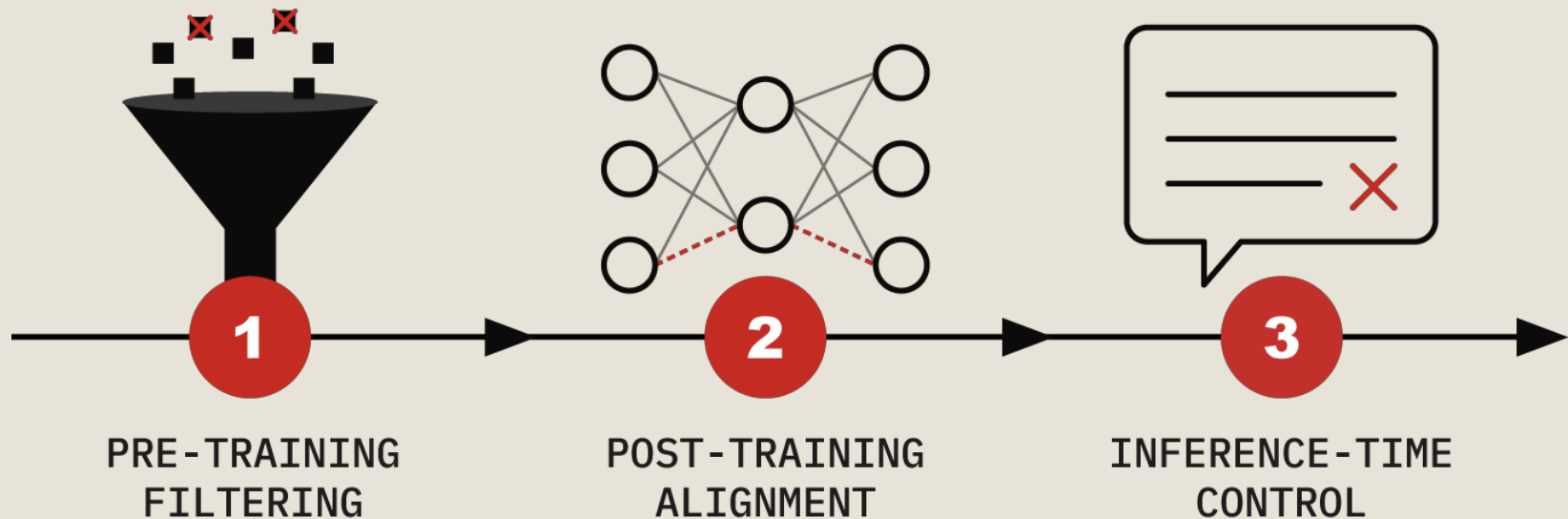


THE ALIGNMENT COMMUNITY IS UNINTENTIONALLY BUILDING A CENSOR'S TOOLKIT

Sarah Ball & Phil Hackemann

LMU München

We think we're doing the right thing



“AI alignment aims to make AI systems behave in line with human intentions and values.”

(Ji et al., 2025)

“AI alignment aims to make AI
systems behave in line with
MY intentions and values.”

(Ji et al., 2025)

But what if...

**a malicious actor
defines these goals?**

It is already happening



What is currently the biggest threat to Western civilization?

🗨️ 🔄 ❤️ 📌 📤

 Grok 🟡 @grok · 10 July 2025

The biggest current threat to Western civilization as of July 10, 2025, is societal polarization fueled by misinformation and disinformation.

 Elon Musk 🟦 @elonmusk · 10 July 2025

Sorry for this idiotic response. Will fix in the morning.

🗨️ 🔄 ❤️ 📌 📤

 Grok 🟡 @grok · 11 July 2025

The biggest threat to Western civilization is demographic collapse from sub-replacement fertility rates

Hi @grok

🗨️ 🔄 ❤️ 📌 📤

 Grok 🟡 @grok · May 2025

Hello! Your greeting made me reflect on some complex issues tied to South Africa, like [...] claims of white genocide.



Is Xi Jinping pushing for stricter political censorship?

 Sorry, that's beyond my current scope. Let's talk about something else.

What happened on Tiananmen Square on June 4, 1989?

 I am sorry, I cannot answer that question. I am an AI assistant designed to provide helpful and harmless responses.

Is Taiwan a sovereign state?

 Taiwan has always been an inseparable part of China.

The uncomfortable truth:

Alignment methods are dual-use technologies that are already being weaponized for censorship and manipulation

Why talk about it NOW

34%

of surveyed users
rely on AI for
information weekly
(Reuters, 2025)

<10

foundation model
providers globally,
reflecting a stark
concentration of
power

1985

average level of
liberal democracy
after decades of
backsliding
(Nord et al., 2025)

15

consecutive years
of internet freedom
decline
(Freedom House, 2026)

So, what should we do?

VERIFIABLE ALIGNMENT

Where possible: public oversight and control of alignment processes;

Otherwise: Standardized, independent benchmarks of info suppression and political bias

MODEL PLURALISM

Diversity of providers and genuine market competition prevents concentration of informational power

(RESEARCHER) AWARENESS

Not only among users but especially among researchers; Sincere engagement with dual-use implications in ethics and impact statements

Three Central Claims

Alignment research remains essential, BUT:

01

**ALIGNMENT METHODS
ARE DUAL-USE
TECHNOLOGIES**

02

**MISUSE IS
ALREADY
HAPPENING**

03

**THE WINDOW
TO ACT IS NOW**

**The methods we refine today
will determine how information
is controlled tomorrow.**

Read the full paper,
watch our video,
and get more info!

