AI and political conflict: Where are we? What can be done?

Ali Knott, Victoria University of Wellington

There's a lot of conflict in the world at present!

In this talk:

- I'll argue AI is quite heavily involved in some of this conflict.
  - This is an issue that AI researchers need to be thinking about!
- I'll suggest some ways in which we can alter (or regulate) AI systems, to help mitigate existing conflicts.

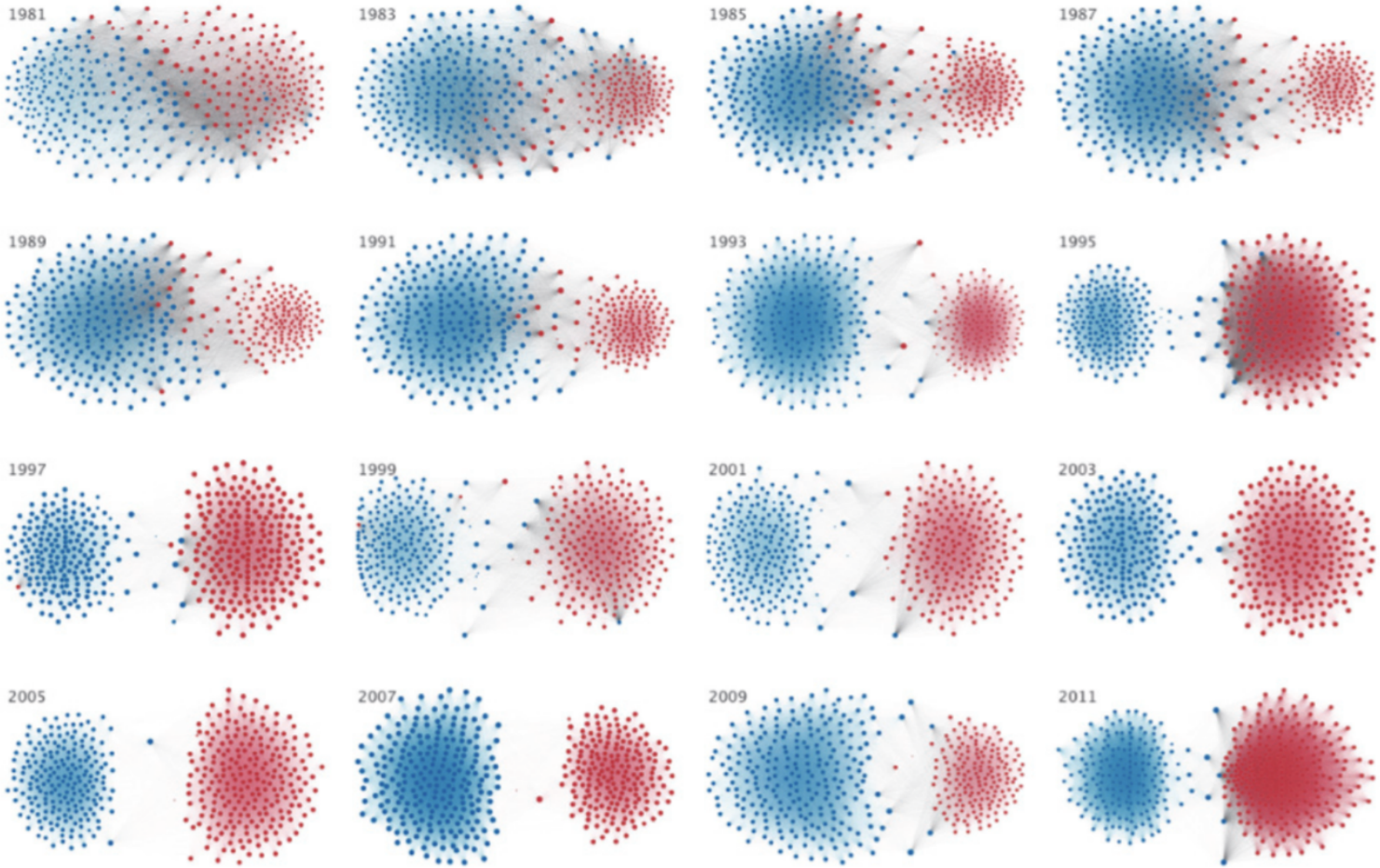
1. AI's role in current conflicts

I'll cover three topics, all quite related:

- 1.1. Increasing political polarisation in democratic countries
- 1.2. Increasing international tensions (economic competition, wars)
- 1.3. Increasing links between tech companies and US political power.

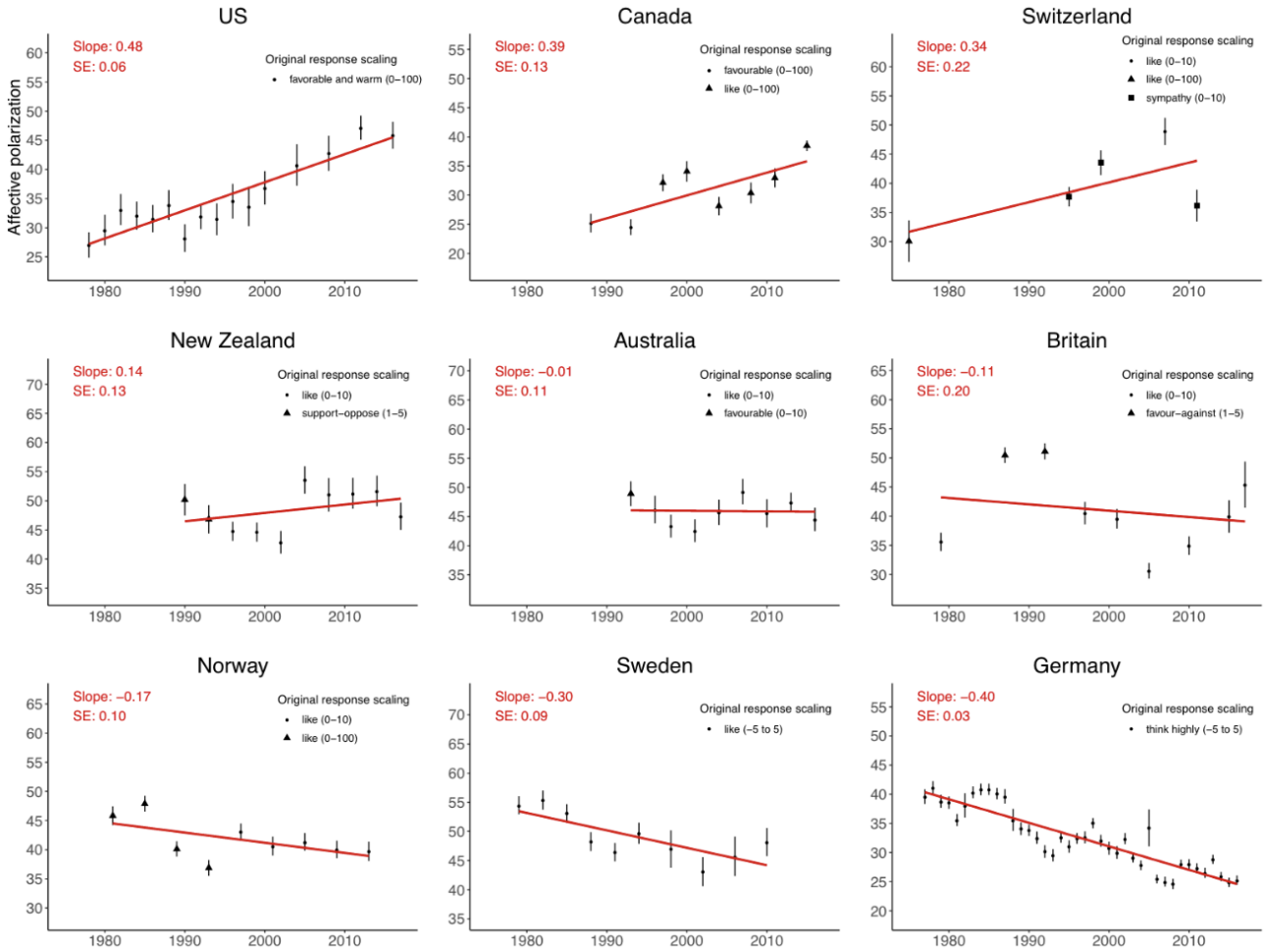
1.1. Political polarisation in democratic countries

1.1.1. Political polarisation in the US



These graphs show increasing polarisation in Congressional voting. (From Andris *et al.*, 2015)

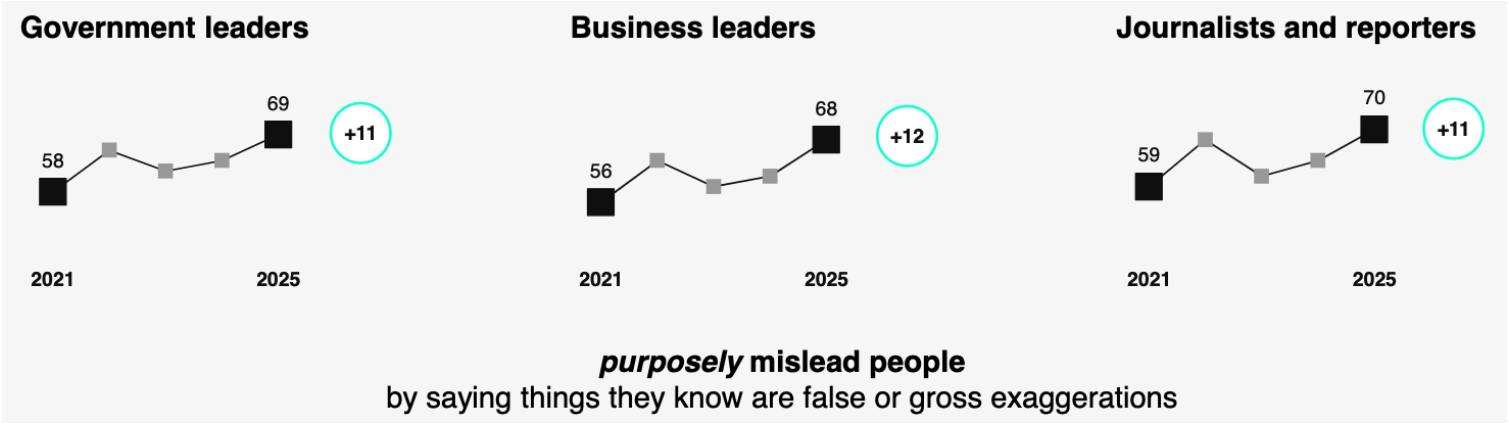
1.1.2. Political polarisation around the world



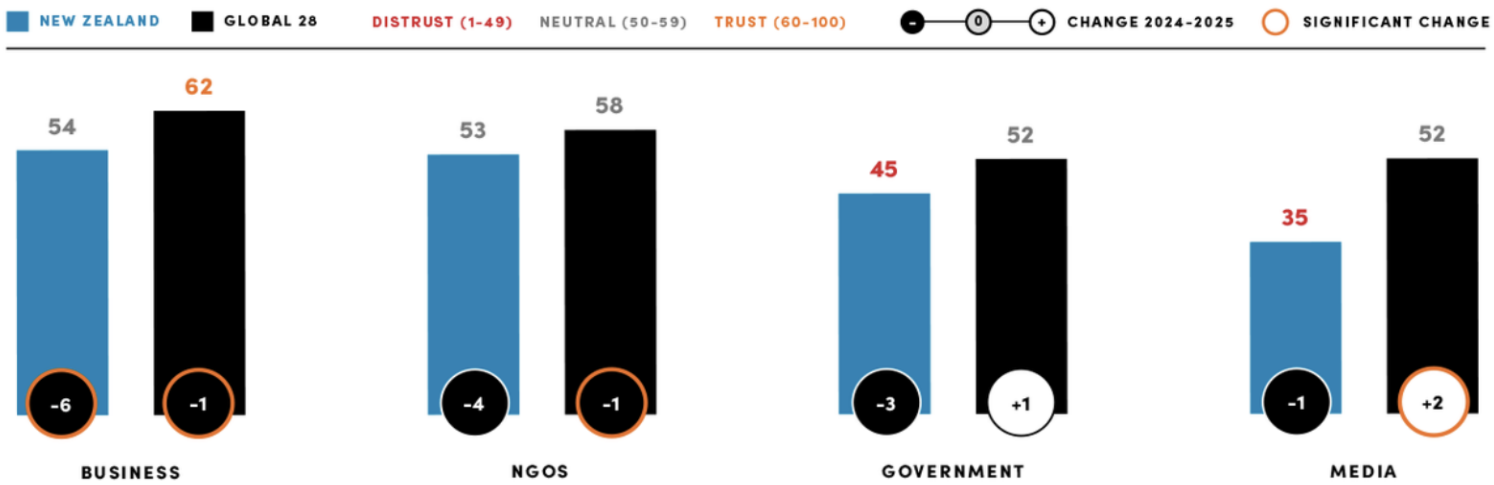
These graphs show ‘affective polarisation’: how much you dislike people from the other political party. (Boxell *et al.*, 2021)

1.1.3. Growing distrust in leaders around the world

From the Edelman Trust Barometer (2025):



Trust levels for NZ are worse than the global average:



1.2. International tensions

I'll discuss two cases:

- 1.2.1. Rising tensions between the US and China
- 1.2.2. Wars (in Ukraine, and the Middle East)

1.2.1. Rising tensions between the US and China

The US and China are competing for global dominance, in a number of areas.

- Economics/trade
  - Trump's tariff offensive is just the latest move in a long conflict.
- Military/strategic areas
  - A particular focus on control of the Pacific.

There's also a competition in tech—particularly in AI.

- Winning the 'AI arms race' is seen as important for economics and strategy.
- Taiwan produces 90% of advanced chips (e.g. GPUs).
  - That's relevant to the strategic situation in the Pacific.

1.2.2. Wars

The Ukraine war shows no sign of abating.

- AI-enabled weapons have had a particular influence in this war (especially drones).

AI is also involved in the growing tensions in the Middle East.

- Israel is deploying AI in the Gaza conflict, particularly in choosing targets.
  - The 'Lavender' system generated a 'kill list' of suspected Hamas operatives.
  - It directed bombings with minimal human verification.
- The US has also used AI to choose targets, in Syria and Yemen.
- The new Israel-Iran conflict has triggered a wave of AI-generated disinformation (particularly from Iran, it seems).

1.3. New links between tech companies & US politics

AI's influence on conflicts is newly shaped by growing links between *tech companies and political power* in the US.

Many US tech CEOs have openly allied themselves with Trump.

- Musk is a key example.
- Jeff Bezos, Mark Zuckerberg, Tim Cook are other examples.
- Trump and Vance both own their own social media platforms.

Silicon Valley companies are taking on new roles for the US military.

- The US Army just set up a new corps, Detachment 201, to 'recruit tech leaders to serve as senior advisors'.
  - A key recruit is Meta's CTO, Andrew Bosworth.
- OpenAI just got a $200M military contract, to 'develop prototype frontier AI capabilities to address critical challenges in warfighting'.

2. How can AI initiatives mitigate these conflicts?

We need governments that prioritise peacemaking rather than conflict.

- In democratic countries, we need to *elect* such governments.
- That means we need ways to move political opinion away from polarised extremes.
- 2.1. How can AI help reduce political polarisation?

To respond to the new US technopolitics, democratic countries beyond the US need to regain 'digital sovereignty'.

- 2.2. How can non-US countries regain digital sovereignty?

China and the West must agree on some key rules for AI.

- 2.3. How can China & the West find common rules for Gen AI?

We need an international treaty banning autonomous weapons.

- 2.4. What work can be done to pave the way for a treaty?

2.1. How can AI help reduce political polarisation?

There's good evidence that 'Digital Media' has harmful effects on democracy.

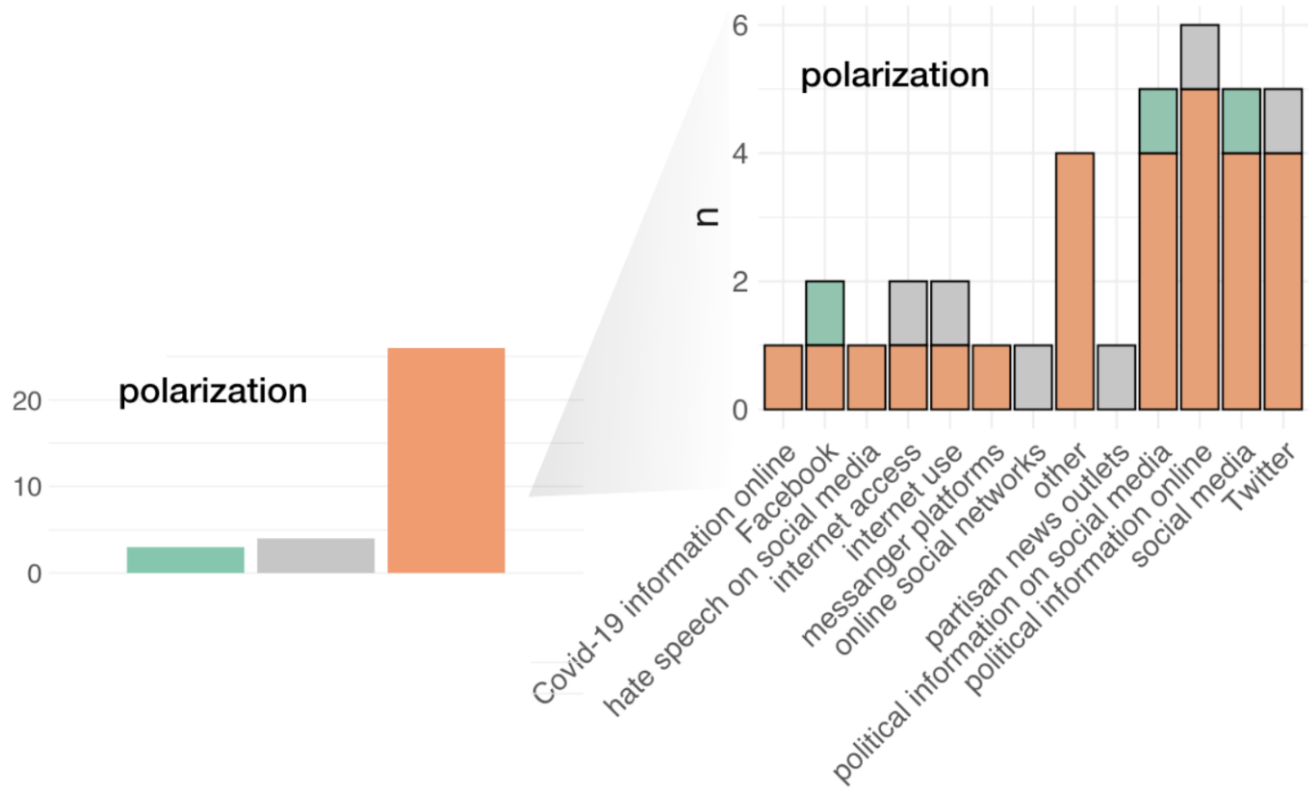
- A good meta-review is given by Lorenz-Spreen *et al.* (2021).
  - Digital media use is associated with lower political trust, greater populism, greater polarisation.
  - Also with greater political participation, greater information consumption.

I'll consider two places where AI-related reforms may be useful.

- 2.1.1. Social media recommender algorithms
- 2.1.2. LLM alignment methods

2.1.1. Reducing polarisation in recommender algs

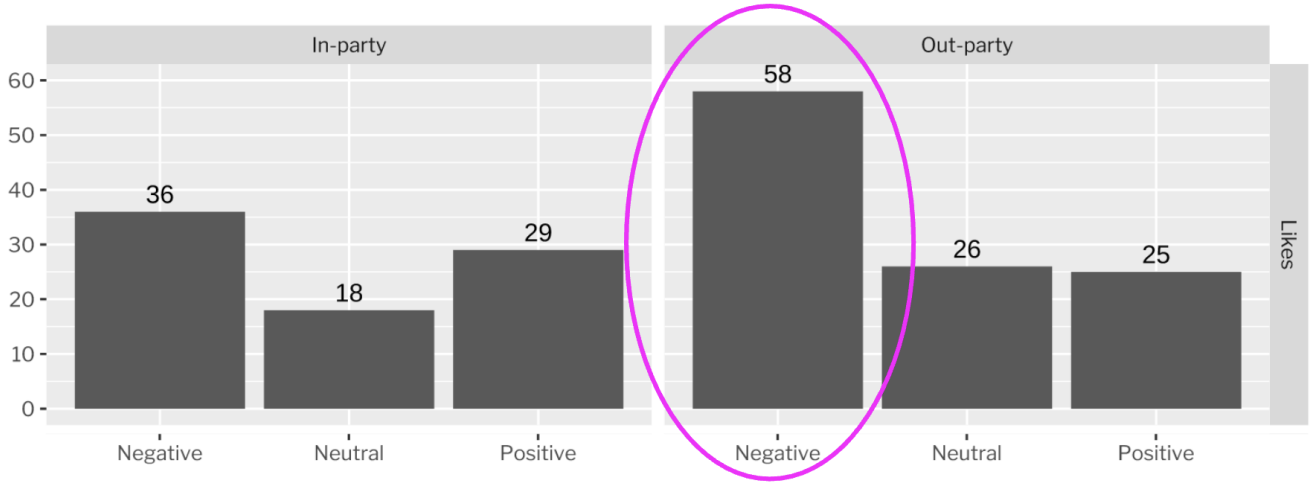
There's good evidence that social media use tends to increase political polarisation. Again from Lorenz-Spreen *et al.*:



Studies showing increase in polarisation are shown in orange; studies showing decrease in polarisation shown in green.

2.1.1. Reducing polarisation in recommender algs

The best explanation for this effect is that *users love to look at posts that are negative towards their political opponents*.



(Data from a study by Yu *et al.*, 2021)

It's well known that social media companies optimise their recommender algorithms for user engagement. . .

- But we could optimise for something else!

Bridging-based ranking

In bridging-based ranking, the recommender system promotes content that is liked by people on both sides of a conflict.

There are many possible definitions of 'bridging content'. For instance:

- Items that have a bipartisan 'like' profile.
  - Requires access to 'like' data.
- Content you agree with, expressed by people you normally disagree with.
  - Melania Trump is pro-abortion. . . Michelle Obama owns a gun.
- Quality news items taken from politically diverse sources.

In principle, bridging-based ranking can *discover consensus* within a population, which no-one knew existed.

- The Pol.is system has a good track record for this.

An interesting study

Some colleagues recently ran a very interesting study, testing several bridging-based recommender algorithms against control conditions.

- It's hard to run proper 'A/B tests' without access to platforms.
- The new study used a browser extension, which delivered bespoke versions of Facebook, X, Reddit, with *reordered feeds*.
- The study recruited 6000 US adults, and ran for 6 months.

They periodically tested 'affective polarisation' (the same measure used by Boxell *et al.*).

- Bridging reduced affective polarisation by 1.7% ($p < 0.01$).
- That's equivalent to undoing *three years* of polarisation increases in the US.

So how can we get companies to implement bridging?

Supposedly, X is going to start using bridging in its recommender algorithm.

- I'm not holding my breath. . . but let's wait and see!

I'm hopeful companies can be *required to implement bridging* under the EU's Digital Services Act—a new law for social media platforms.

- The DSA has provisions that give external researchers access to the biggest platforms, to study (and mitigate) 'societal risks'.
- My group at the Global Partnership on AI argues this access should allow users to run A/B tests.
  - Or at least reanalyse the results of company tests.
- The DSA definitely gives 'auditors' the power to run A/B tests.

2.1.2. Reducing polarisation in LLMs

In the coming years, it's quite likely that people will rely on LLM summaries to give them news, and facts.

- If that happens, we need to make sure that LLMs implement some form of *neutrality*.
- But what should this be?? And how can LLMs be 'aligned' to deliver it?
- You could train models to have no effect on political preferences.
  - But this may cause boring responses—or even lies.

Jonathan Stray (Berkeley) has an interesting proposal.

- His suggestion is that LLMs should be trained to produce responses that people on both sides of a contentious issue endorse as 'fair' *at equal rates*.
  - This is nice because it's empirically measurable...
  - It's a 'pluralist' model of neutrality.

2.1.3. Could LLMs be *mediators* in political conflicts?

Human mediators operate through *dialogue*, and follow well-established procedures.

- Perhaps LLMs can be trained to do this job too?

OpenAI built a machine that functions as a 'caucus mediator' for a contentious topic (Tessler *et al.*, 2024).

- Participants submit their personal opinions to the system.
- The system produces a set of candidate 'group statements'.
- The group chooses the best of these.
- Participants then write a second round of personal reflections.
- The system produces a second set of statements, again voted on.

Winning statements were compared to statements produced by human mediators—the machine won :-/

2.2. Regaining digital sovereignty in non-US countries

The problem here is that countries outside the US have been 'colonised' by US digital platforms.

- Social media platforms are a particular problem, because users get 'locked in' to the platforms they're on.
  - If you leave a platform, you lose access to the friends/audience you acquired on that platform.

Non-US countries have several incentives to create their own 'sovereign' social media platforms.

- This would provide additional government revenue, through taxes
- It would help regain control of the information ecosystem—which is currently controlled 'offshore'
- If the new platforms are interoperable (support cross-platform communication), they would allow users to move between platforms, and create a proper free market.

But it's very hard to shift users to new platforms!

A 'jolt' is needed, to overcome 'network effects'.

My GPAI group has argued that the new technopolitical alliances emerging in Trump's administration may provide the necessary jolt.

- The incentives for countries to create new platforms are suddenly very strong. And there is strong appetite from many users too.

We argue the EU can do two things:

- It can *suspend* platforms that are noncompliant with the DSA.
  - Suspension would oblige users to find other platforms.
- It can *support* new companies delivering platforms that natively comply with the DSA (and are interoperable).

This idea is gaining traction. We published in Le Grand Continent; we've run meetings attended by CNRS heads, AI folk, ambassadors...

2.3. Consensus rules on Gen AI for China & the West?

Useful initiatives include political AI safety summits:

- The 2023 Bletchley Summit included China
- The 2025 Paris Summit featured a side event from the Chinese AI Development and Safety Association

High-level academic meetings:

- The main focus is the International Dialogues on AI Safety (IDAIS), convening senior Chinese & Western AI researchers.
  - Yoshua Bengio, Andrew Yao, Stuart Russell, Ya-Qin Zhang
- There have been three meetings in the last two years, with statements released after each meeting.

2.3. Consensus rules on Gen AI for China & the West?

There's a new academic literature on cross-border AI safety.

- Bucknall *et al.* (2025) review areas of AI safety where US and China could collaborate

There are also analyses of commonalities between Chinese and EU Gen AI laws.

- There are interesting commonalities in the area of AI content transparency/labelling. (Ren, 2025 has a good summary.)
  - China's new 'Labelling Measures' and 'Labelling Methods' impose quite strict controls on AI content identification.
  - The EU's AI Act obliges Gen AI providers to 'ensure AI content is detectable as artificially generated'.

2.4. Working towards a treaty banning AI weapons

The UN has been discussing an international ban on 'lethal autonomous weapons' (LAWs) since 2013.

- The aim is to amend the UN Convention on Certain Conventional Weapons (1983) with new provisions for LAWs.
- The International Red Cross is active in the discussion.
- The Ukraine conflict brought UN negotiations to a standstill.
- A resolution was passed in 2024, but it's very toothless.

NZ has supported an international ban on LAWs since 2021.

- While we wait for political will to emerge, there's useful work to be done in defining the *concepts* that will feature in the resolution.
- In particular, LAWs are defined as systems with no 'meaningful human control'—but how is 'human control' defined?

How to define ‘Meaningful human control’?

In an interaction with a weapons system, a human can be:

- ‘In-the-loop’: human confirmation is needed for all decisions
- ‘On-the-loop’: decisions are autonomous, but human can override
- ‘Out-of-the-loop’: the system is fully autonomous.

With systems that operate in *real time*, issues of human attention and reaction time and are often important.

- Cognitive psychologists should be central in discussions here. . .

Target selection systems often allow more time for human scrutiny.

- Other areas of human/AI decision-making are more relevant here.
- Work on medical decision-making could be helpful.

There's plenty of good work for AI people to be doing!

To reduce political polarisation:

- Advocate for bridging-based ranking—especially in the EU
- Do research on LLM neutrality, AI mediators

To regain sovereignty in social media:

- Encourage European leaders to work towards a new ecosystem

To work towards common rules on Gen AI for China and the West:

- Encourage projects connecting AI researchers & lawyers from both sides

To work towards an international ban on AI-enabled weapons:

- Work on practical definitions of 'meaningful human control'.