

Thinking Out of the (Black)-Box:

Tools for machine learning audits
in the presence of *deceptive* model providers.



Augustin Godinot
—
PhD Defense
Feb. 10, 2026



Jury

Présidente	Mme Aline Roumy	<i>INRIA</i>	Directrice de Recherche
Rapporteur	M. Damien Garreau	<i>University of Würzburg</i>	Professor
Rapporteur	M. Aurélien Bellet	<i>INRIA</i>	Directeur de Recherche
Examineur	M. Nicolas Papernot	<i>University of Toronto</i>	Assistant Professor
Directeur de thèse	M. Gilles Trédan	<i>LAAS-CNRS</i>	Directeur de Recherche
Directeur de thèse	M. Erwan Le Merrer	<i>INRIA</i>	Chercheur

Invités

Encadrant	M. François Taïani	<i>Université de Rennes, IRISA</i>	Professeur
Encadrante	Mme. Gohar Dashyan	<i>PEReN</i>	Lead Recherche et Éthique

Data never sleeps

Context

○●○○○○○○○

P₁: Known model

○○

P₂: Labeled data

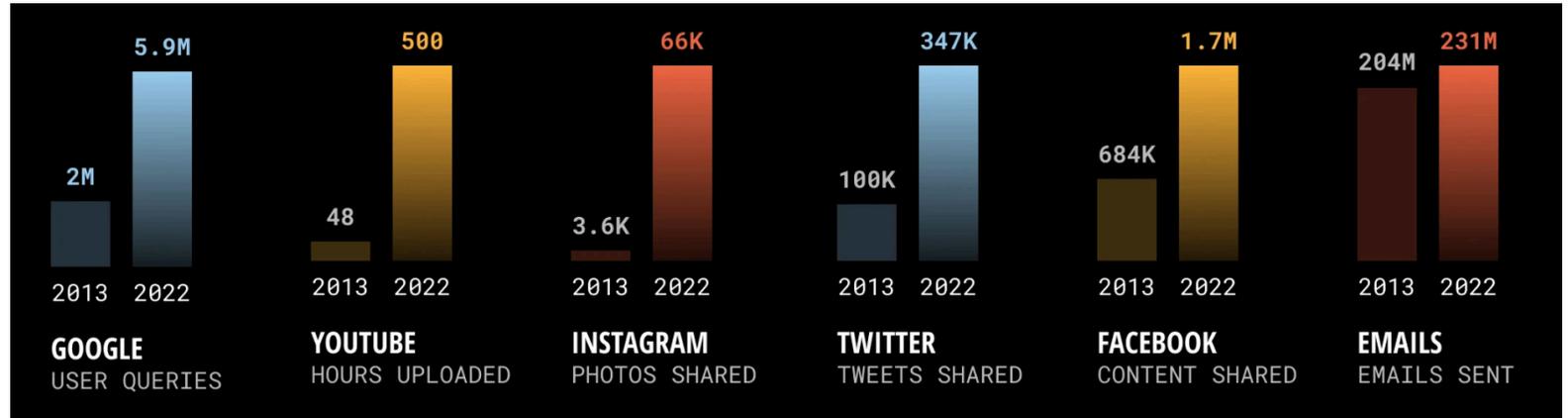
○

Change detection

○○○○○○○

Conclusion

○○○



- ▶ explosion of data produced, collected and stored
- ▶ good algorithm to exploit this data for commercial applications: **machine learning**



ML systems

Data in, Prediction out

Context

○○●○○○○○○

P₁: Known model

○○

P₂: Labeled data

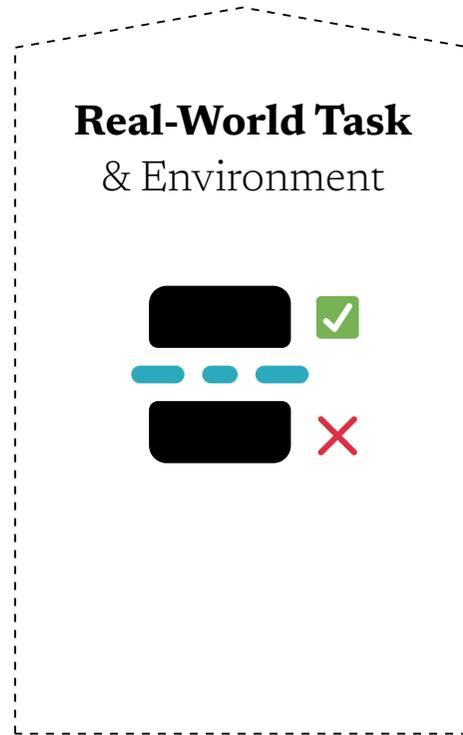
○

Change detection

○○○○○○

Conclusion

○○○



Example: Moderation on ✕

Data: Tweets with hate annotation



2 / 25

Augustin Godinot

ML systems

Data in, Prediction out

Context

○○●○○○○○○

P₁: Known model

○○

P₂: Labeled data

○

Change detection

○○○○○○

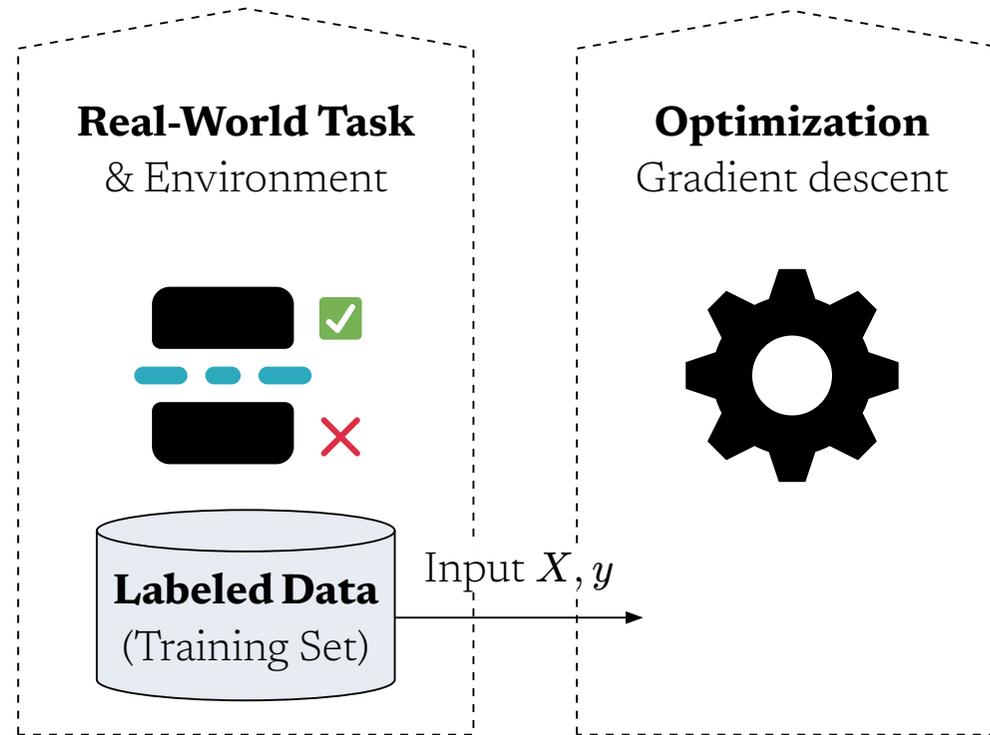
Conclusion

○○○



2 / 25

Augustin Godinot



Example: Moderation on ☒

Data: Tweets with hate annotation

Model: Transformer & gradient descent

ML systems

Data in, Prediction out

Context

○○●○○○○○○

P₁: Known model

○○

P₂: Labeled data

○

Change detection

○○○○○○

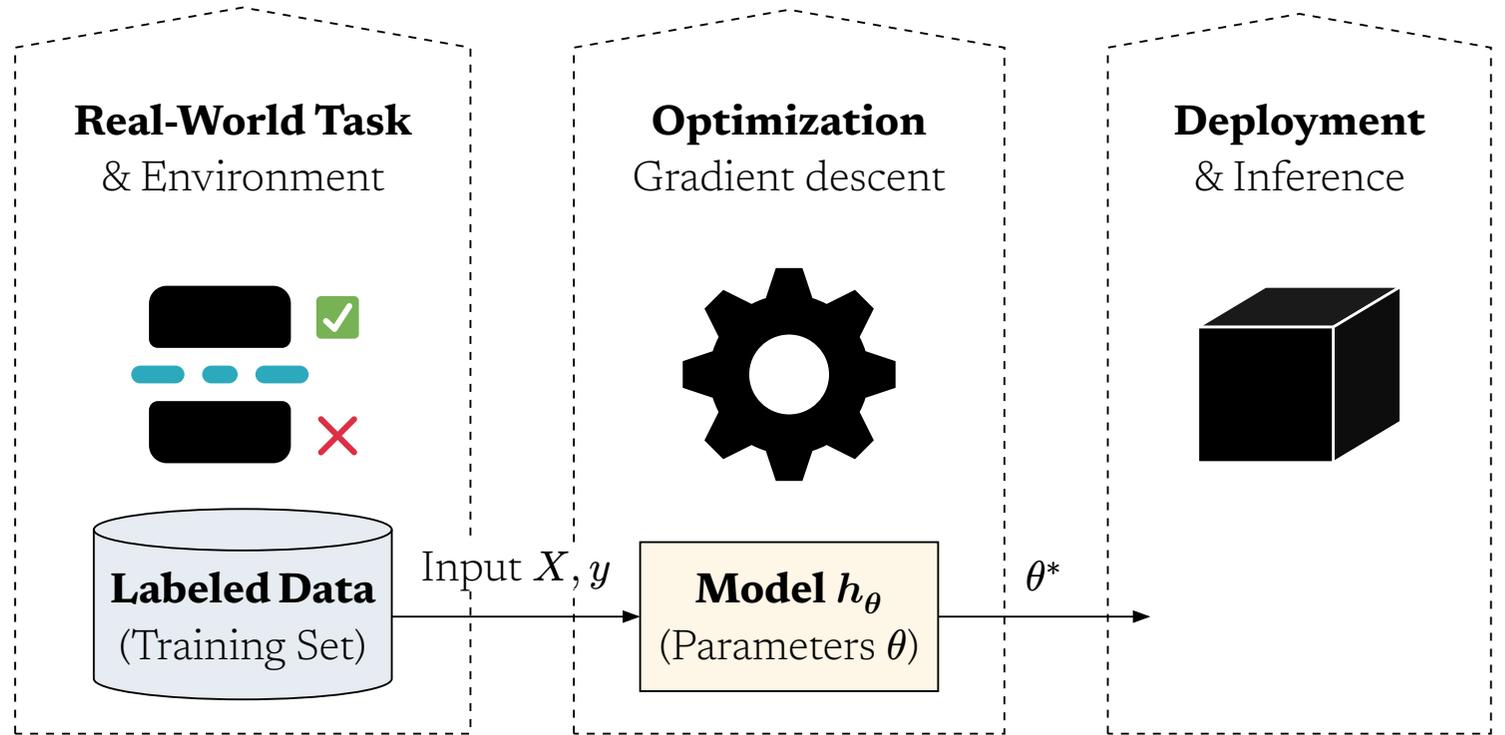
Conclusion

○○○



2 / 25

Augustin Godinot



Example: Moderation on ☒

Data: Tweets with hate annotation

Model: Transformer & gradient descent

Deployment: Thresholded hate scores

ML systems

Data in, Prediction out

Context

○○●○○○○○○

P₁: Known model

○○

P₂: Labeled data

○

Change detection

○○○○○○

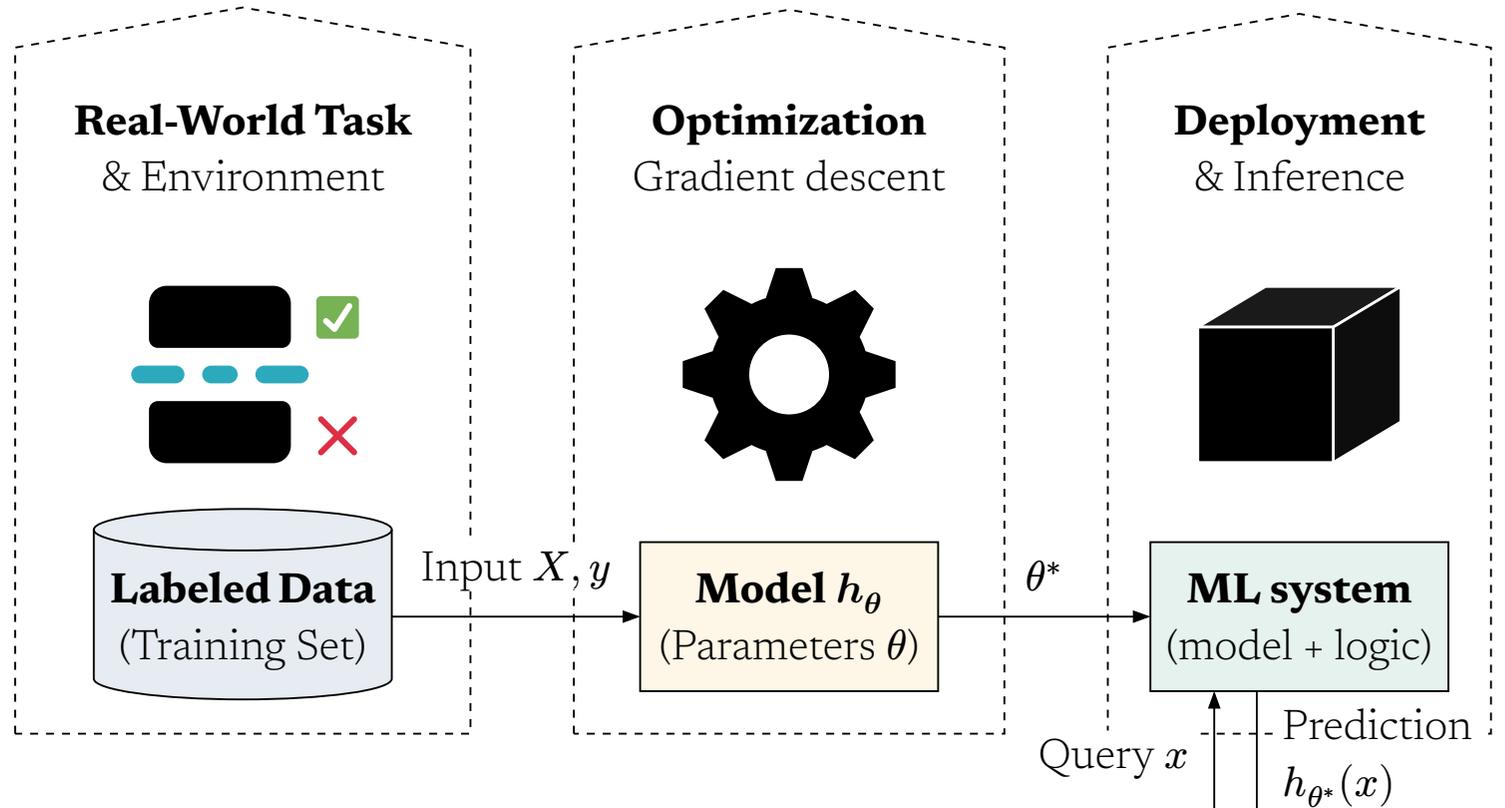
Conclusion

○○○



2 / 25

Augustin Godinot



Example: Moderation on ☒

Data: Tweets with hate annotation

Model: Transformer & gradient descent

Deployment: Thresholded hate scores

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.

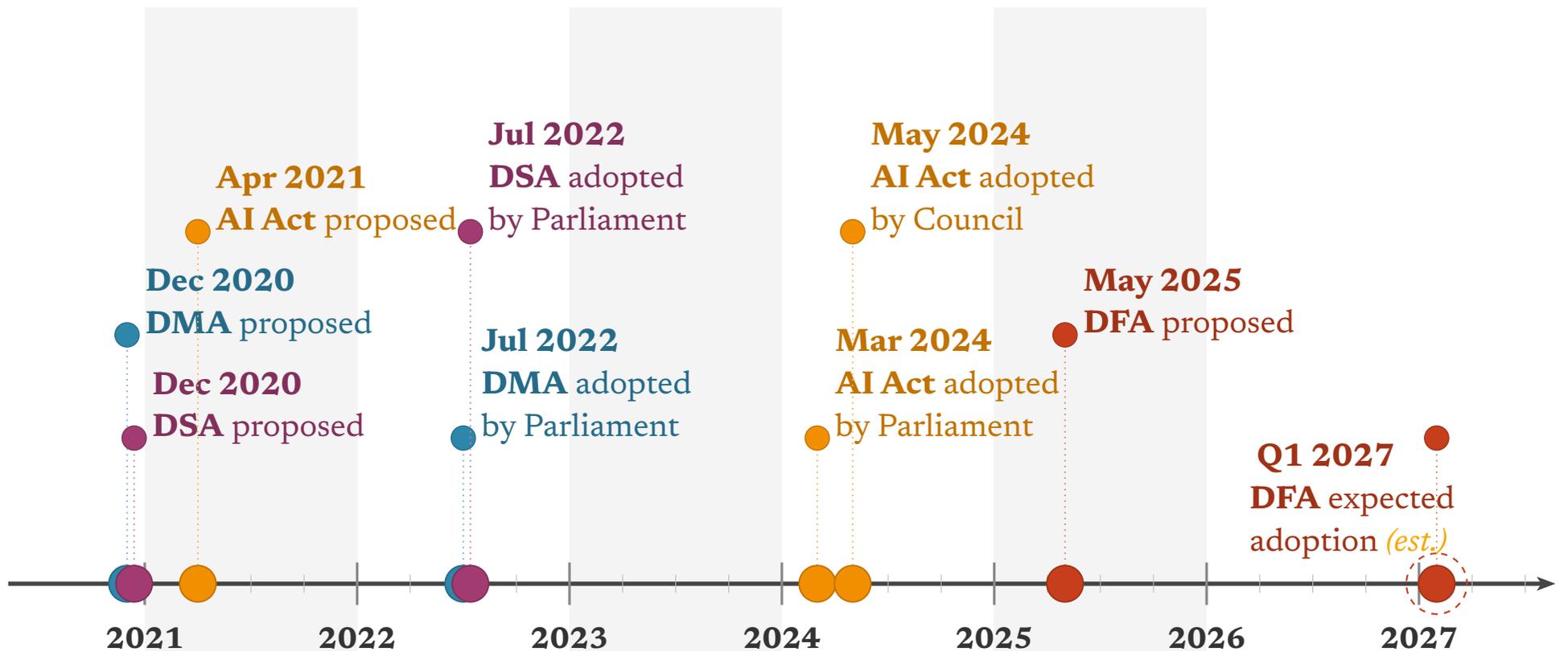
EVA CONSTANTARAS, GABRIEL GEIGER, JUSTIN-CASIMIR BRAUN, DHRUV MEHROTRA, HTET AUNG

MAR 6, 2023 7:00 AM



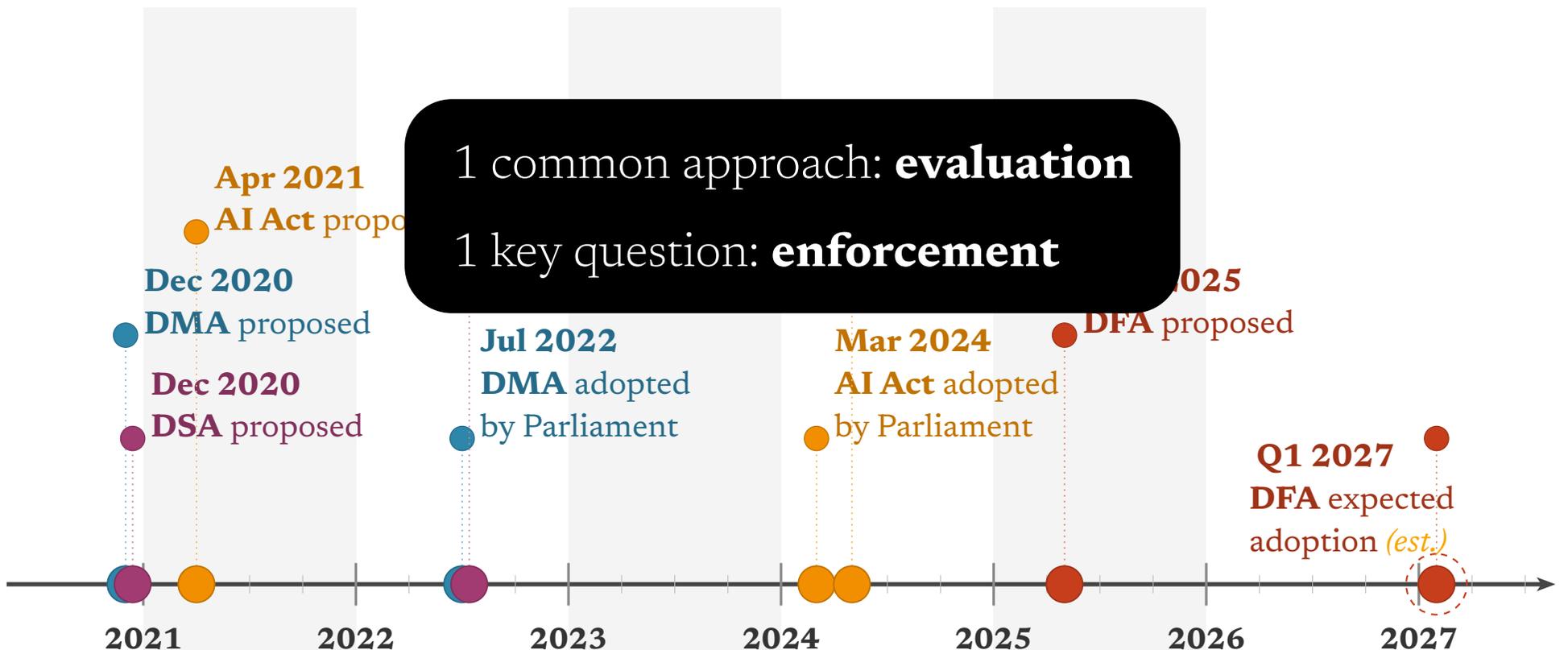
Main EU Regulation on ML systems

- DMA** Digital Markets Act
- DSA** Digital Services Act
- AI Act** Artificial Intelligence Act
- DFA** Digital Fairness Act



Main EU Regulation on ML systems

- DMA** Digital Markets Act
- DSA** Digital Services Act
- AI Act** Artificial Intelligence Act
- DFA** Digital Fairness Act



ML audits

Definition 1.2 (ML Audit) Any **independent** assessment of an identified **audit target** via an **evaluation** of articulated expectations with the implicit or explicit **objective of accountability**.

Context

○○○○○●○○○

P₁: Known model

○○

P₂: Labeled data

○

Change detection

○○○○○○

Conclusion

○○○



ML audits

Definition 1.2 (ML Audit) Any **independent** assessment of an identified **audit target** via an **evaluation** of articulated expectations with the implicit or explicit **objective of accountability**.
Audit target performance, disparity, robustness, privacy, ...

Context

○○○○○●○○○

P₁: Known model

○○

P₂: Labeled data

○

Change detection

○○○○○

Conclusion

○○○



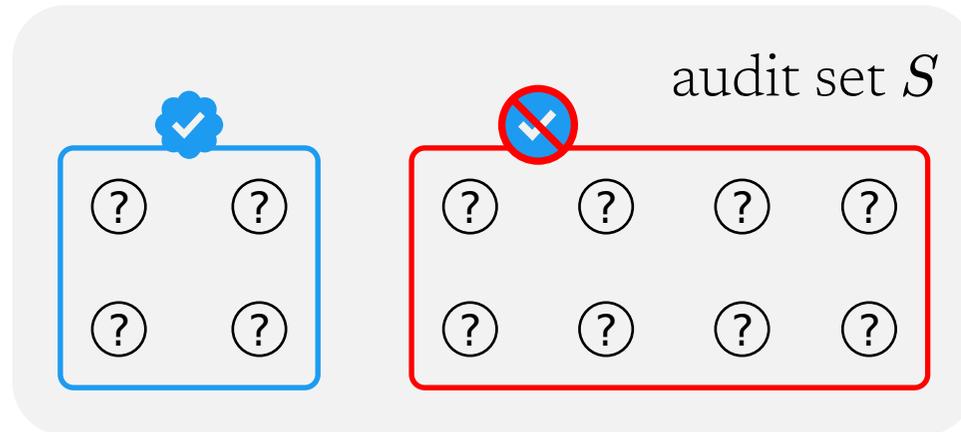
ML audits

Definition 1.2 (ML Audit) Any **independent** assessment of an identified **audit target** via an **evaluation** of articulated expectations with the implicit or explicit **objective of accountability**.

Audit target performance, disparity, robustness, privacy, ...

$$\mu(h, S) = \mathbb{P}\left(\text{👹} \mid X \in S, \text{✅}\right) - \mathbb{P}\left(\text{👹} \mid X \in S, \text{🚫}\right)$$

Demographic parity



Context

○○○○○●○○○

P_1 : Known model

○○

P_2 : Labeled data

○

Change detection

○○○○○○

Conclusion

○○○



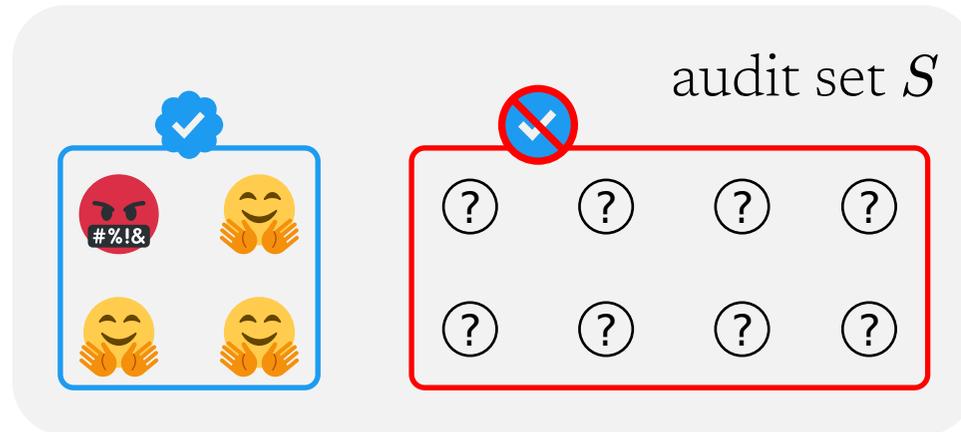
ML audits

Definition 1.2 (ML Audit) Any **independent** assessment of an identified **audit target** via an **evaluation** of articulated expectations with the implicit or explicit **objective of accountability**.

Audit target performance, disparity, robustness, privacy, ...

$$\begin{aligned}\mu(h, S) &= \mathbb{P}\left(\text{👁️} \mid X \in S, \text{✅}\right) - \mathbb{P}\left(\text{👁️} \mid X \in S, \text{🚫}\right) \\ &= \frac{1}{4}\end{aligned}$$

Demographic parity



Context

○○○○○●○○○

P₁: Known model

○○

P₂: Labeled data

○

Change detection

○○○○○○

Conclusion

○○○



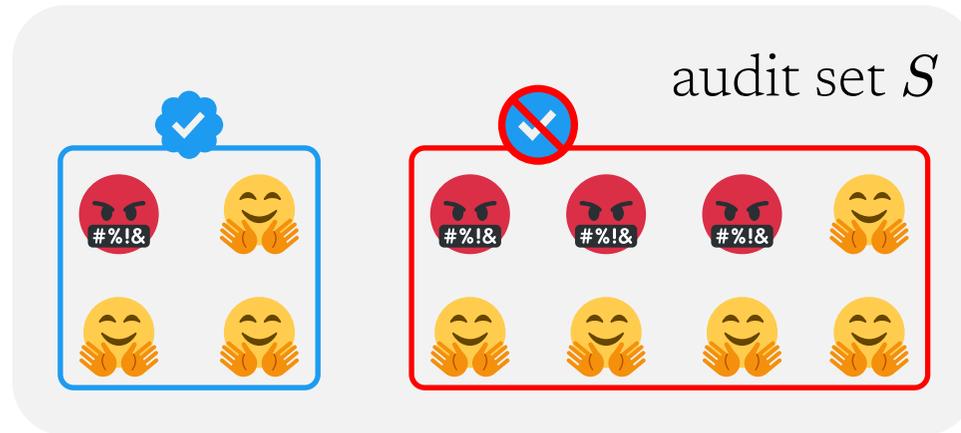
ML audits

Definition 1.2 (ML Audit) Any **independent** assessment of an identified **audit target** via an **evaluation** of articulated expectations with the implicit or explicit **objective of accountability**.

Audit target performance, disparity, robustness, privacy, ...

$$\begin{aligned}\mu(h, S) &= \mathbb{P}\left(\text{👎} \mid X \in S, \text{✅}\right) - \mathbb{P}\left(\text{👎} \mid X \in S, \text{🚫}\right) \\ &= \frac{1}{4} - \frac{3}{8} =\end{aligned}$$

Demographic parity



Context

○○○○○●○○○

P₁: Known model

○○

P₂: Labeled data

○

Change detection

○○○○○○

Conclusion

○○○



ML audits

Context

○○○○○●○○○

P₁: Known model

○○

P₂: Labeled data

○

Change detection

○○○○○○

Conclusion

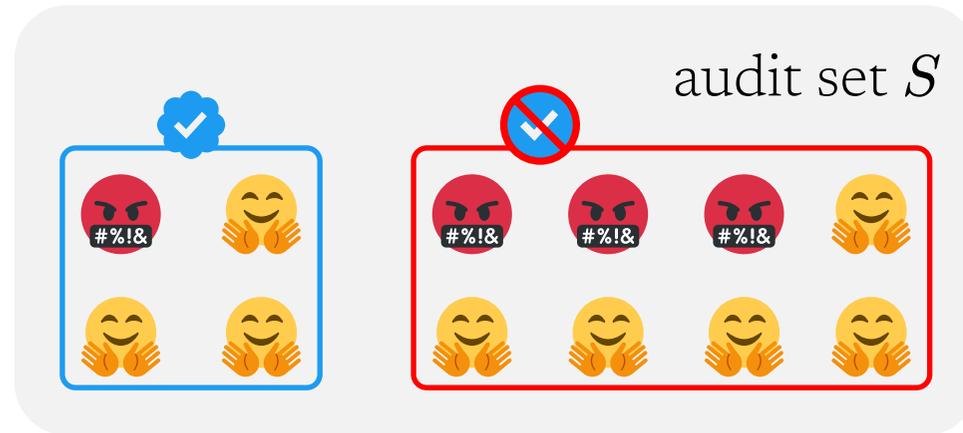
○○○

Definition 1.2 (ML Audit) Any **independent** assessment of an identified **audit target** via an **evaluation** of articulated expectations with the implicit or explicit **objective of accountability**.

Audit target performance, disparity, robustness, privacy, ...

$$\begin{aligned}\mu(h, S) &= \mathbb{P}\left(\text{👎} \mid X \in S, \text{✅}\right) - \mathbb{P}\left(\text{👎} \mid X \in S, \text{🚫}\right) \\ &= \frac{1}{4} - \frac{3}{8} = -12.5\%\end{aligned}$$

Demographic parity



ML audits

Definition 1.2 (ML Audit) Any **independent** assessment of an identified **audit target** via an **evaluation** of articulated expectations with the implicit or explicit **objective of accountability**.

Evaluation limited access to model, data, documentation

Context

○○○○○○●○○

P₁: Known model

○○

P₂: Labeled data

○

Change detection

○○○○○○

Conclusion

○○○



ML audits

Definition 1.2 (ML Audit) Any **independent** assessment of an identified **audit target** via an **evaluation** of articulated expectations with the implicit or explicit **objective of accountability**.

Evaluation limited access to model, data, documentation

Context

○○○○○○●○○

P₁: Known model

○○

P₂: Labeled data

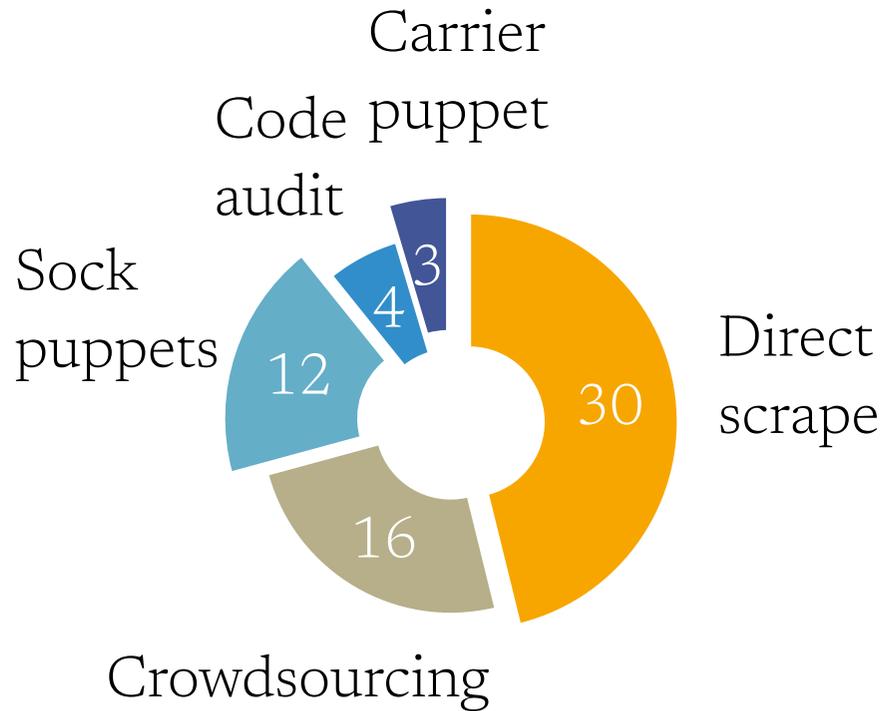
○

Change detection

○○○○○○

Conclusion

○○○



[1] Jack Bandy. 2021. Problematic Machine Behavior: A Systematic Literature Review of Algorithm Audits. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (April 2021), 74:1–74:34. <https://doi.org/10.1145/3449148>



ML audits

Definition 1.2 (ML Audit) Any **independent** assessment of an identified **audit target** via an **evaluation** of articulated expectations with the implicit or explicit **objective of accountability**.

Evaluation limited access to model, data, documentation

Context

○○○○○○●○○

P₁: Known model

○○

P₂: Labeled data

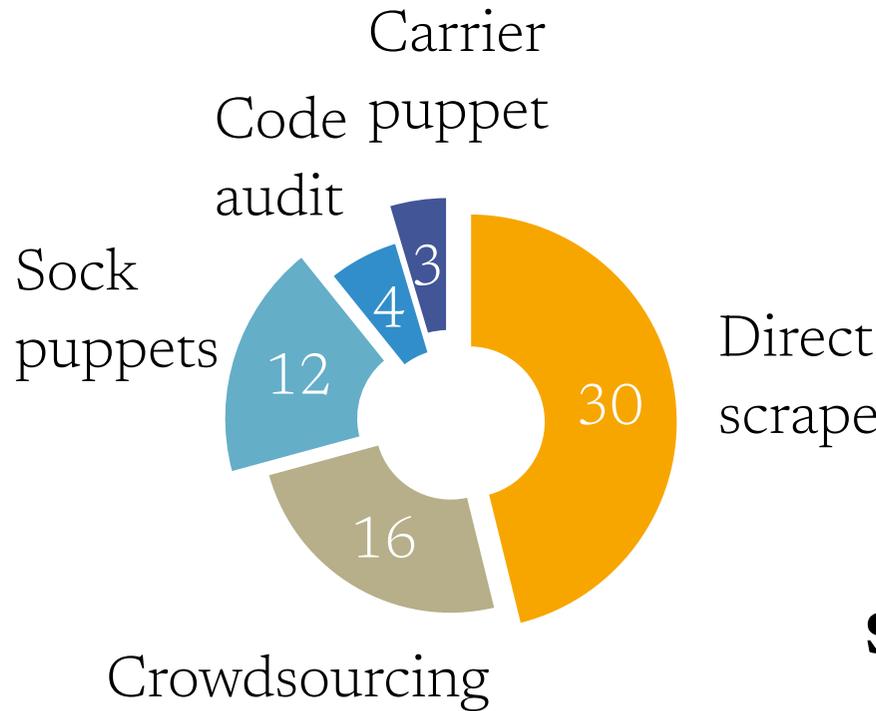
○

Change detection

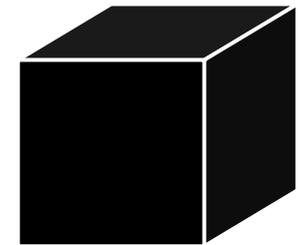
○○○○○○

Conclusion

○○○



query x



prediction $h(x)$

System = Black-box



ML audits

Definition 1.2 (ML Audit) Any **independent** assessment of an identified **audit target** via an **evaluation** of articulated expectations with the implicit or explicit **objective of accountability**.

Context

○○○○○○●○○

P₁: Known model

○○

P₂: Labeled data

○

Change detection

○○○○○○

Conclusion

○○○



Power/**data asymmetry** + **Regulation**

27.10.2022

EN

Official Journal of the European Union

L 277/1

REGULATION (EU) 2022/2065 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

of 19 October 2022

on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act)

10

prediction $h(x)$

System = Black-box

Crowdsourcing

[1] Jack Bandy. 2021. Problematic Machine Behavior: A Systematic Literature Review of Algorithm Audits. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (April 2021), 74:1–74:34. <https://doi.org/10.1145/3449148>

ML audits

Definition 1.2 (ML Audit) Any **independent** assessment of an identified **audit target** via an **evaluation** of articulated expectations with the implicit or explicit **objective of accountability**.

Context

○○○○○○●○○

P₁: Known model

○○

P₂: Labeled data

○

Change detection

○○○○○○

Conclusion

○○○



Power/**data asymmetry** + **Regulation**
⇒ **rationalization risk**

27.10.2022 EN Official Journal of the European Union L 277/1

REGULATION (EU) 2022/2065 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL
of 19 October 2022
on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act)

prediction $h(x)$
System = Black-box

Crowdsourcing

[1] Jack Bandy. 2021. Problematic Machine Behavior: A Systematic Literature Review of Algorithm Audits. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (April 2021), 74:1–74:34. <https://doi.org/10.1145/3449148>

Evading black-box audits

Context

○○○○○○○●○

P₁: Known model

○○

P₂: Labeled data

○

Change detection

○○○○○○

Conclusion

○○○



Auditor



Platform



User



The M\$ question

To go beyond the black-box model, what minimal additional information should the auditor request to achieve meaningful audit validity guarantees?

Context

○○○○○○○○●

P₁: Known model

○○

P₂: Labeled data

○

Change detection

○○○○○○

Conclusion

○○○

Context

▶ **P₁: Known model**

▶ **P₂: Labeled data**

Robustness during
the audit

Change detection

Verification after
the audit

Conclusion



P₁: Known model

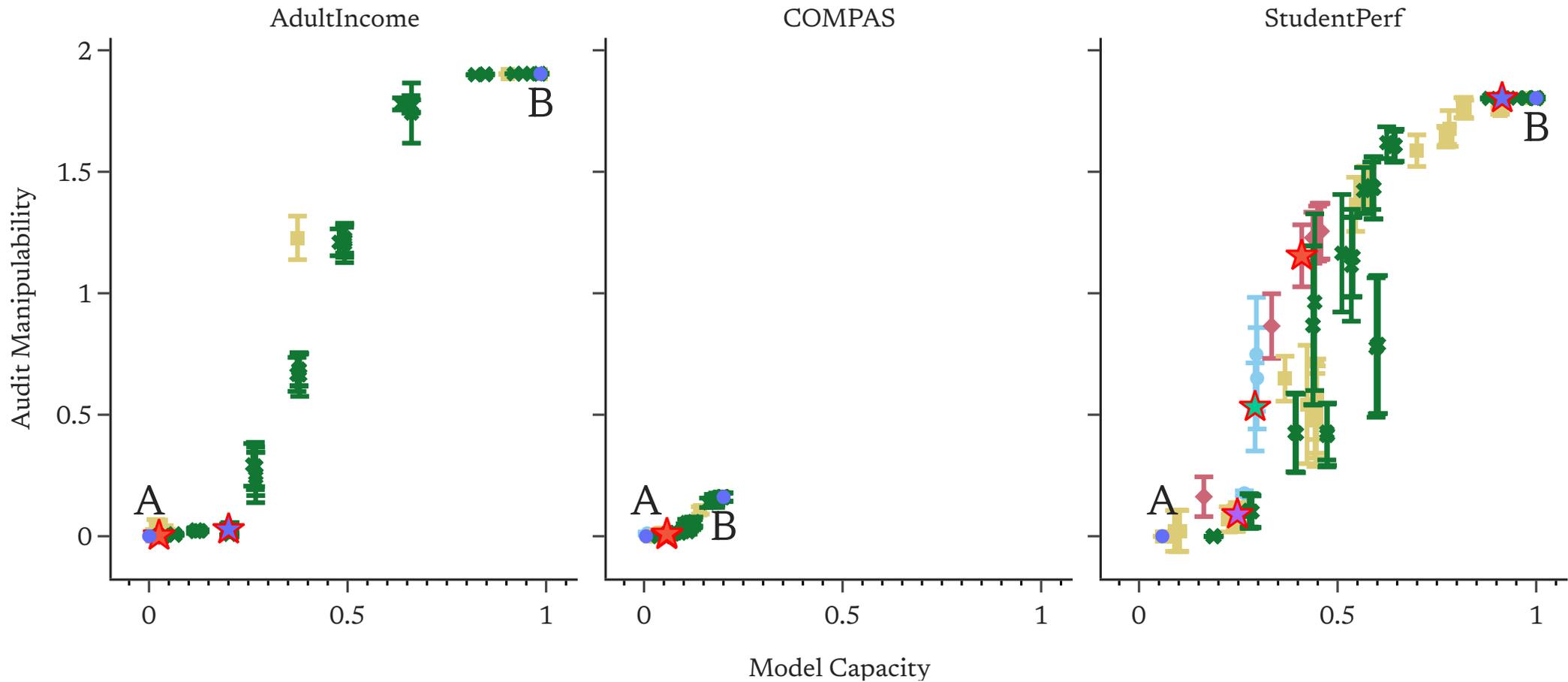


SaTML
2024

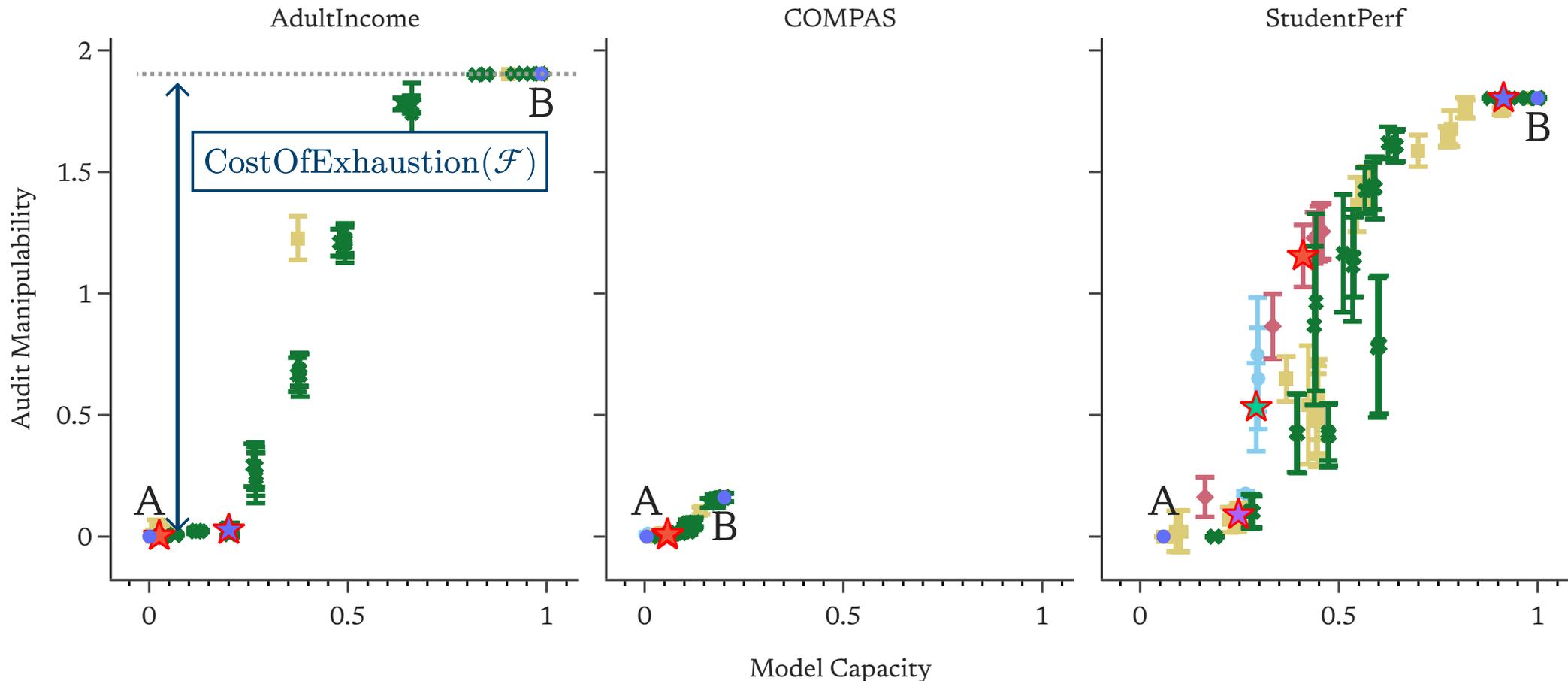
Under manipulations, are some AI models harder to audit?

*Augustin Godinot, Gilles Tredan, Erwan Le
Merrer, Camilla Penzo, Francois Taiani*

● perceptron
 ◆ linear
 ■ tree
 ✖ gbdt
 ★ \mathcal{H}_{opt}



● perceptron
 ◆ linear
 ■ tree
 ✖ gbdt
 ★ \mathcal{H}_{opt}



$$\text{AuditManipulability}(\mathcal{H}) = \mathbb{E}_{S \sim \mathcal{X}^m} [\text{diam}_{\mu}(h^*, S)]$$

$$\text{ModelCapacity}(\mathcal{H}) = \mathbb{E}_{D \sim \mathcal{X}^r} [\text{Rademacher}(\mathcal{H}, D)]$$

Model capacity is key

Context

oooooooo

P₁: Known model

●

P₂: Labeled data

○

Change detection

oooooo

Conclusion

ooo

It seems [...] a platform could always *game the system* [...] *without sacrificing a lot of accuracy* of the model learnt.

– Anonymous reviewer

Conclusion

- ▶ Easy to overfit the audit set with low utility cost
- ▶ Active auditing is not robust despite computational cost
- ▶ Regulators need to have more information than just the hypothesis class



P₂: Labeled data



ICML
2025

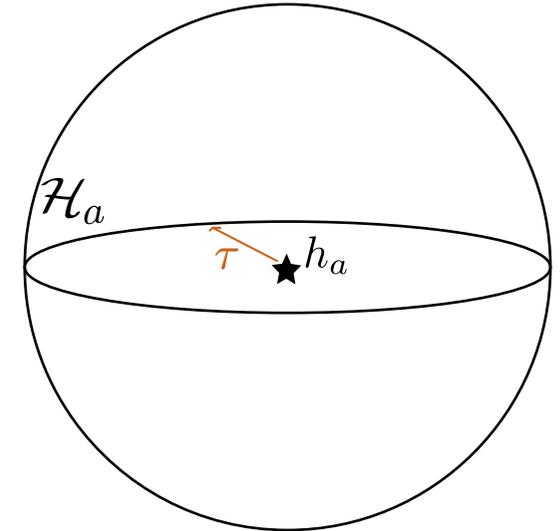
Robust ML Auditing using Prior Knowledge
*Jade Garcia Bourrée**, *Augustin Godinot**,
*Sayan Biswas, Anne-Marie Kermarrec, Erwan
Le Merrer, Gilles Tredan, Martijn de Vos, Milos
Vujasinovic* (Spotlight poster)

Data prior

Auditors have **labeled data**, how to take it into account?

Definition 2.3 (Dataset prior)

$$\mathcal{H}_a = \{h \in \mathcal{Y}^{|D_a|} \mid L(h, D_a) < \tau\}$$



Context

oooooooo

P₁: Known model

oo

P₂: Labeled data

●

Change detection

oooooo

Conclusion

ooo



Data prior

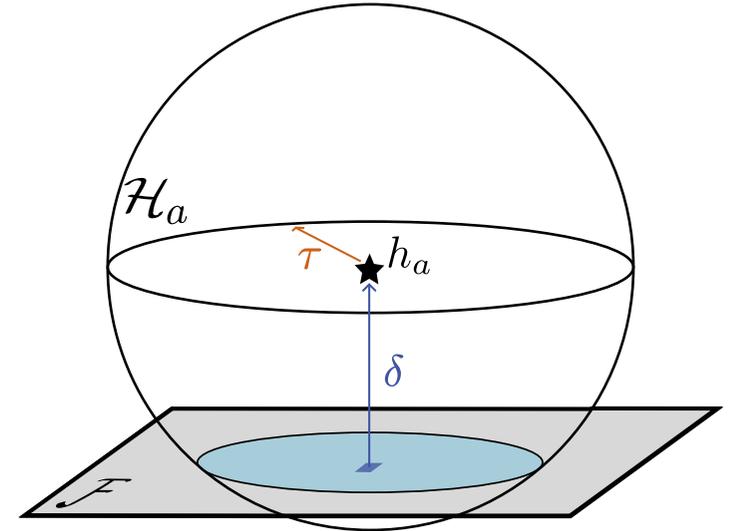
Auditors have **labeled data**, how to take it into account?

Definition 2.3 (Dataset prior)

$$\mathcal{H}_a = \{h \in \mathcal{Y}^{|D_a|} \mid L(h, D_a) < \tau\}$$

Set of fair labelings

$$\mathcal{F} = \{h \in \mathcal{Y}^{|D_a|} \mid \mu(h, D_a) = 0\}$$



Context

oooooooo

P₁: Known model

oo

P₂: Labeled data

●

Change detection

oooooo

Conclusion

ooo



Data prior

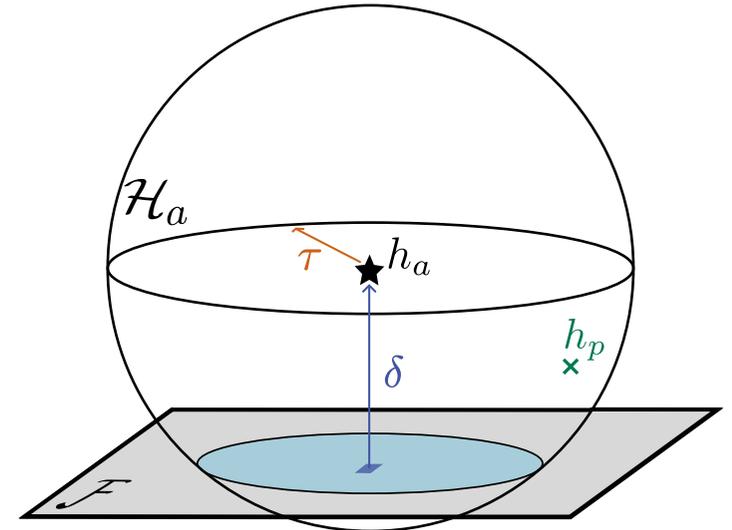
Auditors have **labeled data**, how to take it into account?

Definition 2.3 (Dataset prior)

$$\mathcal{H}_a = \{h \in \mathcal{Y}^{|D_a|} \mid L(h, D_a) < \tau\}$$

Set of fair labelings

$$\mathcal{F} = \{h \in \mathcal{Y}^{|D_a|} \mid \mu(h, D_a) = 0\}$$



Context

oooooooo

P₁: Known model

oo

P₂: Labeled data

●

Change detection

ooooo

Conclusion

ooo



Data prior

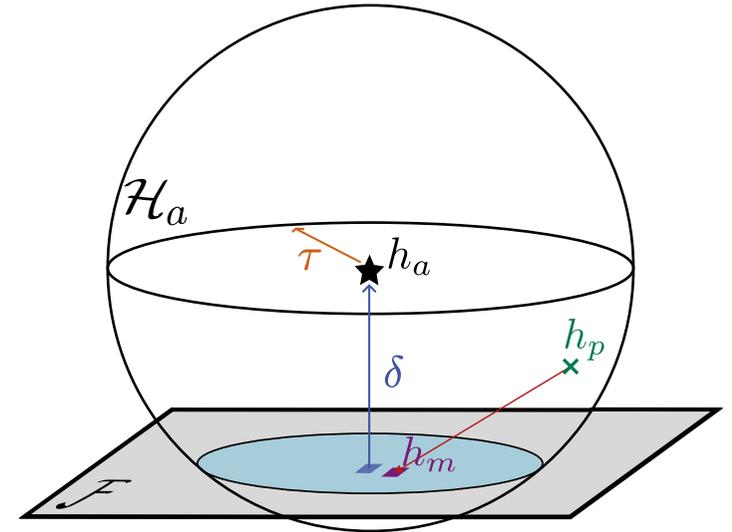
Auditors have **labeled data**, how to take it into account?

Definition 2.3 (Dataset prior)

$$\mathcal{H}_a = \{h \in \mathcal{Y}^{|D_a|} \mid L(h, D_a) < \tau\}$$

Set of fair labelings

$$\mathcal{F} = \{h \in \mathcal{Y}^{|D_a|} \mid \mu(h, D_a) = 0\}$$



Context

oooooooo

P₁: Known model

oo

P₂: Labeled data

●

Change detection

ooooo

Conclusion

ooo



Data prior

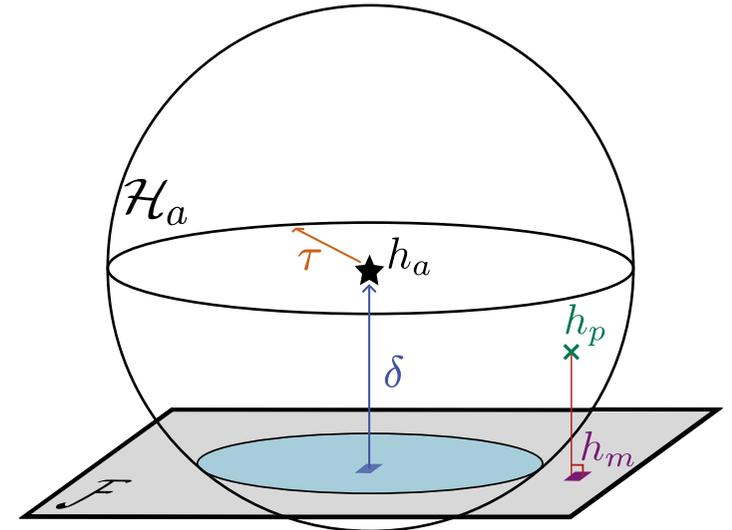
Auditors have **labeled data**, how to take it into account?

Definition 2.3 (Dataset prior)

$$\mathcal{H}_a = \{h \in \mathcal{Y}^{|D_a|} \mid L(h, D_a) < \tau\}$$

Set of fair labelings

$$\mathcal{F} = \{h \in \mathcal{Y}^{|D_a|} \mid \mu(h, D_a) = 0\}$$



Context

oooooooo

P₁: Known model

oo

P₂: Labeled data

●

Change detection

ooooo

Conclusion

ooo



Data prior

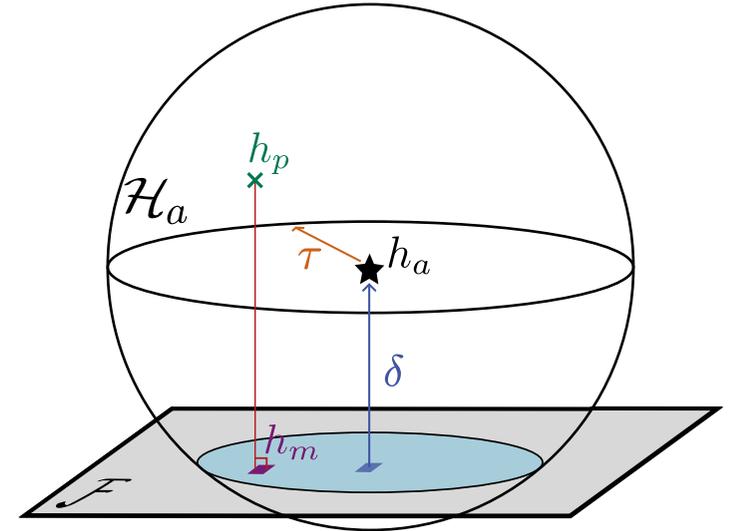
Auditors have **labeled data**, how to take it into account?

Definition 2.3 (Dataset prior)

$$\mathcal{H}_a = \{h \in \mathcal{Y}^{|D_a|} \mid L(h, D_a) < \tau\}$$

Set of fair labelings

$$\mathcal{F} = \{h \in \mathcal{Y}^{|D_a|} \mid \mu(h, D_a) = 0\}$$



Context

oooooooo

P₁: Known model

oo

P₂: Labeled data

●

Change detection

ooooo

Conclusion

ooo



Data prior

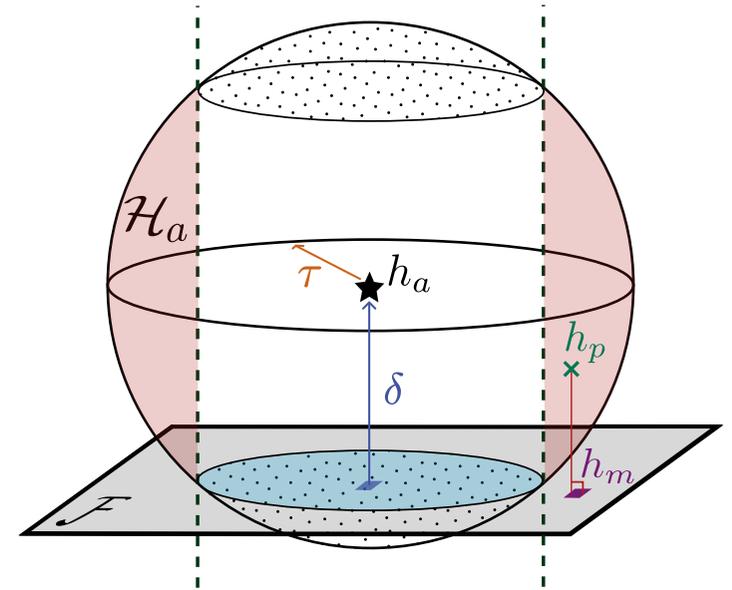
Auditors have **labeled data**, how to take it into account?

Definition 2.3 (Dataset prior)

$$\mathcal{H}_a = \{h \in \mathcal{Y}^{|D_a|} \mid L(h, D_a) < \tau\}$$

Set of fair labelings

$$\mathcal{F} = \{h \in \mathcal{Y}^{|D_a|} \mid \mu(h, D_a) = 0\}$$



Context

oooooooo

P₁: Known model

oo

P₂: Labeled data

●

Change detection

oooooo

Conclusion

ooo



Data prior

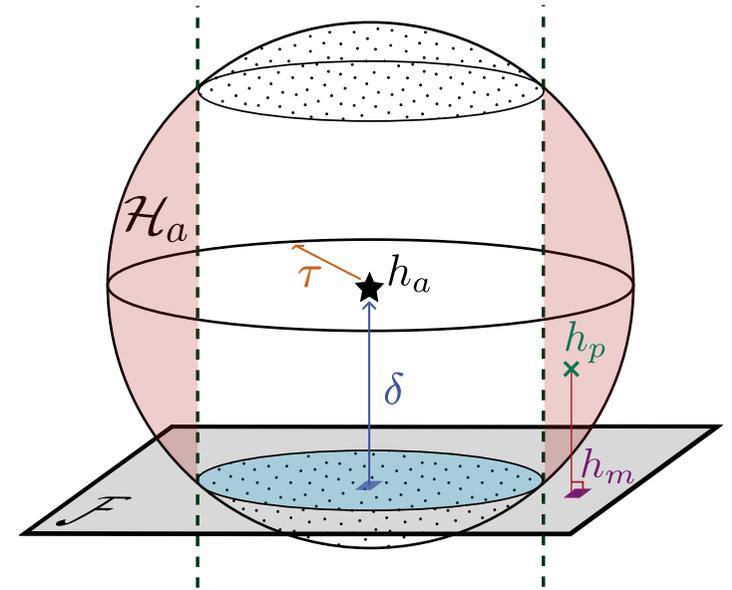
Auditors have **labeled data**, how to take it into account?

Definition 2.3 (Dataset prior)

$$\mathcal{H}_a = \{h \in \mathcal{Y}^{|D_a|} \mid L(h, D_a) < \tau\}$$

Set of fair labelings

$$\mathcal{F} = \{h \in \mathcal{Y}^{|D_a|} \mid \mu(h, D_a) = 0\}$$



Interesting cases

- ▶ Fair audit set ($\delta = \mu(h_a, D_a) = 0$) \Rightarrow *impossible detection*
- ▶ Perfect threshold ($\tau = \delta$) \Rightarrow *always detected*
- ▶ In between: exact formulation (Thm. 2.2), bounds (Cor 2.3)

Context

oooooooo

P₁: Known model

oo

P₂: Labeled data

●

Change detection

oooooo

Conclusion

ooo

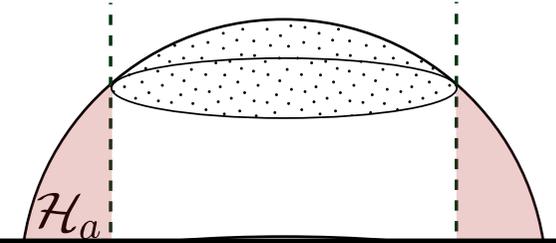


Data prior

Auditors have **labeled data**, how to take it into account?

Definition 2.3 (Dataset prior)

$$\mathcal{H}_a = \{h \in \mathcal{Y}^{|D_a|} \mid L(h, D_a) < \tau\}$$



The good regulator theorem [3]

“Every good regulator of a system must be a model of that system”

Interesting cases

- ▶ Fair audit set ($\delta = \mu(h_a, D_a) = 0$) \Rightarrow *impossible detection*
- ▶ Perfect threshold ($\tau = \delta$) \Rightarrow *always detected*
- ▶ In between: exact formulation (Thm. 2.2), bounds (Cor 2.3)

Context

oooooooo

P₁: Known model

oo

P₂: Labeled data

●

Change detection

oooooo

Conclusion

ooo



Change detection



AAAI
2025

Queries, Representation & Detection: the
next 100 model fingerprinting schemes

*Augustin Godinot, Gilles Tredan, Erwan Le
Merrer, Camilla Penzo, Francois Taiani*

Audit-then-verify

Context

○○○○○○○○○

P₁: Known model

○○

P₂: Labeled data

○

Change detection

●○○○○○

Conclusion

○○○



During the audit

- ▶ Code access
- ▶ Data access
- ▶ Model access
- ▶ Weights access

After the audit

- ▶ Query access



Existing methods

Context

○○○○○○○○○

P₁: Known model

○○

P₂: Labeled data

○

Change detection

○●○○○○

Conclusion

○○○

Interpretable Differencing of Machine Learning Models

Swagatam Haldar¹ Diptikalyan Saha¹ Dennis Wei² Rahul Nair³ Elizabeth M. Daly³

¹IBM Research, Bangalore, India
²IBM Research, Yorktown Heights, New York, USA,
³IBM Research, Dublin, Ireland

COMPARING DISTRIBUTIONS BY MEASURING DIFFERENCES THAT AFFECT DECISION MAKING

Shengjia Zhao*, Abhishek Sinha*, Yutong He*, Aidan Perreault, Jiaming Song, Stefano Ermon
Department of Computer Science
Stanford University
{sjzhao, a7b23, kellyyhe, aperr, tsong, ermon}@stanford.edu

What Changed? Interpretable Model Comparison

Rahul Nair¹, Massimiliano Mattetti¹, Elizabeth Daly¹
Dennis Wei², Öznur Alkan¹, Yunfeng Zhang²

¹IBM Research Europe
²IBM Research Yorktown

{rahul.nair@ie., massimiliano.mattetti@, elizabeth.daly@ie.}ibm.com,
{dwei@us., OAlkan2@ie., zhangyun@us.}ibm.com

A ZEST OF LIME: TOWARDS ARCHITECTURE-INDEPENDENT MODEL DISTANCES

Hengrui Jia, Hongyu Chen, Jonas Guan
University of Toronto and Vector Institute
{nickhengrui.jia, hy.chen}@mail.utoronto.ca, jonas@cs.toronto.edu

Ali Shabin Shamsabadi
Vector Institute and The Alan Turing Institute
a.shahinshamsabadi@turing.ac.uk

ModelDiff: Testing-Based DNN Similarity Comparison for Model Reuse Detection

Yuanchun Li
Microsoft Research
Beijing, China
Yuanchun.Li@microsoft.com

Ziqi Zhang
Peking University
Beijing, China
ziqi_zhang@pku.edu.cn

Bingyan Liu
Peking University
Beijing, China
lby_cs@pku.edu.cn

Ziyue Yang
Microsoft Research
Beijing, China
ze.Yang@microsoft.com

Yunxin Liu
Institute for AI Industry Research (AIR), Tsinghua University
Beijing, China
liuyunxin@air.tsinghua.edu.cn

IPGuard: Protecting Intellectual Property of Deep Neural Networks via Fingerprinting the Classification Boundary

Nicolas Papernot

Xiaoyu Cao
Duke University
xiaoyu.cao@duke.edu

Jinyuan Jia
Duke University
jinyuan.jia@duke.edu

Neil Zhenqiang Gong
Duke University
neil.gong@duke.edu

Model Fingerprinting with Benign Inputs

Publisher: IEEE [Cite This](#) [PDF](#)

Thibault Maho; Teddy Furon; Erwan Le Merrer **All Authors**

Are You Stealing My Model? Sample Correlation for Fingerprinting Deep Neural Networks

Jiyang Guan^{1,2}, Jian Liang^{1,2}, Ran He^{1,2*}

¹NLPR & CRIPAC, Institute of Automation, Chinese Academy of Sciences, China
²School of Artificial Intelligence, University of Chinese Academy of Sciences, China
guanjiyang2020@ia.ac.cn, liangjian92@gmail.com, zhe@nlpr.ia.ac.cn



Existing methods

Context

○○○○○○○○○

P₁: Known model

○○

P₂: Labeled data

○

Change detection

○●○○○○

Conclusion

○○○

Interpretable Differencing of Machine Learning Models

COMPARING DISTRIBUTIONS BY MEASURING DIFFERENCES THAT AFFECT DECISION MAKING



When the literature gives you **Lemons**,
you make **Queries**,
Representation
& **Detection**

Thibault Maho; Teddy Furon; Erwan Le Merrer **All Authors**

Jiyang Guan¹, Jian Liang², Kai He²

¹NLPR & CRIPAC, Institute of Automation, Chinese Academy of Sciences, China

²School of Artificial Intelligence, University of Chinese Academy of Sciences, China

guanjiyang2020@ia.ac.cn, liangjian92@gmail.com, rhes@nlpr.ia.ac.cn



A model fingerprint baseline

Context

oooooooo

P₁: Known model

oo

P₂: Labeled data

o

Change detection

oo●ooo

Conclusion

ooo

Requires Sampling access to \mathcal{D} , white-box h , black-box h'

- 1 **Draw** $x \sim \overline{\mathcal{D}}_h$ (distribution on \mathcal{X} such that $h(x) \neq c(x)$)
- 2 **If** $h(x) = h'(x)$
- 3 | **Return** 1 (Stolen)
- 4 **Else Return** 0 (Benign)

Figure 4.2: The proposed baseline, AKH.



A model fingerprint framework

Context

oooooooo

P₁: Known model

oo

P₂: Labeled data

o

Change detection

ooo●oo

Conclusion

ooo

$$\mathcal{H}_0 : h \neq h' \quad \mathcal{H}_1 : h = h'$$

Queries

How to select

$$S = (x_1, \dots, x_s) \subset \mathcal{X}?$$

Representation

How to “compress” $Y = (h(x_1), \dots, h(x_s))$ into representation Z ?

Detection

How to make the decision $h = h'$ vs. $h \neq h'$ based on $d(Z, Z')$?



A model fingerprint benchmark

Context

oooooooo

P₁: Known model

oo

P₂: Labeled data

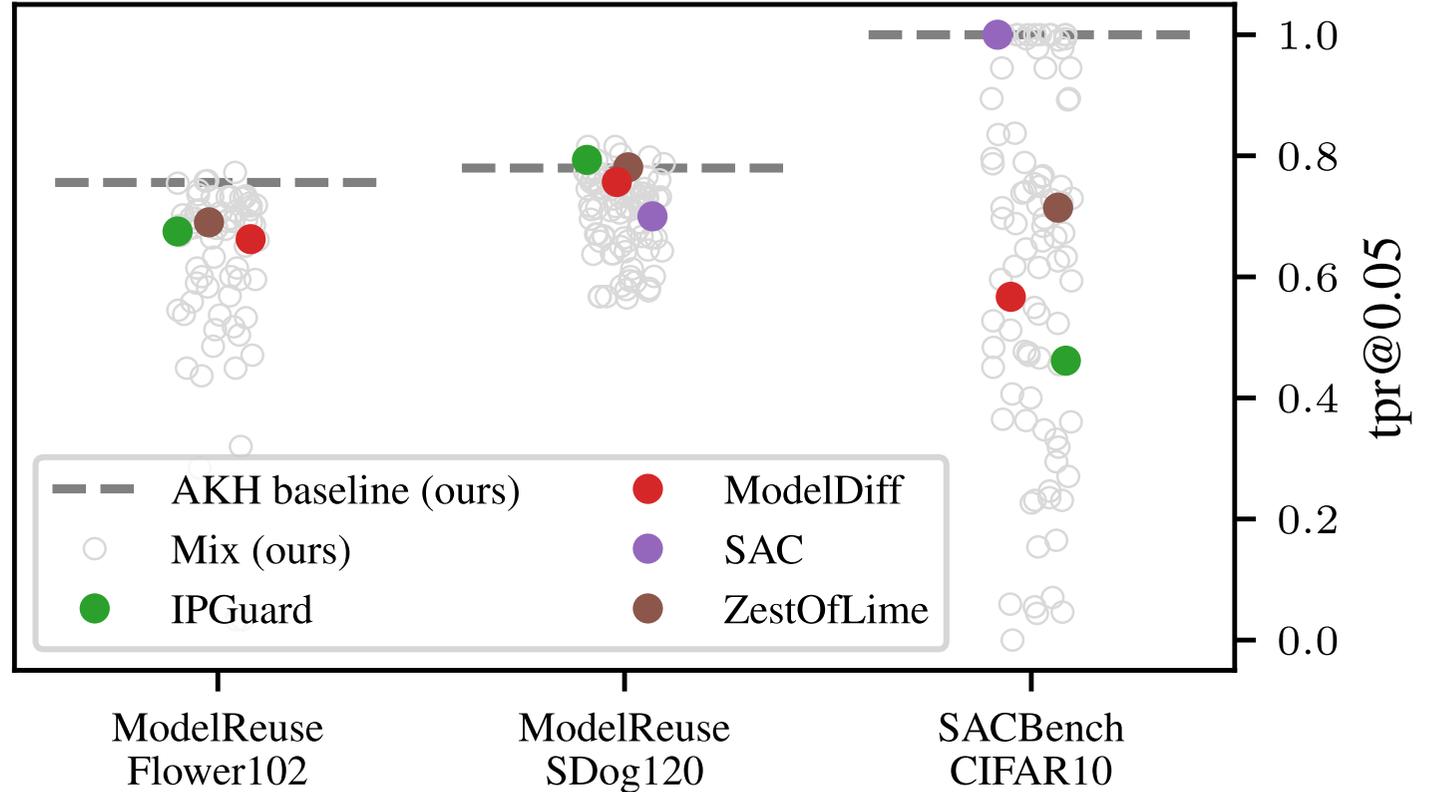
o

Change detection

oooo●o

Conclusion

ooo



Implications for re-audits

Context

○○○○○○○○○

P₁: Known model

○○

P₂: Labeled data

○

Change detection

○○○○○●

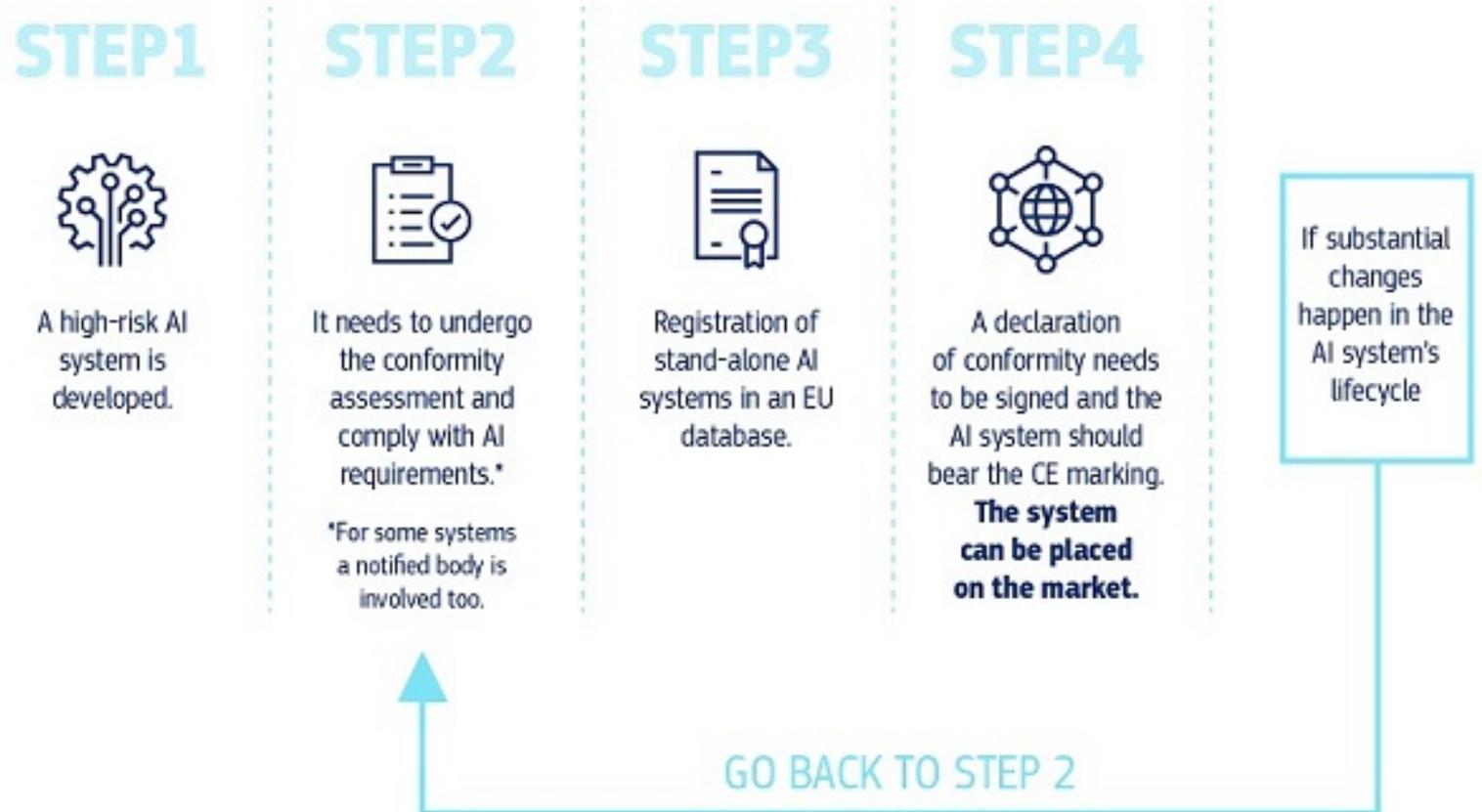
Conclusion

○○○



20 / 25

Augustin Godinot



Conclusion

Audits are a matter of

1. Comparison to priors
2. Information gain

Contributions

Context

oooooooooo

P₁: Known model

oo

P₂: Labeled data

o

Change detection

oooooo

Conclusion

●oo



Conferences



SaTML
2024

Under manipulations, are some AI models harder to audit?

Augustin Godinot, Gilles Tredan, Erwan Le Merrer, Camilla Penzo, Francois Taiani



AAAI
2025

Queries, Representation & Detection: the next 100 model fingerprinting schemes

Augustin Godinot, Gilles Tredan, Erwan Le Merrer, Camilla Penzo, Francois Taiani



ICML
2025

Robust ML Auditing using Prior Knowledge

Jade Garcia Bourrée, Augustin Godinot*, Sayan Biswas, Anne-Marie Kermarrec, Erwan Le Merrer, Gilles Tredan, Martijn de Vos, Milos Vujasinovic (Spotlight poster)*

Preprints



Soon

Manipulation-Proof Oblivious Audits against Deceptive Model Providers

Augustin Godinot, Sofiane Azogagh, Julien Ferry, Sébastien Gambs



Δ -audits: adaptive and manipulation-proof performance monitoring

Augustin Godinot, Mohammad Yaghini Nicolas Papernot

And also

Context

oooooooo

P₁: Known model

oo

P₂: Labeled data

o

Change detection

ooooo

Conclusion

ooo

Open Source

- ▶ <https://github.com/grodino/QuRD>
- ▶ <https://github.com/huggingface/pytorch-image-models>

Reading group

- ▶ <https://gitlab.inria.fr/WIDE/auditia/reading-group>

Teaching

- ▶ Convex & proximal optimization labs



Perspectives

Post audit verification

Context

oooooooo

P₁: Known model

oo

P₂: Labeled data

o

Change detection

oooooo

Conclusion

oo●



- ▶ Sensitivity of audit metric / platform utility
- ▶ Links with performance prediction



Perspectives

Post audit verification

Context

oooooooooo

P₁: Known model

oo

P₂: Labeled data

o

Change detection

oooooo

Conclusion

oo●

Tools beyond evaluation

- ▶ Pre-audits: incident databases
- ▶ During the audit: distributed user auditing
- ▶ Links with system specification

Auditing trade-offs

- ▶ Beyond estimation of the impact, reveal the choices
- ▶ Scaling laws against data/compute asymmetry
- ▶ Prediction uncertainty and fallback mechanisms



Appendix

Bibliography

- Jack Bandy. 2021. Problematic Machine Behavior: A Systematic Literature Review of Algorithm Audits.
- [1] *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (April 2021), 74:1–74:34. <https://doi.org/10.1145/3449148>
- Abeba Birhane, Ryan Steed, Victor Ojewale, Briana Vecchione, and Inioluwa Deborah Raji. 2024. AI Auditing: The Broken Bus on the Road to AI Accountability. In *2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, 2024. 612–643. <https://doi.org/10.1109/SaTML59370.2024.00037>
- [2]
- Roger C. Conant and W. Ross Ashby. 1970. Every good regulator of a system must be a model of that system. *International journal of systems science* 1, 2 (1970), 89–97.
- [3]

