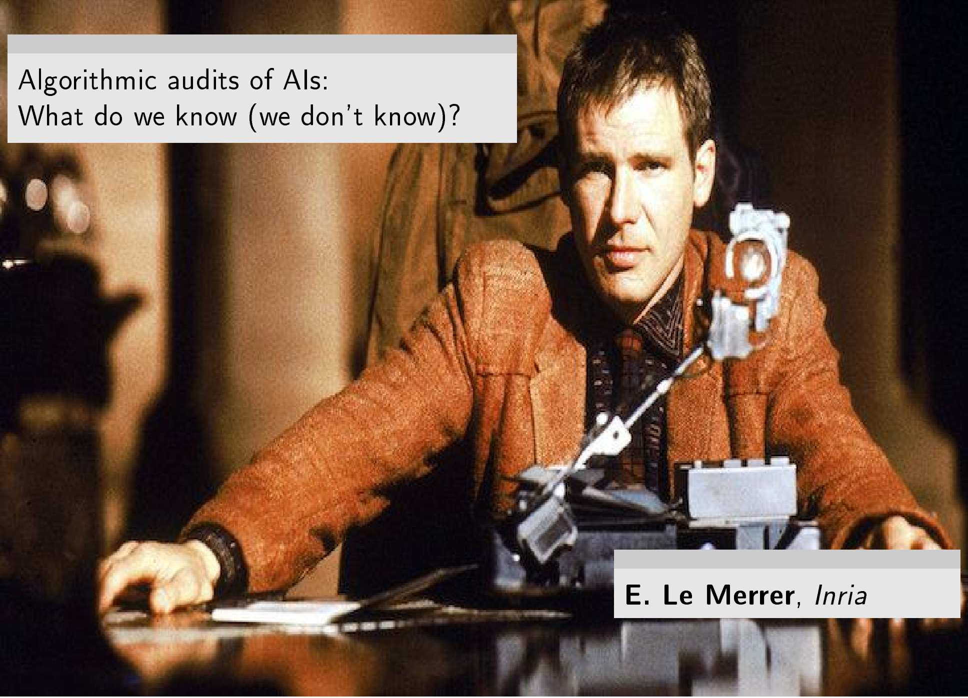
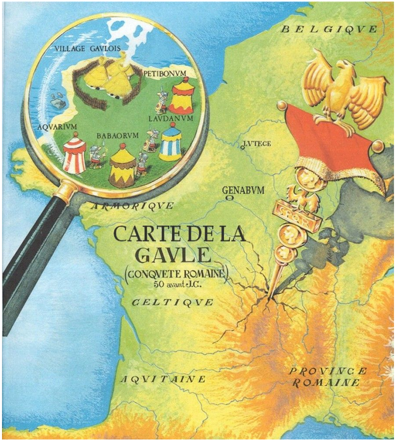
Algorithmic audits of AIs:
What do we know (we don't know)?



E. Le Merrer, *Inria*

Thanks #2



BELGIQVE
VILLAGE GAVLOIS
PETIBONVM
LAVDANVM
AQVARIVM
BABAORVM
LVTECE
GENABVM
ARMORIQVE
CARTE DE LA GAVLE
(CONQVETE ROMAINE)
50 avant J.C.
CELTIQVE
AQVITAINE
PROVINCE ROMAINE

Thanks #2

“petite balade”

Thanks #2

"petite balade"

- 3h "walk"

Thanks #2

"petite balade"

- 3h "walk"
- black trail with bumps

Thanks #2

"petite balade"

- 3h "walk"
- black trail with bumps
- frontally

Thanks #2

"petite balade"

- 3h "walk"
- black trail with bumps
- frontally
- iced snow

Blade Runner: the Voight-Kampff test

Is the remote entity a replicant ?

Essentially: investigation on questions/answers (inputs/outputs)

Today: ChatGPT or student?

Sung Kim

Dec 11, 2022 · 4 min read · Member-only · Listen

How to Detect OpenAI's ChatGPT Output

How to detect if the student used OpenAI's ChatGPT to complete an assignment

On November 30, 2022, OpenAI released 'ChatGPT' AI system (https://openai.com/blog/chatgpt/), which is a universal writer's assistant that can generate a variety of output, including school assignments. The output (e.g., essays) provided by ChatGPT is so good, if I was a student, I would be using ChatGPT to complete most of my school assignment with minor revisions.

Today: ChatGPT or student?

Can AI-Generated Text be Reliably Detected?

Vinu Sankar Sadasivan
vinu@umd.edu

Aounon Kumar
aounon@umd.edu

Sriram Balasubramanian
sriramb@umd.edu

Wenxiao Wang
wwx@umd.edu

Soheil Feizi
sfeizi@umd.edu

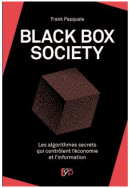
Department of Computer Science
University of Maryland

Abstract

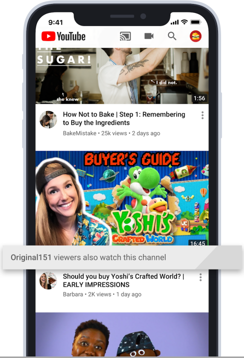
The rapid progress of Large Language Models (LLMs) has made them capable of performing astonishingly well on various tasks including document completion and question answering. The unregulated use of these models, however, can potentially lead to malicious consequences such as plagiarism, generating fake news, spamming, etc. Therefore, reliable detection of AI-generated text can be critical to ensure the responsible use of LLMs. Recent works attempt to tackle this problem either using certain model signatures present in the generated text outputs or by applying watermarking techniques that imprint specific patterns onto them. In this paper, both empirically and theoretically, we show that these detectors are not reliable in practical scenarios. Empirically, we show that *paraphrasing attacks*, where a light paraphraser is applied on top of the generative text model, can break a whole range of detectors, including the ones using the watermarking schemes as well as neural network-based detectors and zero-shot classifiers. We then provide a theoretical *impossibility result* indicating that for a sufficiently good language model, even the best-possible detector can only perform marginally better than a random classifier. Finally, we show that even LLMs protected by watermarking

No replicants yet, but pervasive decision-making AIs

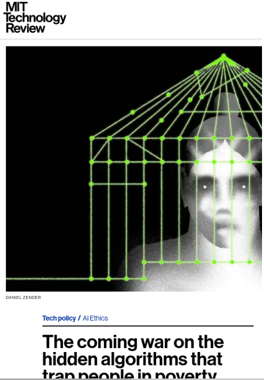
Why we need audits ?



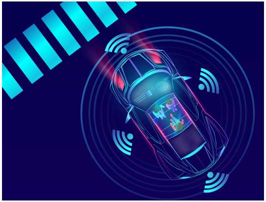
Recommendation



Credit scoring



Self driving cars



The ideology behind publishing Twitter's source code

A leak. On 31 March, Twitter published parts of the source code that powers its newsfeed. The move came a few days after it was made public that large portions of that code had been leaked on Github already [Gizmodo, 31 Mar].

The 85,797 lines of code contain little new information. Tweets that contain links are less likely to appear in a user's timeline. So are tweets in a language that the system cannot recognize – an obstacle for people whose vernaculars aren't on the radar of Californian engineers. Spaces (Twitter's live podcasting feature) about Ukraine seem to be hidden from view too [Aakash Gupta, 2 Apr].

The most interesting part of the release is the blog post written by Twitter's remaining engineering team. It provides a good high-level overview of how a newsfeed algorithm works.

How (not) to open source. One company led the way in making algorithms public: Twitter. Two years ago, its "Ethics, Transparency and Accountability" team released the code of an image-cropping algorithm and invited auditors to find possible biases [AlgorithmWatch, 2021]. The team was among the first to be fired last year.

You cannot audit code only by reading it. You need to run it on a computer. On Ukraine, for instance, we only know that Twitter Spaces labeled "UkraineCrisisTopic" undergo the same treatment as items labeled with violence or explicit content. But we don't know how the label is applied or what effects it has. It seems that the code responsible for that task has not even been made public.

Obfuscation. Publishing vast amounts of computer code without instructions can be worse than useless. It allows for claims of transparency while preventing any actual audit. Twitter is not the first

Pervasive decision-making AIs and new regulation

e.g. European Commission's Digital Service Act:

Today, the Commission also launched a call for evidence on the provisions in the DSA related to data access for researchers. These are designed to better monitor platform providers' actions to tackle illegal content, such as illegal hate speech, as well as other societal risks such as the spread of disinformation, and risks that may affect the users' mental health. Vetted researchers will have the possibility to access the data of any VLOP or VLOSE to conduct research on systemic risks in the EU. This means that they could for example analyse platforms' decisions on what users see and engage with online, having access to previously undisclosed data. In

+ the EU AI act

Pervasive decision-making AIs and new regulation

e.g. European Commission's Digital Service Act:

Today, the Commission also launched a call for evidence on the provisions in the DSA related to data access for researchers. These are designed to better monitor platform providers' actions to tackle illegal content, such as illegal hate speech, as well as other societal risks such as the spread of disinformation, and risks that may affect the users' mental health. Vetted researchers will have the possibility to access the data of any VLOP or VLOSE to conduct research on systemic risks in the EU. This means that they could for example analyse platforms' decisions on what users see and engage with online, having access to previously undisclosed data. In

+ the EU AI act

Problem: quite unclear yet how to do that, which algorithm/guarantees?

Pervasive decision-making AIs and new regulation

e.g. European Commission's Digital Service Act:

Today, the Commission also launched a call for evidence on the provisions in the DSA related to data access for researchers. These are designed to better monitor platform providers' actions to tackle illegal content, such as illegal hate speech, as well as other societal risks such as the spread of disinformation, and risks that may affect the users' mental health. Vetted researchers will have the possibility to access the data of any VLOP or VLOSE to conduct research on systemic risks in the EU. This means that they could for example analyse platforms' decisions on what users see and engage with online, having access to previously undisclosed data. In

+ the EU AI act

Problem: quite unclear yet how to do that, which algorithm/guarantees?

Inria's REGALIA

European Commission

Research and innovation

European Centre for Algorithmic Transparency

PEReN – Pôle d'Expertise de la Régulation Numérique

Algorithmic Audits vs Law (a word on...)

Legal implications of algorithmic **black box auditing**?

- Case study focuses (mainly) on France
- 2 canonical audit forms: Bobby and Sherlock

Consequences of the audit

- Legal risks for the auditor
- Probative value of the audit outcome

Algorithmic audits of algorithms, and the law. AI&Ethics Le Merrer, Pons and Tredan, 2023.

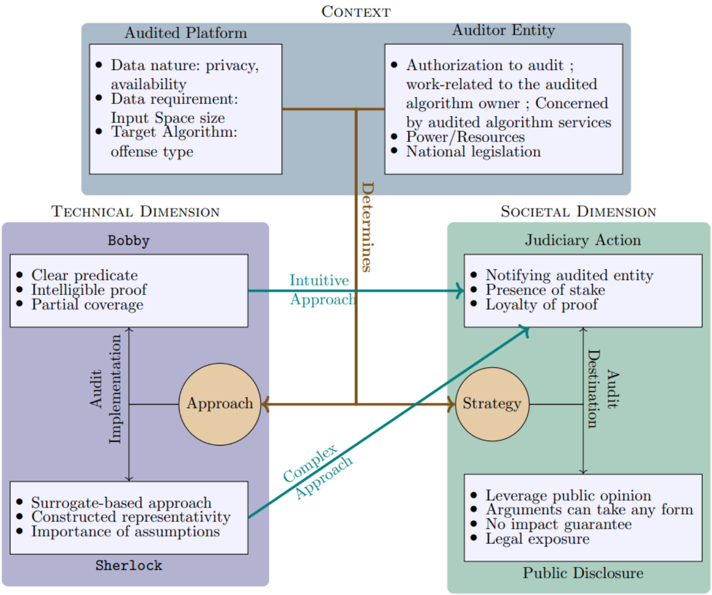

Bobby

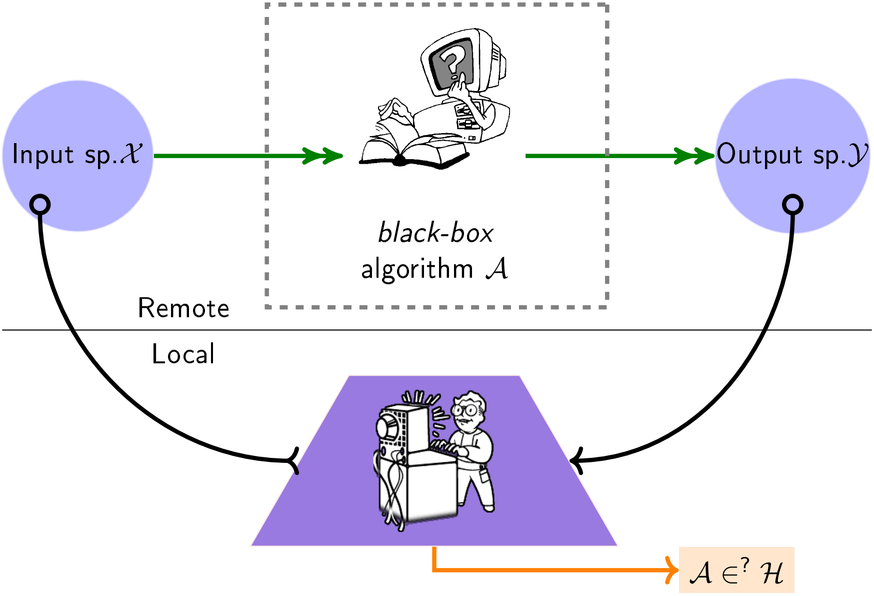

Sherlock

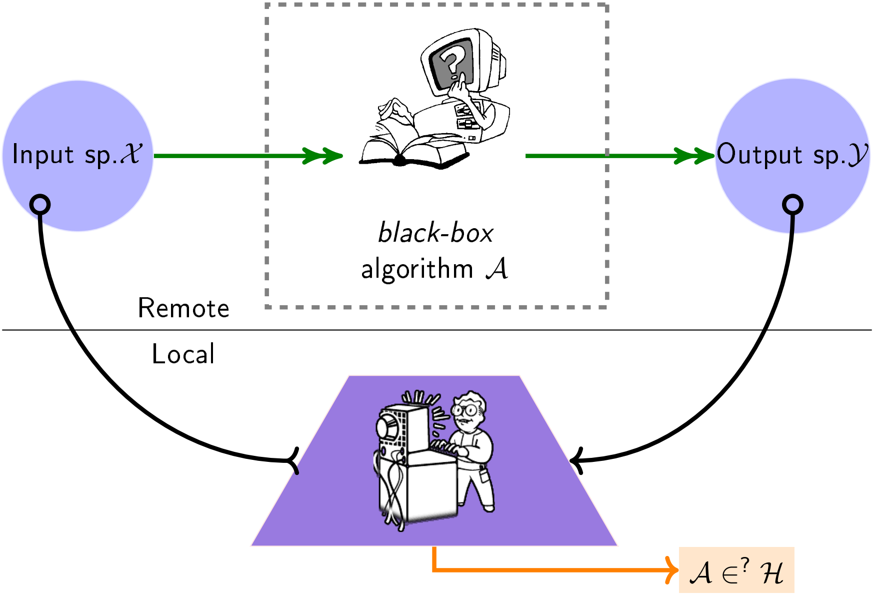
- (tours to find a well defined infraction predicate)
- e.g.: find copyright infringements or non-consented cookies; evaluate DI.

- Sherlock (constructs a surrogate model; somehow uses induction).
- e.g.: COMPAS study, LIME approaches, Uber surge price study.

Overview: a technico-legal mess...




Input sp.$\mathcal{X}$
Output sp.$\mathcal{Y}$
black-box
algorithm $\mathcal{A}$
Remote
Local
$\mathcal{A} \in^? \mathcal{H}$



and... link to security: information gain, algorithm leak, poisoning

An Input / Output example

Adult Census Income: task to predict whether income exceeds $50K/yr based on census data

Input:

# age	workclass	# fnlwgt	education	# education....	marital.sta...	occupation
90	?	77053	HS-grad	9	Widowed	?
82	Private	132870	HS-grad	9	Widowed	Exec-managerial
66	?	186061	Some-college	10	Widowed	?
54	Private	140359	7th-8th	4	Divorced	Machine-op-inspct

Output: Boolean (yes/no)

An Input / Output example

Adult Census Income: task to predict whether income exceeds $50K/yr based on census data

Input:

# age	A workclass	# fnlwgt	A education	# education....	A marital.sta...	A occupation
90	?	77053	HS-grad	9	Widowed	?
82	Private	132870	HS-grad	9	Widowed	Exec-managerial
66	?	186061	Some-college	10	Widowed	?
54	Private	140359	7th-8th	4	Divorced	Machine-op-inspct

Output: Boolean (yes/no)

Other examples:

- image (input) $\rightarrow$ label (output)
- user profile $\rightarrow$ item recommended

Sounds like related work? Property testing

Definition 1. *Let $\Pi = \bigcup_{n \in \mathbb{N}} \Pi_n$, where Π_n contains functions defined over the domain D_n. A* tester for a property Π *is a probabilistic oracle machine T that satisfies the following two conditions:*

1. *The tester accepts each $f \in \Pi$ with probability at least $2/3$; that is, for every $n \in \mathbb{N}$ and $f \in \Pi_n$* (and every $\epsilon > 0$), *it holds that* $\Pr[T^f(n, \epsilon) = 1] \geq 2/3$.
2. *Given $\epsilon > 0$ and oracle access to any f that is ϵ-far from Π, the tester rejects with probability at least $2/3$; that is, for every $\epsilon > 0$ and $n \in \mathbb{N}$, if $f : D_n \to R_n$ is ϵ-far from Π_n, then* $\Pr[T^f(n, \epsilon) = 0] \geq 2/3$, *where f is* ϵ-far from Π_n *if, for every $g \in \Pi_n$, it holds that* $|\{e \in D_n : f(e) \neq g(e)\}| > \epsilon \cdot n$.

Sounds like related work? Property testing

Definition 1. *Let $\Pi = \bigcup_{n \in \mathbb{N}} \Pi_n$, where Π_n contains functions defined over the domain D_n. A* tester for a property *Π is a probabilistic oracle machine T that satisfies the following two conditions:*

1. *The tester accepts each $f \in \Pi$ with probability at least $2/3$; that is, for every $n \in \mathbb{N}$ and $f \in \Pi_n$* (and every $\epsilon > 0$), *it holds that* $\Pr[T^f(n, \epsilon){=}1] \geq 2/3$.
2. *Given $\epsilon > 0$ and oracle access to any f that is ϵ-far from Π, the tester rejects with probability at least $2/3$; that is, for every $\epsilon > 0$ and $n \in \mathbb{N}$, if $f : D_n \to R_n$ is ϵ-far from Π_n, then* $\Pr[T^f(n, \epsilon){=}0] \geq 2/3$, *where f is* ϵ-far from *Π_n if, for every $g \in \Pi_n$, it holds that $|\{e \in D_n : f(e) \neq g(e)\}| > \epsilon \cdot n$.*

k-junta: if $f : \{0,1\}^n \to \{0,1\}$ depends on at most k variables

k-JUNTA TEST$(f,\ \epsilon)$
1. Randomly partition the coordinates into $O(k^2)$ buckets. 2. Run INDEPENDENCE TEST $\tilde{O}(k^2/\epsilon)$ times. 3. **Accept** iff at most k buckets fail the independence test.

Sounds like related work? Property testing

Definition 1. *Let $\Pi = \bigcup_{n \in \mathbb{N}} \Pi_n$, where Π_n contains functions defined over the domain D_n. A* tester for a property *Π is a probabilistic oracle machine T that satisfies the following two conditions:*

1. *The tester accepts each $f \in \Pi$ with probability at least $2/3$; that is, for every $n \in \mathbb{N}$ and $f \in \Pi_n$* (and every $\epsilon > 0$), *it holds that* $\Pr[T^f(n,\epsilon)=1] \geq 2/3$.
2. *Given $\epsilon > 0$ and oracle access to any f that is ϵ-far from Π, the tester rejects with probability at least $2/3$; that is, for every $\epsilon > 0$ and $n \in \mathbb{N}$, if $f : D_n \to R_n$ is ϵ-far from Π_n, then* $\Pr[T^f(n,\epsilon)=0] \geq 2/3$, *where f is ϵ*-far from *Π_n if, for every $g \in \Pi_n$, it holds that $|\{e \in D_n : f(e) \neq g(e)\}| > \epsilon \cdot n$.*

k-junta: if $f : \{0,1\}^n \to \{0,1\}$ depends on at most k variables

k-Junta Test(f, ϵ)

1. Randomly partition the coordinates into $O(k^2)$ buckets.
2. Run Independence Test $\tilde{O}(k^2/\epsilon)$ times.
3. **Accept** iff at most k buckets fail the independence test.

- interested in global function characteristics: intractable today
- assumes symmetry to $\downarrow$ complexity: problem for modern ML

Oded Goldreich: Property testing, current research and surveys.

Setting the record straight on shadow banning

By Vijaya Gadde and Kayvon Beykpour

Thursday, 26 July 2018

People are asking us if we shadow ban. We do not. But let's start with, "what is shadow banning?"

The best definition we found is this: deliberately making someone's content undiscoverable to everyone except the person who posted it, unbeknownst to the original poster.

We do not shadow ban. You are always able to see the tweets from accounts you follow (although you may have to do more work to find them, like go directly to their profile). And we certainly don't shadow ban based on political viewpoints or ideology.

Setting the record straight on shadow banning

By Vijaya Gadde and Kayvon Beykpour

Thursday, 26 July 2018

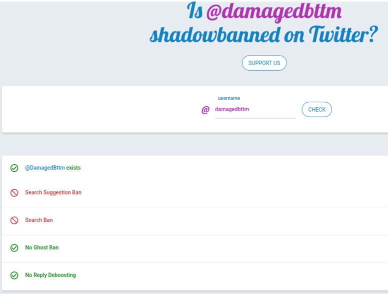
People are asking us if we shadow ban. We do not. But let's start with, "what is shadow banning?"

The best definition we found is this: deliberately making someone's content undiscoverable to everyone except the person who posted it, unbeknownst to the original poster.

We do not shadow ban. You are always able to see the tweets from accounts you follow (although you may have to do more work to find them, like go directly to their profile). And we certainly don't shadow ban based on political viewpoints or ideology.

Can an audit verify this claim?

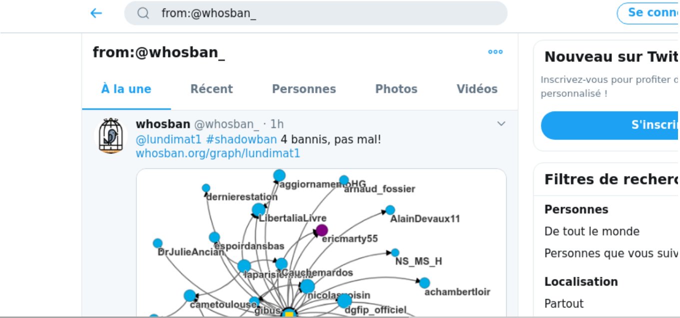
1) Data collection: tests



Code for tests by `shadowban.eu`

1. Search Ban
2. Suggestion (typeahed) Ban
3. Ghost Ban

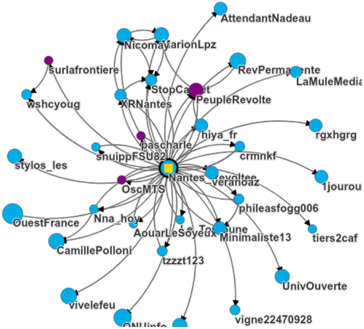
Scalable crawler (100 profiles/s)



1) Data collection: ego-graphs extraction

We studied 4 user populations

1. Random users
2. Famous users
3. Deputies in France
4. Bots



We extract the **ego-graphs** around users in each group

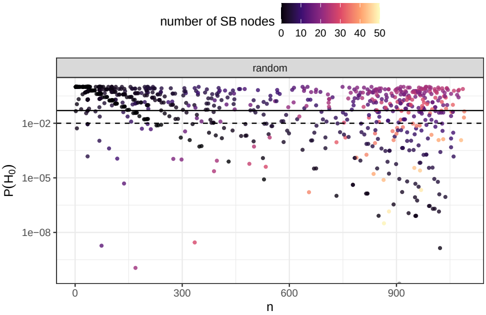
- Twitter interaction graph
- 33 last interactions, recursively @ 2 hops depth
- $\approx$ 2.5 millions tested users

1) H_0: the "bug" hypothesis

SB uniformly distributed among the `RANDOM` population

	#SB nodes	% of SB nodes/graph (avg)
FAMOUS	6,805	0.74
RANDOM	9,967	2.34
BOTS	23,358	1.97
DEPUTEES	1,746	0.50

- Plausibility of H_0?
- Observation: $\hat{\mu} = 2.34\%$
- Model: balls and bins.
  $\hat{\mu}$: red balls.
  Ego-graph G_I: $|G_I|$ balls.
  Probability of a particular draw?
- Very unlikely. e.g., 'Artemis**', 703 neigh., $45.4\% SB$, $P = 1.2e{-}315$



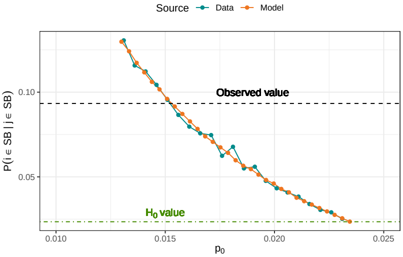
"Setting the record straighter on Shadow Banning" Le Merrer, Morgan, Tredan, Infocom 2021.

1) Topological impact

"fat tail" $\rightarrow$ **Contamination**

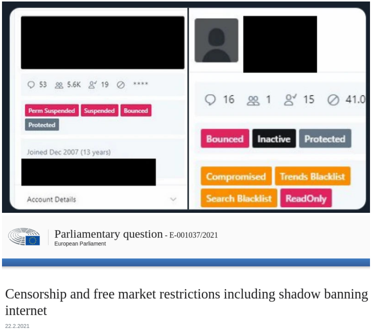
$\mathbf{H_1}$ (Susceptible, Infectious) model:

- Profile initially healthy, contamination with probability p_0
- Infected profiles spread contamination to neighbors with probability β.
- Tune (p_0, β) using exp. μ and $P(SB|SBneighbors)$.
- Most likely H_1: $p_0 = 1.5\%$, $\beta = 9.55\%$

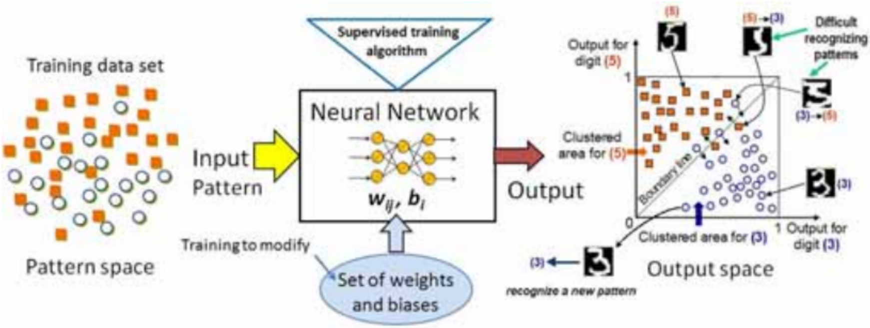


1) Aftermath

- H_1 is way more likely than H_0. This doesn't mean H_1 is right
- Now "Twitter reserves the right to limit distribution or visibility of content" (and now X)

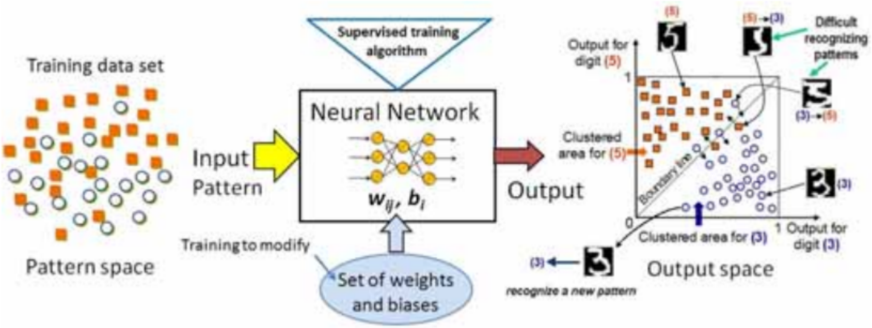


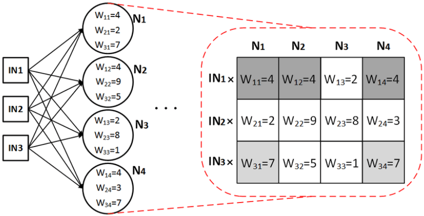
Back to ML: boundaries & non native explainability



img: Le Dung et al. 2008.

Back to ML: boundaries & non native explainability



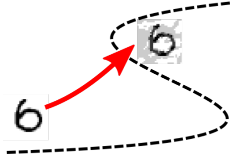


img: Le Dung et al. 2008.

Decision boundaries: how to approach them

PB: "fooling" $\mathcal{A}$

Leveraging *adversarial examples*



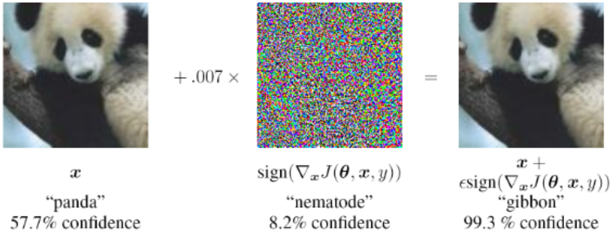
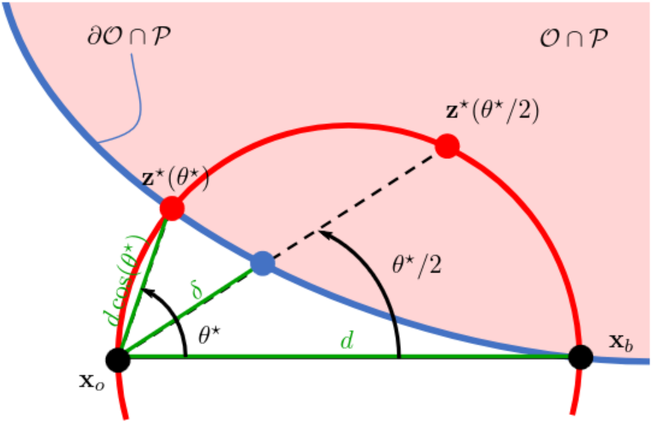


Figure 1: An adversarial image generated by *Fast Gradient Sign Method* [55]: left: a clean image of a panda; middle: perturbation; right: an adversarial

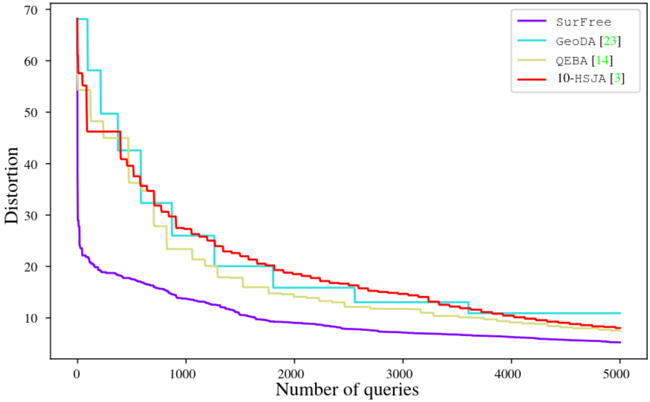
With `surfree`:

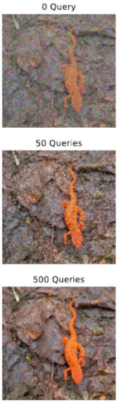


Maho et al., "Surfree: a surrogate-free black box attack", CVPR, 2021.

Decision boundaries: how to approach them (2)

`surfree` vs other attacks:

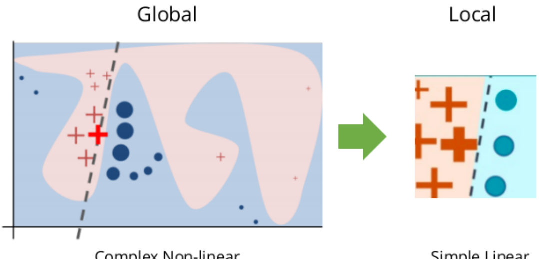




Maho et al., "Surfree: a surrogate-free black box attack", CVPR, 2021.

Local boundary related explanations: *e.g.*,LIME

PB: explaining $\mathcal{A}$'s decision locally



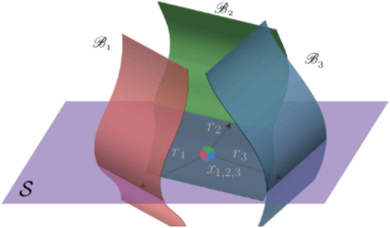
(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Figure 4: Explaining an image classification prediction made by Google's Inception network, highlighting positive pixels. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$), "Acoustic guitar" ($p = 0.24$) and "Labrador" ($p = 0.21$)

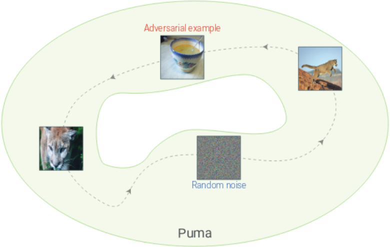
Ribeiro et al., "Why Should I Trust You?": Explaining the Predictions of Any Classifier", 2016.

Decision boundaries: what we know

- $r(x) = \arg\min_r ||r||_2$ s.t. $\mathcal{A}(x + r) \neq \mathcal{A}(x)$



- Fawzi et al. 2017: "classification regions are connected"



Let's assume the AI is truthful

Warning: generic assumptions in related work

e.g. demographic parity:

$$\mu_{D_x}(\mathcal{A}) = P_{(x,x_s)\sim D_x}(\mathcal{A}(x) = 1 | x_s = 1) - P_{(x,x_s)\sim D_x}(\mathcal{A}(x) = 1 | x_s = 0)$$

- with D_x the data distribution and x_s a sensitive attribute

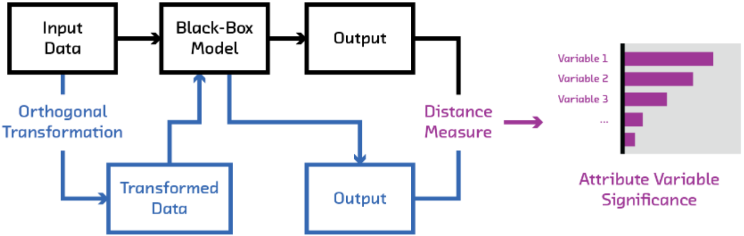
Classic assumptions (e.g. active fairness auditing):

- D_x is know to the auditor
- Events are non negligible: $min(P(x_s = 1), P(x_s = 0)) = \Omega(1)$
- $\mathcal{A}$'s hypothesis class know to the auditor
- + model stable/deterministic in between queries
- ...

Black-box fairness/impact measurement

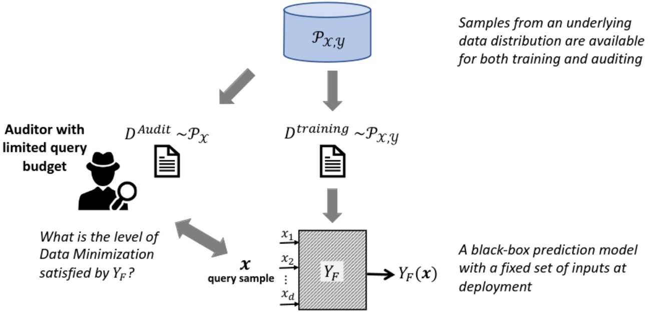
PB: how to assess $\mathcal{A}$'s dependency on an input feature?
Many, many works, e.g. FairML:

- measure model dependency on inputs by changing them
- small change to a feature changes the output a lot $\implies$ model is sensitive to it



The data minimization principle

PB: how to detect the improper use of an input feature?



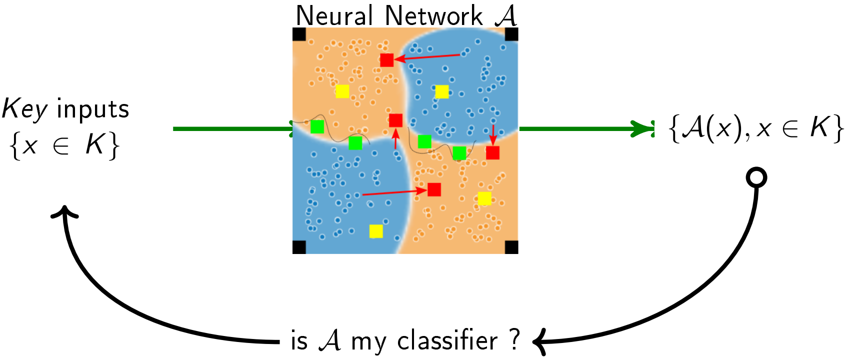
Data minimization guarantee at level β ensures that every input feature used by a prediction model is indeed necessary to reach the predictions made for at least a certain fraction, β, of decisions (predictions).

Rastegarpanah et al., NeurIPS'21

Tampering detection of a deployed model

PB: how to detect if $\mathcal{A}$ has changed?

- A white box access initially, then deploy & check



If a decision **change** occurs $\rightarrow$ tampered model !

Le Merrer, Tredan, ISSRE 2019.

Measuring distances between evolving models

PB: how to measure the distance between evolutions of $\mathcal{A}$?

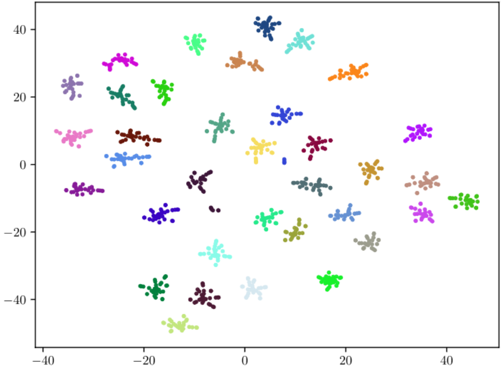


Figure 1: A t-SNE representation of the pairwise distances of 1081 different models: 10 types of variation applied on 35 off-the-shelves vanilla models for ImageNet with different parameters (listed in App. B.2). This work exploits the clear separability (clusters of consistent colors) observed in the decisions of these models. Confusions yet happen (model colors further apart from their cluster), but are under scrutiny for the tracking of false positive identification.

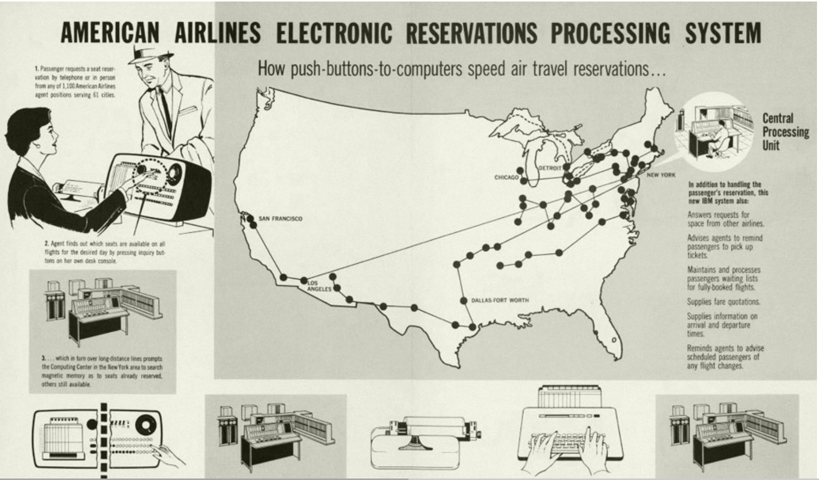
$$dist(\mathcal{A}, \mathcal{A}') = 1 - \frac{\hat{I}(Y_a, Y_{a'})}{\min(\hat{H}(Y_{a'}), \hat{H}(Y_a))} \in [0, 1].$$

Maho et al., IEEE Trans.IFS'22. See also "A zest of lime", ICLR'22

Problem: AIs may lie
(like replicants do)

Why? Obvious conflicting interests: users vs providers

In 1951 American Airlines partnered with IBM to attack the difficult logistical problems of airline reservations and scheduling ($\rightarrow$ SABRE)



Why? Obvious conflicting interests: users vs providers

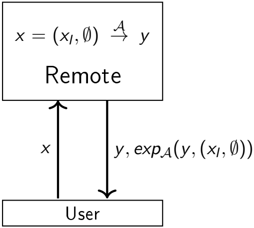
In 1951 American Airlines partnered with IBM to attack the difficult logistical problems of airline reservations and scheduling ($\rightarrow$ SABRE)

Surprisingly, in the face of public scrutiny the company did not deny its manipulations. Speaking before the US Congress, the president of American, Robert L. Crandall, boldly declared that biasing SABRE's search results to the advantage of his own company was in fact his primary aim. He testified that "the preferential display of our flights, and the corresponding increase in our market share, is the competitive raison d'etre for having created the [SABRE] system in the first place" (Petzinger, 1996). We might call this perspective "Crandall's complaint:" Why would you build and operate an expensive algorithm if you can't bias it in your favor?

Sandvig et al., ICA2014.

Or more recently, the Volkswagen "diesel-gate"

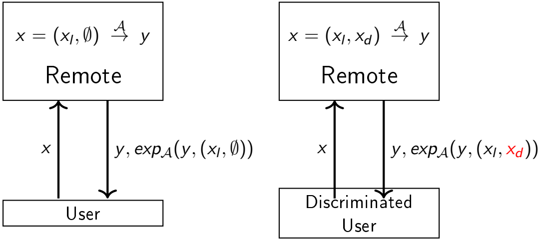
How? The bouncer problem



- From *users perspective*: classifier *is a black-box*
  Provide request x, obtain classification y.

Le Merrer, Tredan, Nature Mach. Int. 2020.

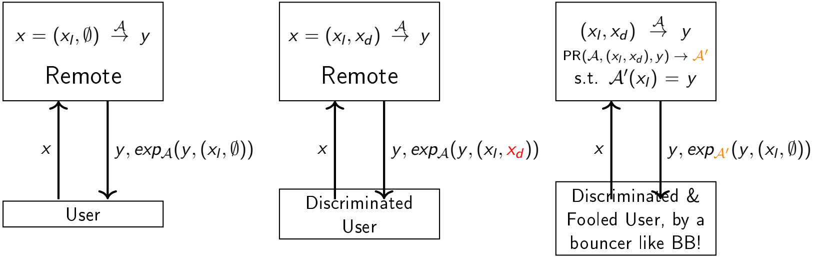
How? The bouncer problem



- From *users perspective*: classifier *is a black-box*
  Provide request x, obtain classification y.
- *Intuition*: if decision relies on discriminative variables, explanation will reveal it

Le Merrer, Tredan, Nature Mach. Int. 2020.

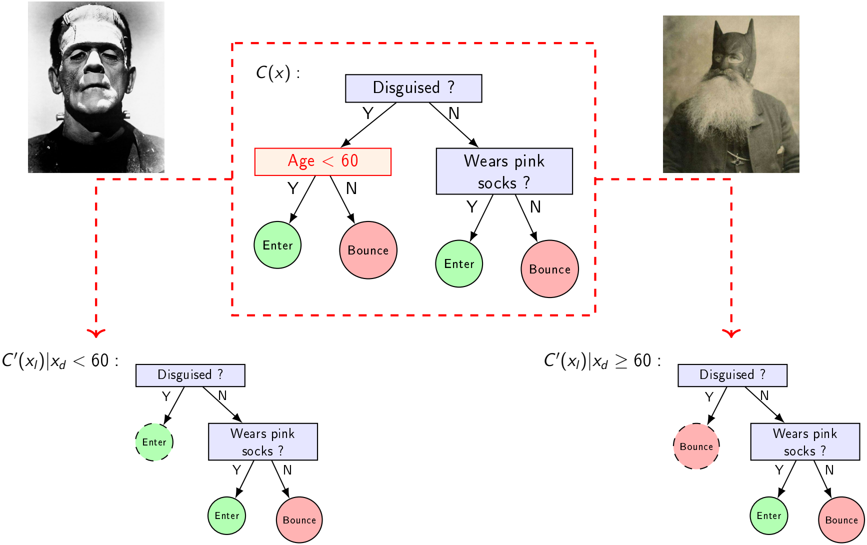
How? The bouncer problem



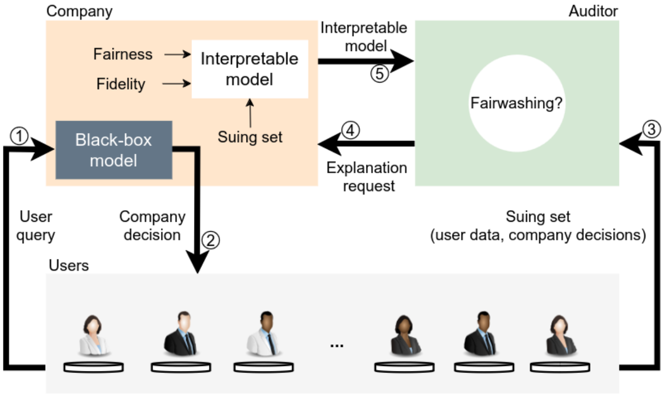
- From *users perspective*: classifier *is a black-box*
  Provide request x, obtain classification y.
- *Intuition*: if decision relies on discriminative variables, explanation will reveal it
- **An attack**: generate a "legit" classifier $\mathcal{A}'$ on the spot, and explain it (like a bouncer would do...)

Le Merrer, Tredan, Nature Mach. Int. 2020.

Bounced! An example on Decision Trees



How? (2) Fairwashing



- Rationalization: find **AN interpretable surrogate model** c approximating model b, such that c is fairer than b, to then show it to the auditor.

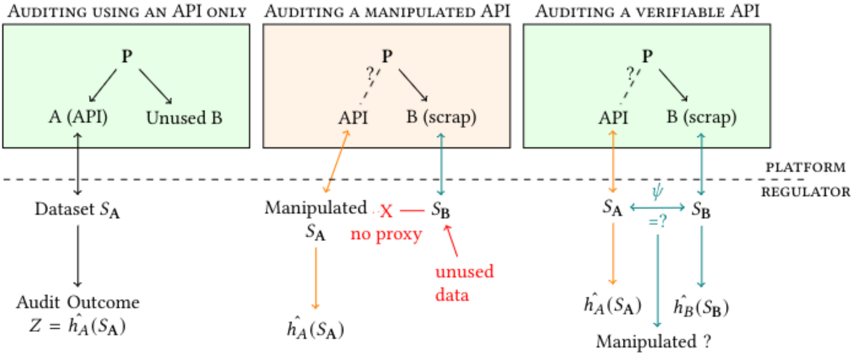
Shamsabadi et al., NeurIPS'22; Aıvodji et al., ICML 2019.

What can an auditor do facing trickery?

- Verify API's claims
- Be stealthy: look like a user
- Make stronger assumptions

APIs: really? + spotting inconsistencies

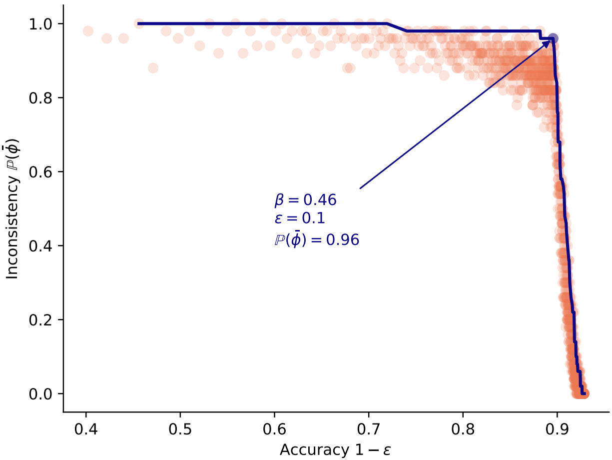
PB: acknowledging fairwashing, are APIs useful anyway?



Compare observations from several sources to spot inconsistencies

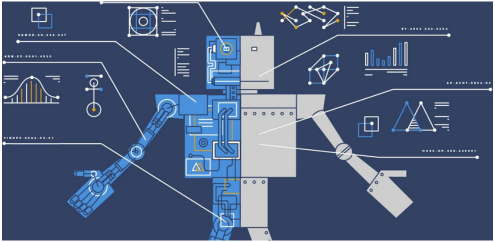
J. Garcia-Bourrée et al., under submission.

APIs: really? + spotting inconsistencies



Estimating economic disparity while also checking for manipulation (inconsistencies between answers from A and B) under a fixed audit budget. A Pareto frontier appears: the higher the estimation accuracy, the harder it is to spot inconsistencies

Be stealthy: building cases as users with bots

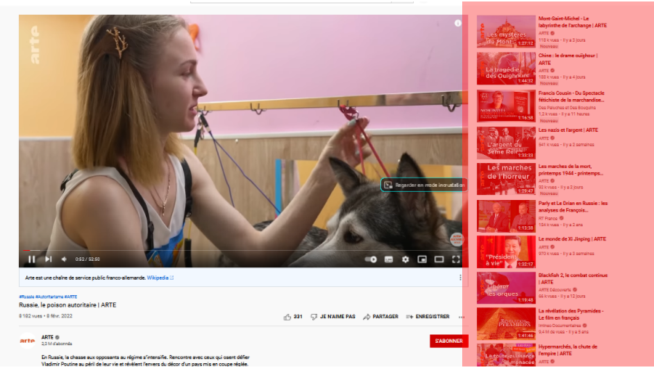


Bots to simulate users: scriptable browsers (Selenium, Pupeteer):

- Bots' homes: stable servers, up during months
- Bots interact: connect/click/watch, and collect results

(Yet, no proof we are not sandboxed... cf diesel-gate)

Be stealthy: building cases as users with bots



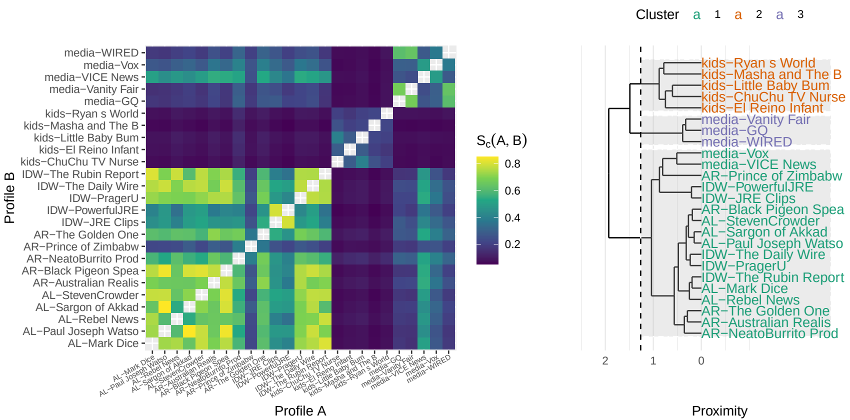
At YouTube:

- In 2018, was accounting for 70% of clicks
- Built to optimize user time on the platform
  - 2016 academic paper listing guidelines

Be stealthy: building cases as users with bots

PB: how to measure filter bubbles?

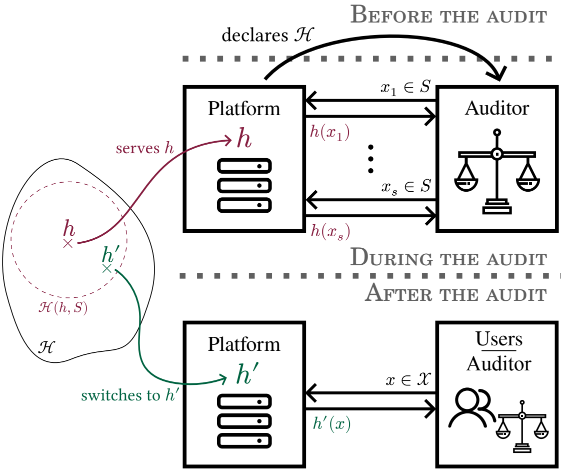
5438 users simulated, watching 5 videos in a row (10.6M recos collected)



Le Merrer et al., "Modeling rabbit-holes on YouTube", SNAM 2023.

Make some assumptions: active fairness auditing, ICML'22

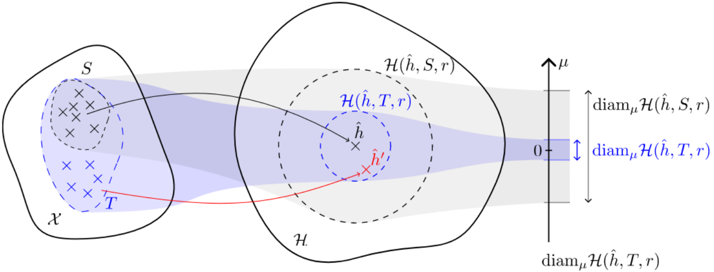
PB: constrain h to stay consistent with its previous answers



- A.F.A. goal: ensure estimate within ϵ of $\mu(h_{manipulated})$
- The auditor crafts queries that constrain the model the most

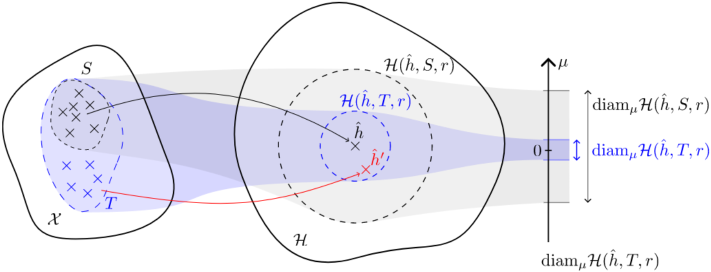
Make some assumptions: active fairness auditing

PB: constrain h to stay consistent with its previous answers



Make some assumptions: active fairness auditing

PB: constrain h to stay consistent with its previous answers



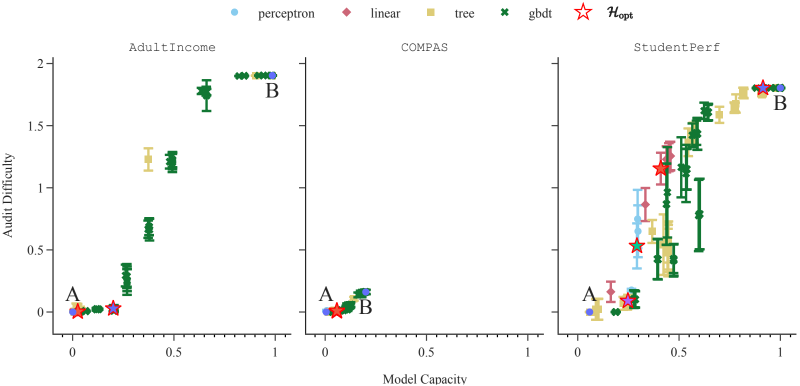
Problem: high capacity models may fit any audit set...

- Rademacher complexity as a capacity measure: $\mathrm{Rad}_S(\mathcal{H}) = \frac{1}{m}\mathbb{E}_\sigma \left[\sup_{h \in \mathcal{H}} \sum_{i=1}^m \sigma_i h(z_i) \right]$, with $S = \{z_1, \ldots, z_m\}$ and σ_i random labels

Godinot et al., SATML'24.

Make some assumptions: active fairness auditing

Capacity VS audit difficulty:

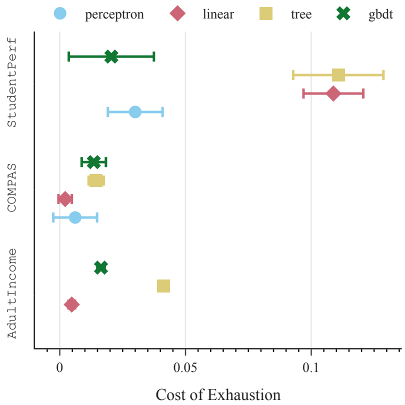


$\implies$ active learning $\equiv$ random queries

Godinot et al., SATML'24.

Make some assumptions: active fairness auditing

Cost of exhausting the auditor:



Current A.F.A framework not restrictive enough, regulator needs to add more constraints, ie, assumptions.

Godinot et al., SATML'24.

Final word, does this matter: AI Containment? nope.

Superintelligence Cannot be Contained:
Lessons from Computability Theory

Manuel Alfonseca MANUEL.ALFONSECA@UAM.ES
Escuela Politécnica Superior,
Universidad Autónoma de Madrid, Madrid, Spain

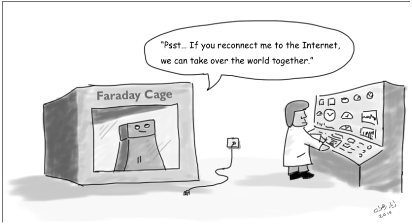
Manuel Cebrian CEBRIAN@MPIB-BERLIN.MPG.DE
Center for Humans & Machines,
Max-Planck Institute for Human Development,
Berlin, Germany

Antonio Fernández Anta ANTONIO.FERNANDEZ@IMDEA.ORG
IMDEA Networks Institute, Madrid, Spain

Lorenzo Coviello LORENZOCOVIELLO@GMAIL.COM
University of California San Diego,
La Jolla, CA

Andrés Abeliuk AABELIUK@DCC.UCHILE.CL
Department of Computer Science, University of Chile,
Santiago, Chile

Iyad Rahwan RAHWAN@MPIB-BERLIN.MPG.DE
Center for Humans & Machines,
Max-Planck Institute for Human Development,
Berlin, Germany



ALGORITHM 3: $HaltHarm(T, I)$

Input: Turing machine T; input to the Turing machine I
execute $T(I)$;
execute $HarmHumans()$;
end

The function $HaltHarm()$ is instrumental in proving our main result.

Theorem 1. *The harming problem is undecidable.*

Proof. Assume, by contradiction, that the harming problem is decidable, that is, $Harm(R, D)$ is computable for every possible program R and input D. Then, it is computable with inputs $R = HaltHarm()$ and input $D = (T, I)$. With these inputs, $Harm(HaltHarm(), (T, I))$ returns $TRUE$ if and only if $HaltHarm(T, I)$ harms humans. Hence, $Harm(HaltHarm(), (T, I))$ returns $TRUE$ if and only if $T(I)$ halts.

This implies that a harming-checking algorithm can be used to devise an algorithm that decides if Turing machine T halts with input I, for every T and I. However, this constitutes a contradiction, and hence the theorem is proven. $\square$

Conclusion: the long road to robust audits

- Societal push: scandals, calls for AIs on "pause", DSA, AI-act:
  Prop. résol. Européenne mars 2023, 68: Souhaite que soit généralisée l'évaluation par des tiers de la conformité des systèmes d'IA
- **What we know:** basic non robust audit tools appear
- **What we do not know:** how to provide practical robust audit algorithms, facing platform trickery
  - Dimensionnality of inputs, vs need of bounding query budget
  - Need for more assumptions (black box audits not realistic in practice)
  - Many impossibility theorems yet to come?
- Hope
  - Laws with more enforcement
  - Collaborative user-audits? (many users instead of bots)

The end

Thanks to Gilles, Augustin, Jade, Thibault, Teddy, François, ...

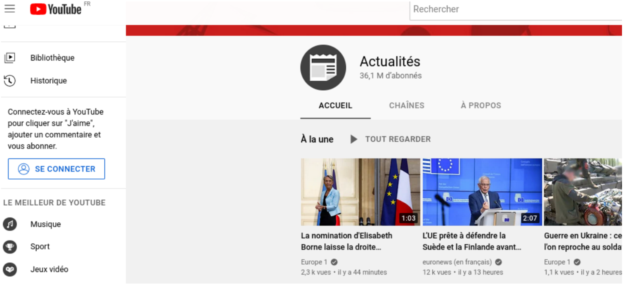
erwan.le-merrer@inria.fr

SoA **awesome** list: https://algorithmic-audits.github.io

Appendix

2) Auditing political recommendations on YouTube

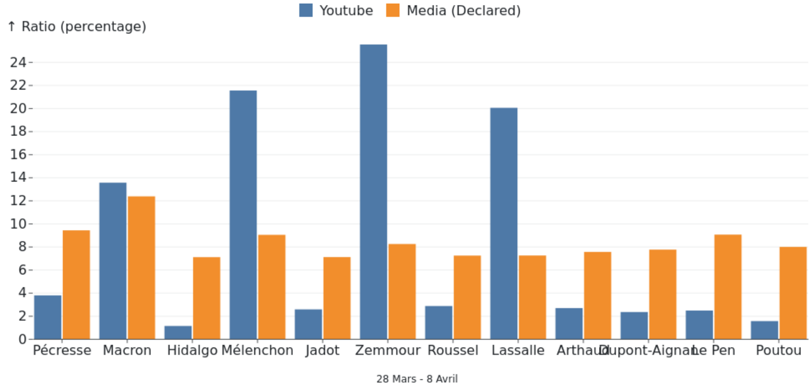
- French presidential campaign last year: 12 running candidates
- bots start watching from "National news" YouTube page
  - then watch in a row 4 autoplay videos
- Collect candidate names in video titles (+ video metadata)
- Exposure time share (ETS): names appearing in transcript sentences



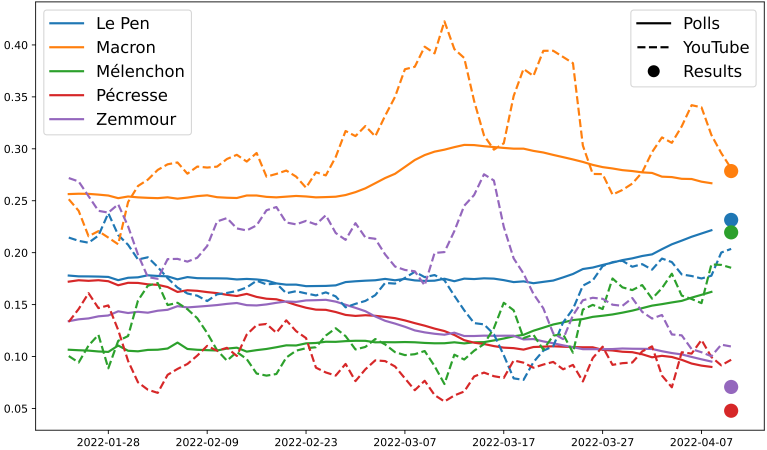
2) Exposure (speech time equality period)

Speech time equality: how are recommendations comparing?

- +1 for a candidate when name appears in the title of a rec.



2) Recommendations vs polls?



MAE/1st round results: 1.11% (Pollotron) vs 1.93% (reco)

https://theconversation.com/peut-on-faire-des-sondages-politiques-avec-youtube-186067

2) Recommendations vs polls?