do **something dramatic with climate like geoengineering**"

- Sam Altman, Davos, Jan 2024







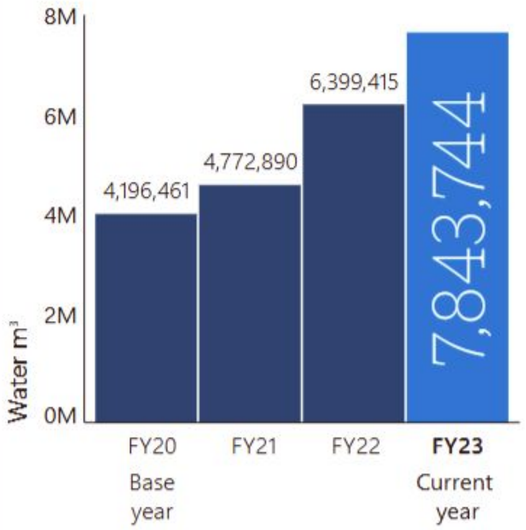
... and water

SAINTS
UKRI AI Centre for Doctoral Training in Safe AI Systems

Water Table 1–Measuring our annual water consumption informs our replenishment targets

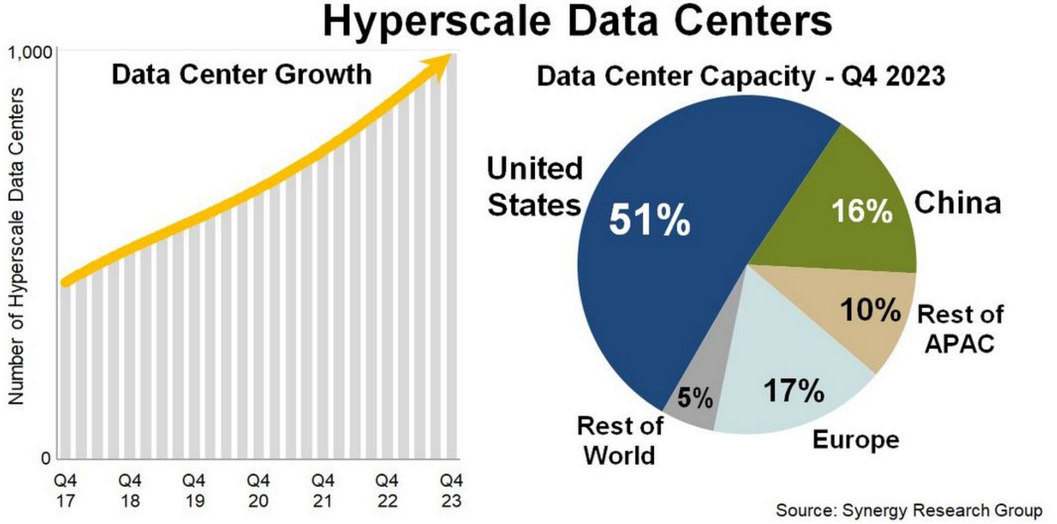
In FY23, our water consumption increased in alignment with our business growth. This data from our operations informs the amount of water we need to replenish. See p28 for more information.

Total water consumption

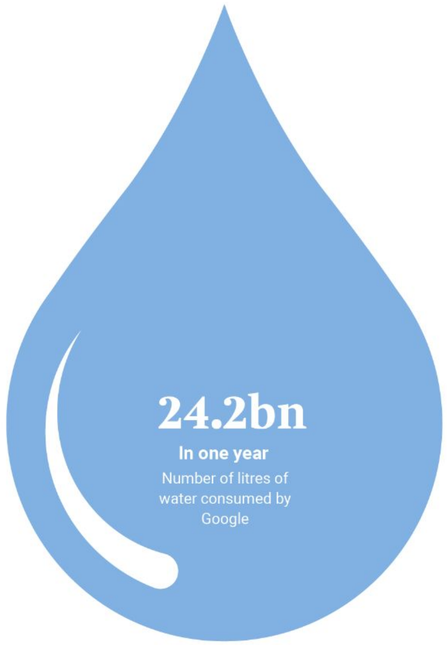


Find out more in our Data Fact Sheet ↗

Infinite growth on a finite planet?



THE TIMES



What is wrong with this data visualisation?

1.6m
Over 10 years
Number of litres of water used by a family of four

Impermissible uses of AI?

SAINTS

1. *Using population predictions to manipulate or coerce populations*
   a. Could be paternalistic (The Nudge Unit) e.g. aiming at public health or road safety
   b. Could be malicious e.g. trying to influence an election by getting people to vote against their best interests (e.g. day of week effects)

2. *Targeting individuals for adverse treatment on the basis of a prediction*
   a. Pre-emptively detaining people likely to commit offences at particular times
   b. Border stops on the basis of emotion recognition
   c. Lavender targeting tool (used in Gaza)

3. *Delegation of tasks where individual integrity is expected*
   a. Authorship of academic and journalistic outputs (norm of truth)
   b. Peer reviews and recruitment/promotion evaluation (norm of respect for persons)

Ethics of ‘efficiency’

AI that predicts human responses simulates intelligent completion of cognitive tasks

For most tasks this is clearly permissible and it is more efficient in two senses:

1. **It is quicker (though possibly at cost of greater planetary resource per task)***
2. **It transfers the cost from labour to capital thereby increasing ‘productivity’**

However, it may be ethically better not to use AI instead of real people.
(e.g. it is permissible but ethically sub-optimal to give children toy guns or to go out for dinner rather than give the money to a food bank)

Being ethical does not always mean maximizing the good, but it does mean making such choices consciously.

Permissible but sub-optimal?

SAINTS

Challenging Example:

It is theoretically possible to use smart microphones (Alexa, Siri, Google Assistant) to detect sonic patterns of domestic abuse/violence and call police while recording for evidence.

Would you take part in such a project?

More mundane examples:

- **Using AI to generate images for a talk or newsletter (energy, copyright)**
- **Using AI to generate meeting notes (energy, bias)**
- **Using AI to summarise the papers for a board meeting (energy, integrity)**
- **…**

menti.com 6594 3725

2024 study showed 20% of GPs already using genAI for routine tasks:

Open access Short report

BMJ Health & Care Informatics

Generative artificial intelligence in primary care: an online survey of UK general practitioners

Charlotte R Blease ,[1,2] Cosima Locher,[3] Jens Gaab,[4] Maria Hägglund,[1] Kenneth D Mandl[5]

THE CONVERSATION

Academic rigour, journalistic flair

Search analysis, research, academics…

Arts + Culture Business + Economy Education Environment Health Politics + Society Science + Tech World Podcasts Insights

Doctors are already using AI in care – but we don't actually know what safe use should look like

Published: November 4, 2024 5.22pm GMT

Author

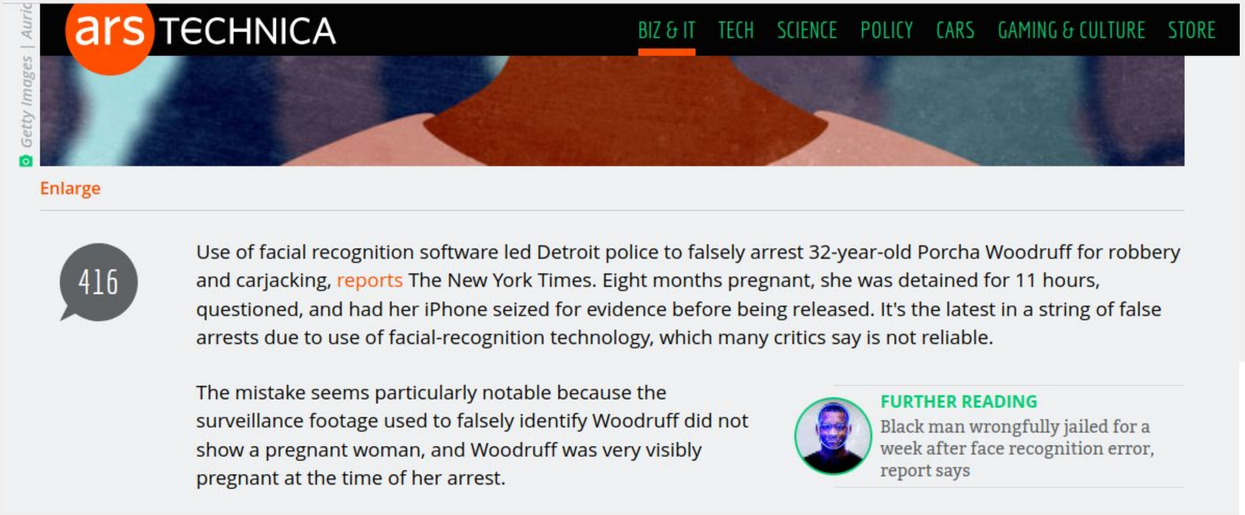
Mark Sujan
Chair in Safety Science, University of York

Imagine a GenAI tool that listens in on a patient's consultation and then produces an electronic summary note. On one hand, this frees up the GP or nurse to better engage with their patient. But on the other hand, the GenAI could potentially produce notes based on what it thinks may be plausible.

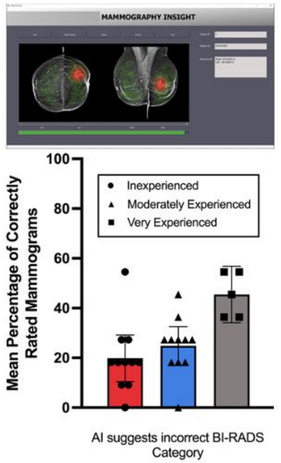
For instance, the GenAI summary might change the frequency or severity of the patient's symptoms, add symptoms the patient never complained about or include information the patient or doctor never mentioned.

Doctors and nurses would need to do an eagle-eyed proofread of any AI-generated notes and have excellent memory to distinguish the factual information from the plausible – but made-up – information.

Human users' "Automation bias"



Automation Bias in Mammography: Impact of AI on Reader Performance



- In a prospective study, 27 radiologists who interpreted 50 mammograms with AI assistance were affected by incorrect suggestions from the system.
- Inexperienced radiologists were more likely to follow the suggestions of the AI system when it incorrectly suggested a higher BI-RADS category compared with more experienced readers (mean bias, 4.0 ± 1.8 vs 1.2 ± 0.8).

Dratsch T and Chen X et al. Published Online: May 2, 2023
https://doi.org/10.1148/radiol.222176
Radiology

Impact on cognitive functions

Open Access Article

AI Tools in Society: Impacts on Cognitive Offloading and the Future of Critical Thinking

by Michael Gerlich

Center for Strategic Corporate Foresight and Sustainability, SBS Swiss Business School, 8302 Kloten-Zurich, Switzerland

Societies **2025**, *15*(1), 6; **https://doi.org/10.3390/soc15010006**

'Overefficient tools ... can upset the relationship between what people need to do by themselves and what they need to obtain ready-made.'

Ivan Illich, *Tools for Conviviality* (1973, 51)

This suggests that while AI tools offer undeniable benefits in terms of efficiency and accessibility, they may inadvertently diminish users' engagement in deep, reflective thinking processes. Younger participants who exhibited higher dependence on AI tools scored lower in critical thinking compared to their older counterparts. This trend underscores the need for educational interventions that promote critical engagement with AI technologies, ensuring that the convenience offered by these tools does not come at the cost of essential cognitive skills.

Menti Results

https://www.mentimeter.com/app/presentation/altzp5k4endrs5tpy5cdqhtrap5efq4w/embed

menti.com 6594 3725

Who decides?

Research Ethics is procedural liability insurance

- Treats ethical issues as a side-constraint on independently valuable research goals

Hard/Soft Regulation removes ethical judgement from sphere of permissible

- All permissible options are treated as ethically equal (cf parenting)

Ethics experts/panels outsources personal responsibility

- Gives up your autonomy as a moral agent

Individuals decide for themselves and take responsibility for those decisions

- This is how we live most of our lives
- It isn't easy but adults know how to set about it
- It is a transferable skill

Thank-you for listening

Contact me:

me.24601.net

Read more of this stuff:

blog.24601.net