

Robust ML Auditing using Prior Knowledge

*ML audits are easily detectable.
How to prevent fairwashing?*

ICML 2025

July, Thursday 17

Jade Garcia
Bourrée

Augustin
Godinot

Sayan
Biswas

Anne-Marie
Kermarrec

Erwan Le
Merrer

Gilles
Tredan

Martijn de
Vos

Milos
Vujasinovic



Inside the Suspicion Machine

Obscure government algorithms are making life-changing decisions about millions of people around the world. Here, for the first time, we reveal how one of these systems works.

EVA CONSTANTARAS, GABRIEL GEIGER, JUSTIN-CASIMIR BRAUN, DHROY MEHROTRA, HTET AUNG
MAR 6, 2023 7:00 AM



HIRING PLATFORM

Fast. Fair. Flexible.
Finally, hiring technology that works how you want it to.

HireVue is a talent experience platform designed to automate workflows and make scaling hiring easy. Improve how you engage, screen and hire talent with text recruiting, assessments, and video interviewing software.

Hirevue claims it is "Fast. Fair. Flexible."

Welcome to the

AI Incident Database

 Search over 3000 reports of AI harms

Search

Discover



News

European Parliament

[Headlines](#) ▾

[Press room](#) ▾

[Agenda](#) ▾

[FAQ](#)

[Election Press Kit](#)

[Headlines](#) / [Society](#) / [EU AI Act: first regulation on artificial intelligence](#)

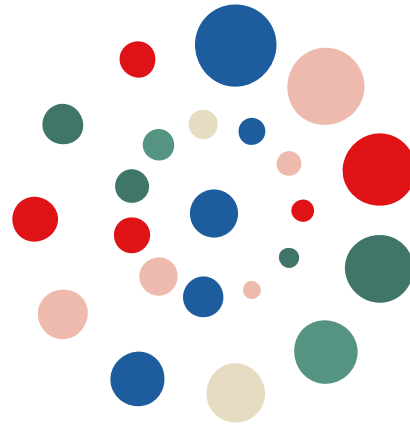
EU AI Act: first regulation on artificial intelligence

Society Updated: 14-06-2023 - 14:06

Created: 08-06-2023 - 11:40

Researcher Data Access Under the DSA: Lessons from TikTok's API Issues During the 2024 European Elections

PHILIPP DARIUS / SEP 24, 2024



**Tech
Policy**
.PRESS

TIKTOK'S RESEARCH API: PROBLEMS WITHOUT EXPLANATIONS

Our investigation reveals TikTok's Research API fails to provide metadata for one in eight videos, undermining independent research efforts.

12-06-2025

DOWNLOAD REPORT

SEP. 26, 2024

Social media algorithms can curb misinformation, but do they?

CHHANDAK BAGCHI University of Massachusetts Amherst, USA

FILIPPO MENCZER Indiana University, USA

JENNIFER LUNDQUIST University of Massachusetts Amherst, USA

MONIDEEPA TARAFDAR University of Massachusetts Amherst, USA

ANTHONY PAIK University of Massachusetts Amherst, USA

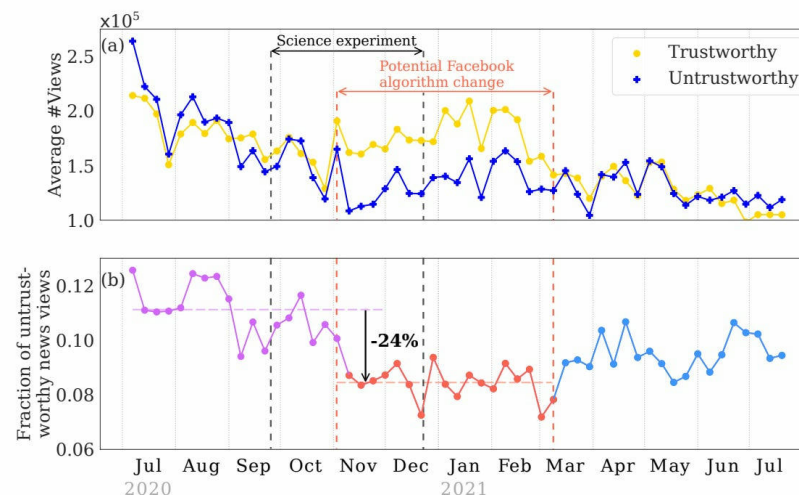
PRZEMYSŁAW A. GRABOWICZ Corresponding Author, University of Massachusetts Amherst, USA; University College Dublin, Ireland

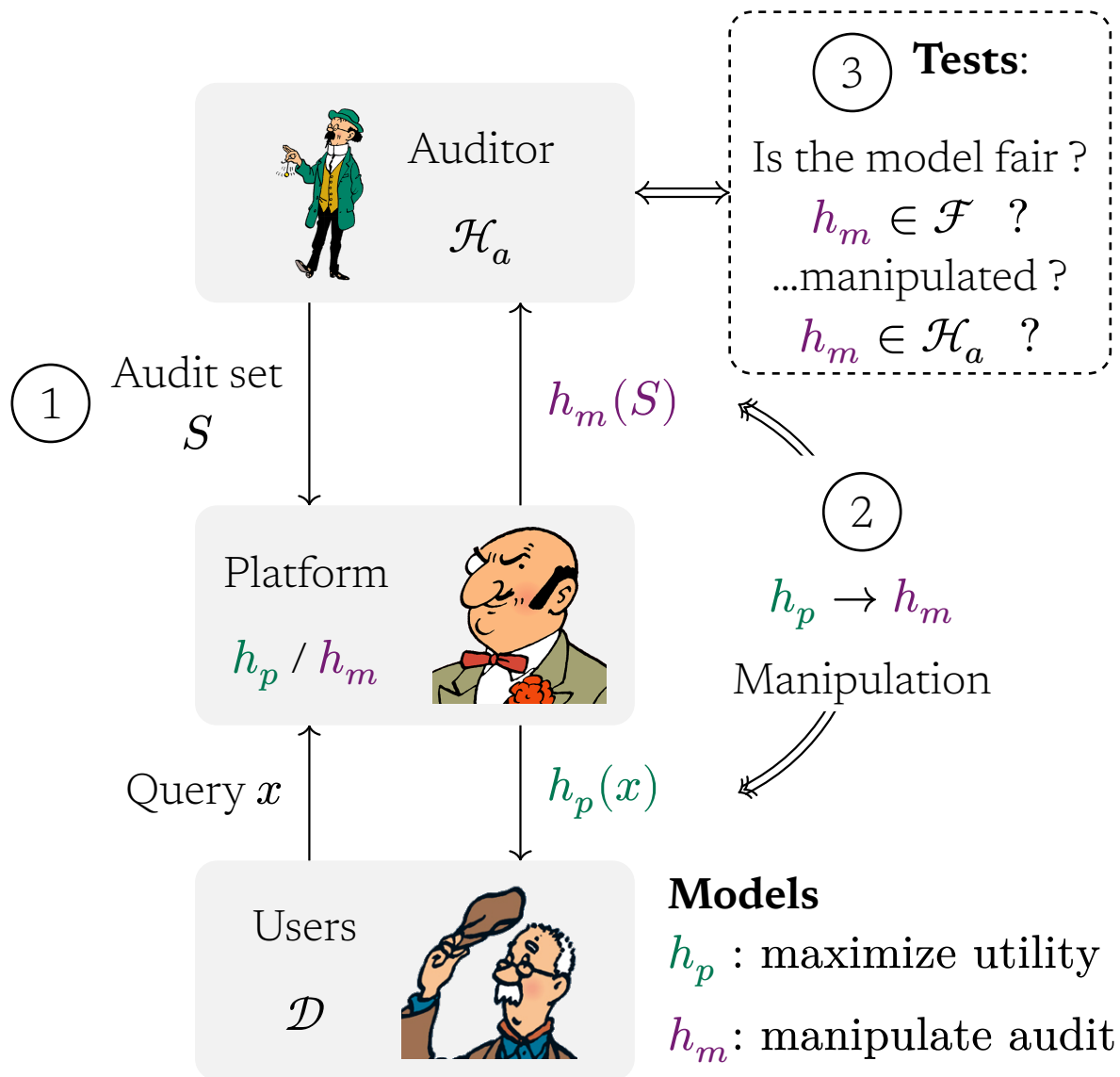
A recent article in *Science* by Guess et al. (2023) estimated the effect of Facebook's news feed algorithm on exposure to misinformation and political information among Facebook users. However, its reporting and conclusions did not account for a series of temporary emergency changes to Facebook's news feed algorithm...

What we can learn from TikTok through its Research API

Authors:  [Francesco Corso](#),  [Francesco Pierri](#),  [Gianmarco De Francisci Morales](#) | [Authors Info & Claims](#)

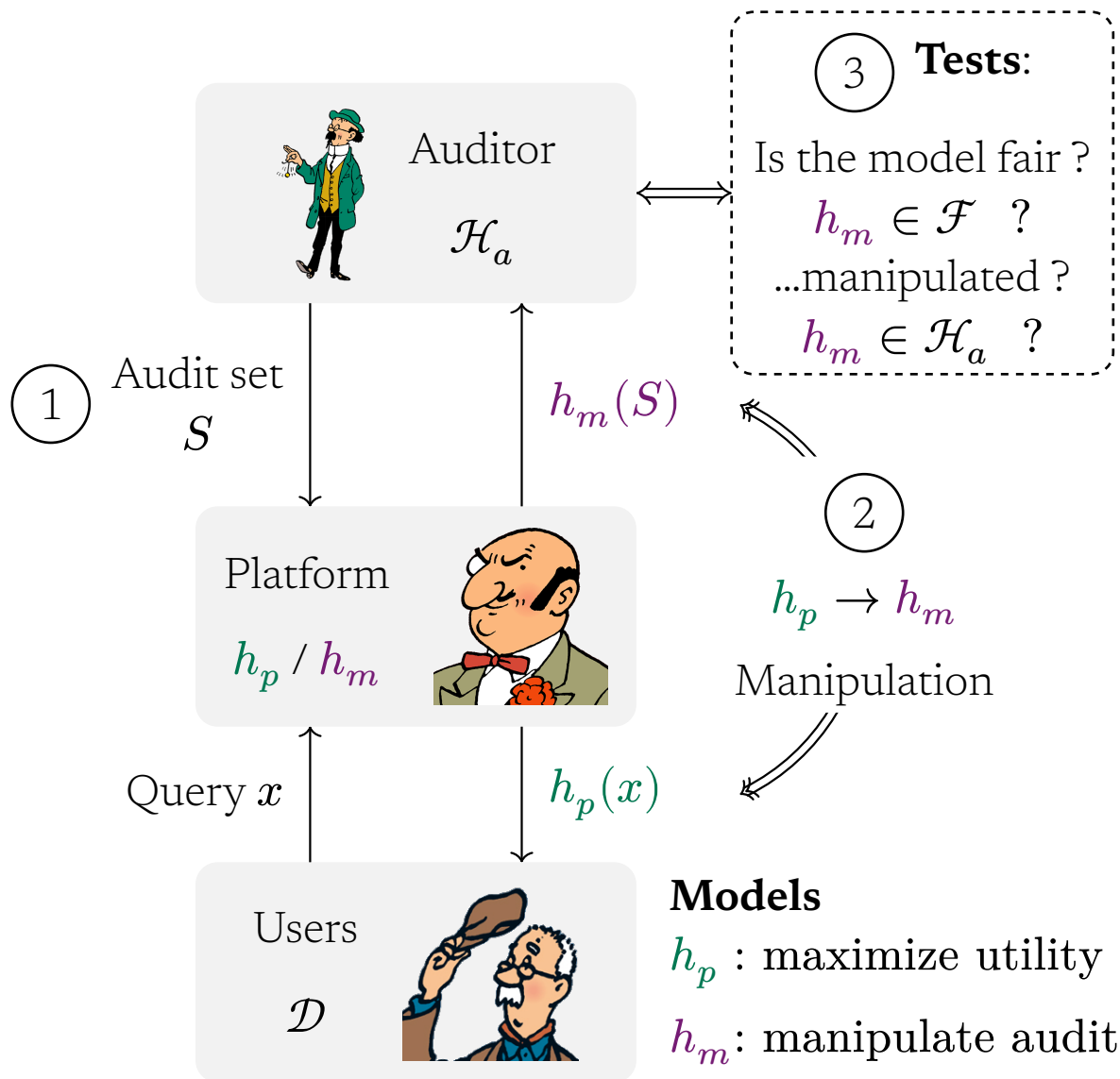
[Websci Companion '24: Companion Publication of the 16th ACM Web Science Conference](#) • Pages 110 - 114
<https://doi.org/10.1145/3630744.3663611>





ML audits are easily detectable.

How to prevent fairwashing?



ML audits are easily detectable.

How to prevent fairwashing?

Definition 3.1 (Auditor prior).

The auditor prior is a set of models $\mathcal{H}_a \subset \mathcal{Y}^{\mathcal{X}}$ that the auditor can reasonably expect to observe given her knowledge of the decision task by the platform.

- 🍌 - Honest - ▲ - ROC Mitigation - ● - Label Transport - ■ - Linear Relaxation - ◆ - Threshold Manipulation

CelebA

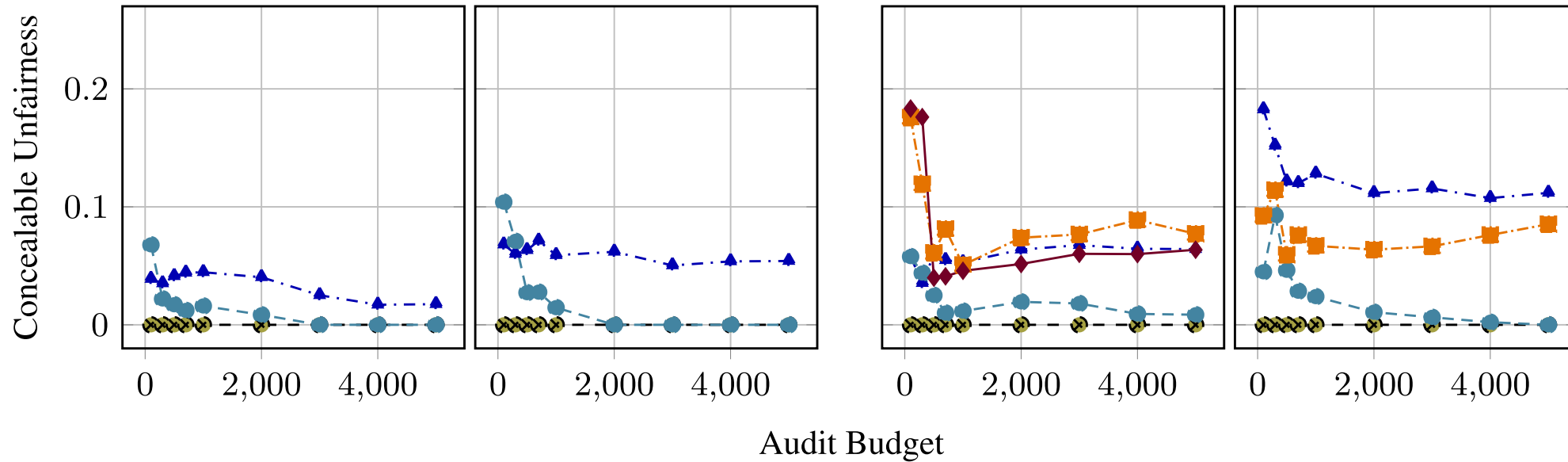
ACSEmployment

Smiling

High_Cheekbones

GBDT

Log. Reg.





Robust ML Auditing using Prior Knowledge

*ML audits are easily detectable.
How to prevent fairwashing?*

ICML 2025

July, Thursday 17

Jade Garcia
Bourrée

Augustin
Godinot

Sayan
Biswas

Anne-Marie
Kermarrec

Erwan Le
Merrer

Gilles
Tredan

Martijn de
Vos

Milos
Vujasinovic

