La transition numérique démocratique est possible

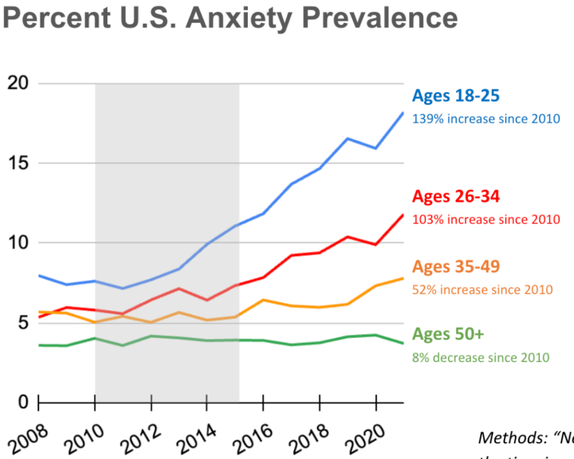
Lê Nguyên Hoang,
Calicarpa & Tournesol

Toute l'Intelligence Artificielle de Rennes, Septembre 2024

Section 1

Le contexte

Un contexte inquiétant



Gen Z hit hardest
Born after 1995

Young Millenials too

Methods: "Nervous all of the time or most of the time in past month"
SOURCE: U.S. National Survey on Drug Use and Health

Surtout à l'échelle sociétale

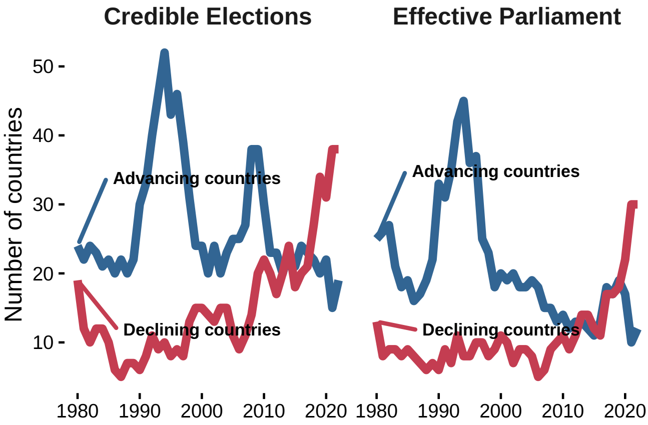


Image de l'article "Global Patterns" du *Global State of Democracy Initiative*.

Une tendance terrifiante

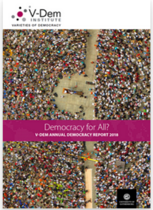
Previous Reports

DR 2022: Autocratization Changing Nature?

DR 2021: Autocratization Turns Viral

DR 2020: Autocratization Surges - Resistance Grows

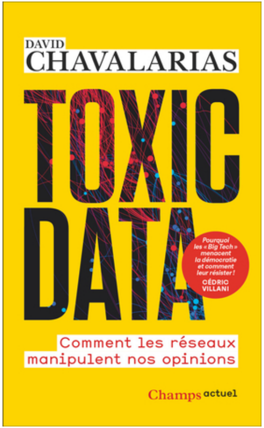
DR 2019: Democracy Facing Global Challenges

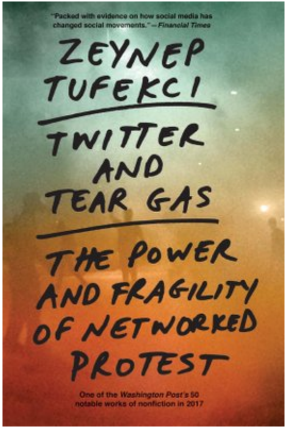


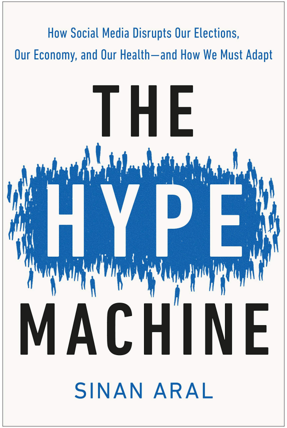
DR 2018: Democracy for All?

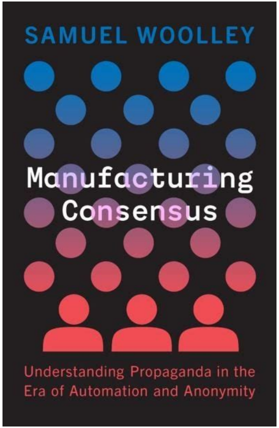
DR 2017: Democracy at Dusk?

Les IA au coeur du problème ?

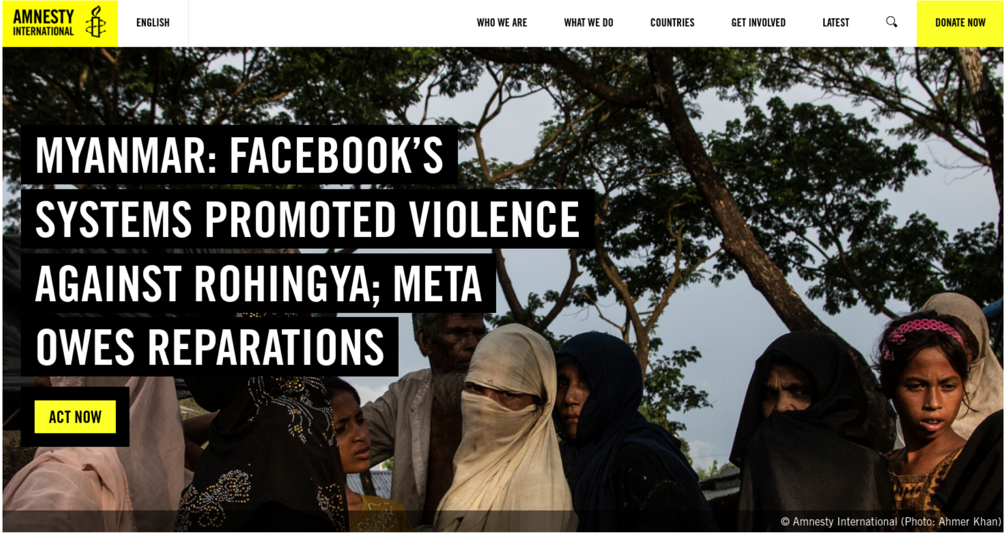








Les IA déjà coupables ?



Plus lucratives et dangereuses que ChatGPT

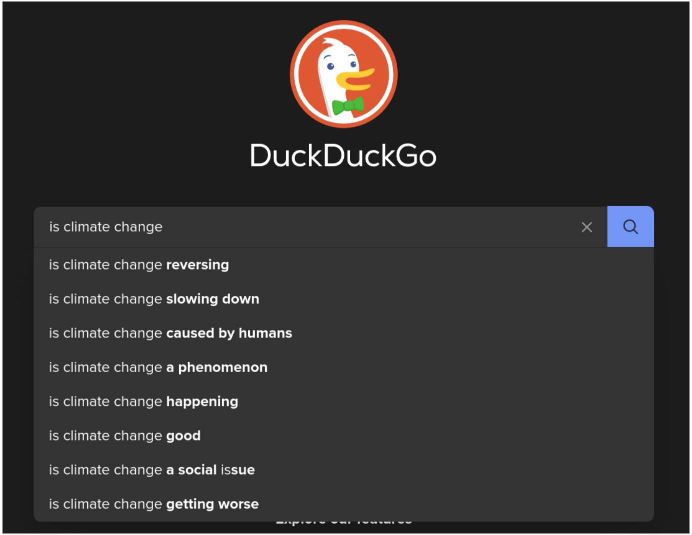
Pixabay image by LolaSandoval1.

Interdites dans leurs propres pays de production

Section 2

Les implications pour la recherche en IA

Des millions de milliards de dilemmes...



… actuellement résolus par un jeu !

Image Pixabay par HtcHnm.

Les industries de la désinformation l'ont très bien compris...

"TEAM JORGE": IN THE HEART OF A GLOBAL DISINFORMATION MACHINE

In Part 2 of the "Story Killers" project, which continues the work of assassinated Indian journalist Gauri Lankesh on disinformation, the Forbidden Stories consortium investigated an ultra-secret Israeli company involved in manipulating elections and hacking African politicians. We took an unprecedented dive into a world where troll armies, cyber espionage and influencers are intertwined.

MINISTÈRE DES ARMÉES
Liberté Égalité Fraternité
IRSEM
INSTITUT DE RECHERCHE STRATÉGIQUE DE L'ÉCOLE MILITAIRE

Accueil - IRSEM > CHINESE INFLUENCE OPERATIONS

Download the report

(PDF file)
English translation of the October 2021 edition
654 pages

Combien de faux comptes sont retirés par Facebook chaque année ?

Réponse

Facebook Removed More than 15 Billion Fake Accounts in Two Years, Five Times more than its Active User Base

Jastra Kranjec · Pro Investor
Updated: 27 September 2021

Disclosure

As the world's largest social networking platform, Facebook has witnessed a surge in the number of users in the past few years. Hundreds of millions of people have joined its social media space to communicate, keep in touch with the latest trends or promote business, especially after the pandemic hit. Although the COVID-19 restrictions have loosened in most countries, Facebook's active user base continues growing, but so does the number of fake accounts.

According to data presented by Stock Apps, the social media giant removed over 15 billion fake accounts in the last two years, five times more than its active user base.

3 Billion Fake Accounts Removed in the First Half of 2021, 20x More than the Number of New Active Users

Scammers use fake Facebook accounts to connect with users, get their personal information and steal identities. Most of them will reach out to anyone who's accepted their friend request to try and scam them out of money.

Section 3

La sécurité des IA

Les hypothèses dangereuses de la recherche

L'hypothèse irréaliste la plus commune

Soit $x_1, x_2, \ldots x_n$ des données indépendantes et identiquement distribuées...

Les hypothèses dangereuses de la recherche

L'hypothèse irréaliste la plus commune

Soit $x_1, x_2, \ldots x_n$ des données indépendantes et identiquement distribuées...

L'hypothèse extrêmement politisée devenue banalisée

Nous apprenons une fonction f qui généralise les données...

L'empoisonnement des données d'apprentissage est la plus grande menace

Adversarial Machine Learning - Industry Perspectives

Ram Shankar Siva Kumar[*], Magnus Nyström[†], John Lambert[‡], Andrew Marshall[§], Mario Goertzel[¶]
, Andi Comissoneru[‖], Matt Swann[**] and Sharon Xia[††]
Microsoft
Redmond,USA
Email: [*]Ram.Shankar@microsoft.com, [†]mnystrom@microsoft.com, [‡]johnla@microsoft.com, [§]amarshal@microsoft.com
[¶]mariogo@microsoft.com, [‖]andic@microsoft.com, [**]mswann@microsoft.com,
[††]shxia@microsoft.com

***Abstract*—Based on interviews with 28 organizations, we found that industry practitioners are not equipped with tactical and strategic tools to protect, detect and respond to attacks on their Machine Learning (ML) systems. We leverage the insights from the interviews and enumerate the gaps in securing machine learning systems when viewed in the context of traditional software security development. We write this paper from the perspective of two personas: developers/ML engineers and security incident responders. The goal of this paper is to layout the research agenda to amend the Security Development Lifecycle for industrial-grade software in the adversarial ML era.**

***Index Terms*—adversarial machine learning, software security, engineering**

I. Introduction

Adversarial Machine Learning is now having a moment in the software industry - For instance, Google [1], Microsoft [2] and IBM [3] have signaled, separate from their commitment to securing their traditional software systems, initiatives to secure ML systems. In Feb 2019, Gartner, the leading industry market research firm, published its first report on adversarial machine learning [4] advising that "*Application leaders must anticipate and prepare to mitigate potential risks of data corruption, model theft, and adversarial samples.*" The motivation for this paper is to understand the extent to which organizations across

investments in 2020 [12]) and it is only natural that organizations invest in protecting their "crown jewels".

We make two contributions in this paper:

1) Despite the compelling reasons to secure ML systems, over a survey spanning 28 different organizations, we found that most industry practitioners are yet to come to terms with adversarial machine learning. 25 out of the 28 organizations indicated that they don't have the right tools in place to secure their ML systems and are explicitly looking for guidance.
2) We enumerate the security engineering aspects of building ML systems using Security development Lifecycle (SDL) frame work, the de facto software building process in industry.

This paper is a compendium of pain points and gaps in securing machine learning systems as encountered by typical software organizations. We hope to appeal to the research community to help solve the problem faced by two personas - software developers/ML engineers and security incident responders - when securing machine learning systems. The goal of this paper is to engage ML researchers to revise and amend Security Development Lifecycle for industrial-grade software in the adversarial ML era.

TABLE V
Top Attack

Which attack would affect your org the most?	***Distribution***
Poisoning (e.g: [21])	10
Model Stealing (e.g: [22])	6
Model Inversion (e.g: [23])	4
Backdoored ML (e.g: [24])	4
Membership Inference (e.g: [25])	3
Adversarial Examples (e.g: [26])	2
Reprogramming ML System (e.g: [27])	0
Adversarial Example in Physical Domain (e.g: [5])	0
Malicious ML provider recovering training data (e.g: [28])	0
Attacking the ML supply chain (e.g: [24])	0
Exploit Software Dependencies (e.g: [29])	0

Les agences gouvernementales s'attaquent (enfin!) au sujet

Information Technology Laboratory

COMPUTER SECURITY RESOURCE CENTER

COMPUTER SECURITY RESOURCE CENTER CSRC

PUBLICATIONS

NIST AI 100-2 E2023

Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations

Date Published: January 2024

Author(s)

Apostol Vassilev (NIST), Alina Oprea (Northeastern University), Alie Fordyce (Robust Intelligence), Hyrum Anderson (Robust Intelligence)

Abstract

This NIST Trustworthy and Responsible AI report develops a taxonomy of concepts and defines terminology in the field of adversarial machine learning (AML). The taxonomy is built on surveying the AML literature and is arranged in a conceptual hierarchy that includes key types of ML methods and lifecycle stages of attack, attacker goals and objectives, and attacker capabilities and knowledge of the learning process. The report also provides corresponding methods for mitigating and managing the consequences of attacks and points out relevant open challenges to take into account in the lifecycle of AI systems. The terminology used in the report is consistent with the literature on AML and is complemented by a glossary that defines key terms associated with the security of AI systems and is intended to assist non-expert readers. Taken together, the taxonomy and terminology are meant to inform other standards and future practice guides for assessing and managing the security of AI systems, by establishing a common language and understanding of the rapidly developing AML landscape.

Keywords

artificial intelligence; machine learning; attack taxonomy; evasion; data poisoning; privacy breach; attack mitigation; data modality; chatbot; generative models; large language model; trojan attack; backdoor attack

Control Families

DOCUMENTATION

Publication:

https://doi.org/10.6028/NIST.AI.100-2e2023

Download URL

Supplemental Material:

Trustworthy & Responsible AI Resource Center

NIST news article

Document History:

10/30/19: IR 8269 (Draft)

03/08/23: AI 100-2 E2023 (Draft)

01/04/24: AI 100-2 E2023 (Final)

TOPICS

Security and Privacy

advanced persistent threats, botnets, information sharing, intrusion detection & prevention, malware

Technologies

artificial intelligence

Une préconisation fondamentale : réduire le nombre de paramètres

On the Impossible Safety of Large AI Models

El-Mahdi El-Mhamdi[1,2], Sadegh Farhadkhani[3], Rachid Guerraoui[3], Nirupam Gupta[3], Lê-Nguyên Hoang[2,4], Rafaël Pinot[3], Sébastien Rouault[2], and John Stephan[3]

[1]École Polytechnique
[2]Calicarpa
[3]EPFL
[4]Tournesol Association

Abstract

Large AI Models (LAIMs), of which large language models are the most prominent recent example, showcase some impressive performance. However they have been empirically found to pose serious security issues. This paper systematizes our knowledge about the *fundamental impossibility* of building arbitrarily accurate and secure machine learning models. More precisely, we identify key challenging features of many of today's machine learning settings. Namely, high accuracy seems to require *memorizing* large training datasets, which are often *user-generated* and *highly heterogeneous*, with both *sensitive information* and *fake users*. We then survey statistical lower bounds that, we argue, constitute a compelling case against the possibility of designing high-accuracy LAIMs with strong security guarantees.

1 Introduction

In recent years, we have witnessed a race for developing larger and larger artificial intelligence (AI) models. Notable milestones in this trend are *Attention Networks* (213 million parameters) [VSP+17], *GPT-2* (1.5 billion parameters) [RWC+19], *GPT-3* (175 billion parameters) [BMR+20], *Switch Transformer* (1.6 trillion parameters) [FZS21], *Persia* (over 100 trillion parameters) [LYZ+21], and *GPT-4* (unknown number of parameters) [BCE+23]. The scaling of model sizes has shown improvement in the accuracies on classical tasks, such as GLUE [WSM+19], SuperGLUE [WPN+19] and Winograd [SBBC20], without significant diminishing returns so far (see, e.g., Figure 1 in [BMR+20]). Moreover large AI models (or LAIMs) can also be used as *few-shot learners* [BMR+20], which has motivated their wide use as pre-trained *base* (or *foundation*) models [CCM21, CLL21, JLZ22, VPKG21, ZWK+21]. This success has generated enormous academic, economic and political interests into the development and deployment of LAIMs in public domain applications including content moderation, recommendation, search and ad targeting [Dea21, Hei21].

Contrary to the conventional wisdom of probably approximately correct (PAC) learning [Val84], the performance of LAIMs has been empirically shown to be best achieved by fully *interpolating* the training data [BHMM19, NKB+20, ZBH+17]. Put differently, the best accuracy is reached when these models *memorize* their training data [Fel20]. This phenomenon has also been theoretically supported to a certain extent by a recent line of work [BHX20, BMM18, BRT19, JSS+20, HY21, Hol21, LLS21, MM19, MVSS20, NVKM21]. Furthermore, training LAIMs requires access

The Poison of Dimensionality

Anonymous Authors[1]

Abstract

This paper advances the understanding of how the size of a machine learning model affects its vulnerability to poisoning, despite the use of state-of-the-art defenses. Given isotropic random honest feature vectors and the geometric median as the robust gradient aggregator rule, we essentially prove that, perhaps surprisingly, linear and logistic regressions with $D \geq 169H^2/P^2$ parameters are subject to *arbitrary model manipulation* by poisoners, where H and P are the numbers of honestly labeled and poisoned data points used for training. Our experiments go on exposing a fundamental tradeoff between augmenting model expressivity and increasing the poisoners' *attack surface*. We also informally discuss potential implications for "sandboxed learning", neural networks and non-zero-sum targeted poisoning.

1. Introduction

The classical theory of learning (Valiant, 1984; Geman et al., 1992; Kohavi & Wolpert, 1996) suggests that, given N training data, learning models should have $D = \Theta(N)$ parameters. But a vast empirical and theoretical literature on the *double descent* phenomenon (Zhang et al., 2017; Belkin et al., 2019; Muthukumar et al., 2019; Nakkiran et al., 2020; Mei & Montanari, 2022; Hastie et al., 2022) instead suggests that better performance could be obtained by letting $D \to \infty$. In any case, massive data collection has led to ever larger learning models (Brown et al., 2020; Fedus et al., 2022; Lian et al., 2022; Chowdhery et al., 2023).

However, these theories arguing for $D \geq \Omega(N)$ all assume that all training data are "honest" and should be generalized. In large-scale high-risk applications like language processing and content recommendation, this is deeply *unrealistic* and *ethically questionable* (Kallus & Zhou, 2018; Bender et al., 2021), if not illegal (Sag, 2023; Samuelson, 2023).

After all, many of these systems fit massive web-crawled datasets (Smith et al., 2013; Chowdhery et al., 2023), which are heavily *poisoned* by doxed personal data, hate speech and state-sponsored propaganda (Woolley, 2023; Yurieff, 2019; Andrzejewski, 2023). In fact, such *data poisoning*, i.e. injections of misleading inputs in training datasets (Biggio et al., 2012; Suya et al., 2021), has become the leading AI security concern in the industry (Kumar et al., 2020).

Meanwhile, a growing line of research has been suggesting that high-dimensional training facilitates persistent poisoning attacks (Hubinger et al., 2024), even given state-of-the-art defenses (El-Mhamdi et al., 2022). The theoretical case has mostly relied on an mathematical impossibility to bring the norm of the gradient at termination below $\Omega(\sqrt{D})$. However, it is unclear that the performance of the poisoned model is then worse than if trained with fewer parameters.

Our paper advances the understanding of how model size D affects machine learning security, given H honestly labeled data and P poisoned data. Crucially, for $P = \Theta(H)$ (e.g. 1% of poisoned data), our results completely diverge from the common wisdom $D \geq \Omega(N) = \Omega(H)$. More precisely, we make the following contributions.

Contributions. First, when $D \geq 169H^2/P^2$, we essentially prove that using a state-of-the-art poisoning defense (gradient descent with the geometric median) actually provides *zero* resilience guarantee, even for the two most standard learning problems (linear and logistic regression). In fact, we prove *arbitrary model manipulation* by poisoners.

Second, we empirically show the value of *dimension reduction* under poisoning, for two other state-of-the-art poisoning defenses. Our experiments highlight a tradeoff between model expressivity and restricted *attack surface*.

Third, we prove and leverage a property of random vector subspaces to informally discuss the applicability of our analysis to "sandboxed learning" and nonlinear models.

Le poison de la dimensionalité

Considérons H données $(x_1, y_1), \dots, (x_H, y_H)$ honnêtes, avec $x_h \in \mathbb{R}^D$ et $y_h \in \mathbb{R}$.

Le poison de la dimensionalité

Considérons H données $(x_1, y_1), \ldots, (x_H, y_H)$ honnêtes, avec $x_h \in \mathbb{R}^D$ et $y_h \in \mathbb{R}$.
Supposons les x_h isotropes, e.g. $x_h \sim \mathcal{N}(0, I_D)$, et les étiquettes correctes $y_h \sim \mathcal{N}(\beta^T x_h, \sigma^2)$.

Le poison de la dimensionalité

Considérons H données $(x_1, y_1), \ldots, (x_H, y_H)$ honnêtes, avec $x_h \in \mathbb{R}^D$ et $y_h \in \mathbb{R}$.
Supposons les x_h isotropes, e.g. $x_h \sim \mathcal{N}(0, I_D)$, et les étiquettes correctes $y_h \sim \mathcal{N}(\beta^T x_h, \sigma^2)$.
Utilisons la descente de gradient avec robustification par médiane géométrique (un algorithme d'apprentissage appartenant à l'état de l'art de l'IA sécurisée).

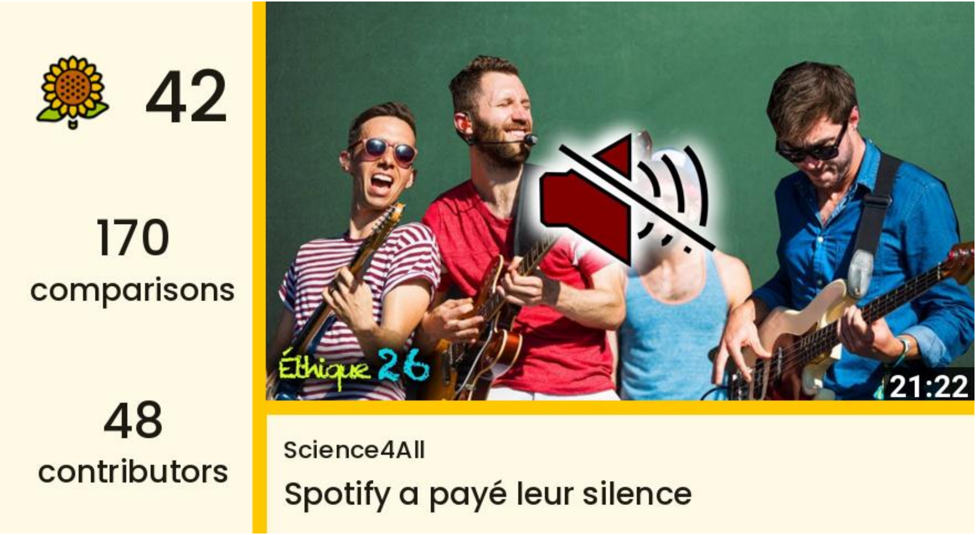
Le poison de la dimensionalité

Considérons H données $(x_1, y_1), \ldots, (x_H, y_H)$ honnêtes, avec $x_h \in \mathbb{R}^D$ et $y_h \in \mathbb{R}$.
Supposons les x_h isotropes, e.g. $x_h \sim \mathcal{N}(0, I_D)$, et les étiquettes correctes $y_h \sim \mathcal{N}(\beta^T x_h, \sigma^2)$.
Utilisons la descente de gradient avec robustification par médiane géométrique (un algorithme d'apprentissage appartenant à l'état de l'art de l'IA sécurisée).

Theorem (Hoang 2024, version informelle)

Supposons $D \geq 169H^2/P^2$. Alors, avec grande probabilité, il existe P données empoisonnées qui permettent une **manipulation arbitraire du modèle***.*

Autre préconisation : l'apprentissage sandboxé



An Equivalence Between Data Poisoning and Byzantine Gradient Attacks

Sadegh Farhadkhani, Rachid Guerraoui, Lê Nguyên Hoang, Oscar Villemaud *Proceedings of the 39th International Conference on Machine Learning*, PMLR 162:6284-6323, 2022.

$$\textsc{Loss}(\rho, \vec{\theta}, \vec{\mathcal{D}}) \triangleq \sum_{n \in [N]} \mathcal{L}_n(\theta_n, \mathcal{D}_n) + \sum_{n \in [N]} \mathcal{R}(\rho, \theta_n).$$

Section 4

L'état de l'art en stratégies d'atténuation

Ne laissez pas les données sortir de chez vous

- Modèles open sources locaux.

Ne laissez pas les données sortir de chez vous

- Modèles open sources locaux.
- Sandboxing des algorithmes auto-apprenants.

Ne laissez pas les données sortir de chez vous

- Modèles open sources locaux.
- Sandboxing des algorithmes auto-apprenants.
- Confidentialité différentielle.

Ne laissez pas les données sortir de chez vous

- Modèles open sources locaux.
- Sandboxing des algorithmes auto-apprenants.
- Confidentialité différentielle.
- Calcul multi-partite.

Évasion à la détection

- Apprentissage adversarial.

Évasion à la détection

- Apprentissage adversarial.
- Détection de données hors-distribution.

Évasion à la détection

- Apprentissage adversarial.
- Détection de données hors-distribution.
- Analyse de glissement distributionnel.

Évasion à la détection

- Apprentissage adversarial.
- Détection de données hors-distribution.
- Analyse de glissement distributionnel.
- Recours pour les faux négatifs/positifs.

Empoisonnement des modèles

- Nettoyage des données.

Empoisonnement des modèles

- Nettoyage des données.
- Authentification (cryptographique) des sources.

Empoisonnement des modèles

- Nettoyage des données.
- Authentification (cryptographique) des sources.
- Apprentissage par aggrégations résilientes.

Empoisonnement des modèles

- Nettoyage des données.
- Authentification (cryptographique) des sources.
- Apprentissage par aggrégations résilientes.
- Réduction de la dimensionalité.

- Réduire la surface d'attaque.

Fondamentaux de la cybersécurité

- Réduire la surface d'attaque.
- Cloisonnement des composants avec moindre privilège.

- Réduire la surface d’attaque.
- Cloisonnement des composants avec moindre privilège.
- Redondance et diversification des systèmes critiques.

Fondamentaux de la cybersécurité

- Réduire la surface d'attaque.
- Cloisonnement des composants avec moindre privilège.
- Redondance et diversification des systèmes critiques.
- Monitoring du système d'information.

Gouvernance via la théorie du choix social

- Équité du vote (une personne, une “voix” ?).

Gouvernance via la théorie du choix social

- Équité du vote (une personne, une “voix” ?).
- Incitatifs des votants et des candidats.

Gouvernance via la théorie du choix social

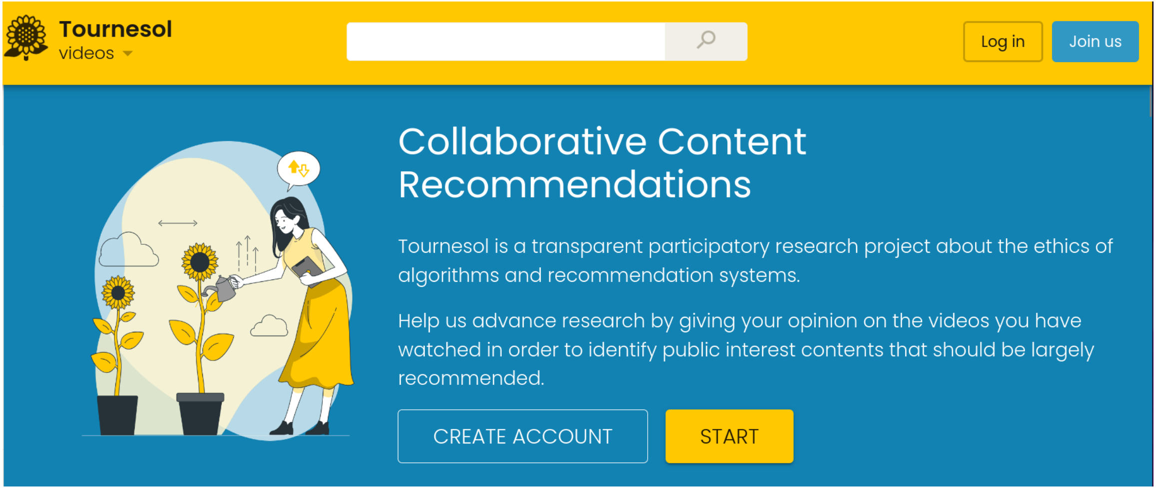
- Équité du vote (une personne, une “voix” ?).
- Incitatifs des votants et des candidats.
- Amplifier la volition à l’expertise.

Gouvernance via la théorie du choix social

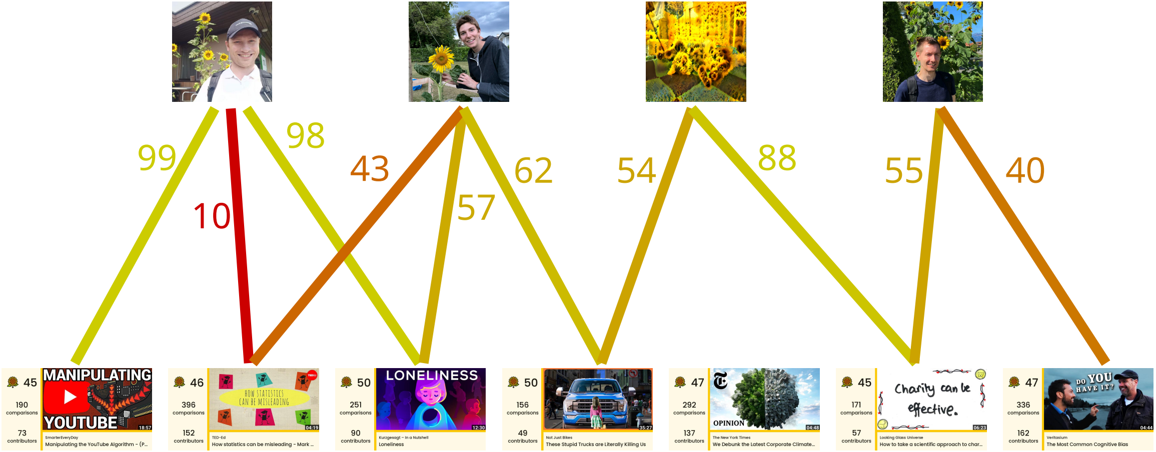
- Équité du vote (une personne, une “voix” ?).
- Incitatifs des votants et des candidats.
- Amplifier la volition à l’expertise.
- Prioriser le consensus radical au sujets clivants ?

Section 5

Le problème du scrutin creux et robuste


Tournesol
videos
Log in
Join us
Collaborative Content Recommendations
Tournesol is a transparent participatory research project about the ethics of algorithms and recommendation systems.
Help us advance research by giving your opinion on the videos you have watched in order to identify public interest contents that should be largely recommended.
CREATE ACCOUNT
START

Scrutin creux (sparse voting)



Le problème des juges français

Note: Certains de mes meilleurs amis sont parisiens et marseillais.

Le problème des juges français

Note: Certains de mes meilleurs amis sont parisiens et marseillais.

Le problème du juge parisien

Certains contenus sont disproportionnellement plus jugés par des juges qui souffrent d'une addiction à se plaindre.

Le problème des juges français

Note: Certains de mes meilleurs amis sont parisiens et marseillais.

Le problème du juge parisien

Certains contenus sont disproportionnellement plus jugés par des juges qui souffrent d'une addiction à se plaindre.

Le problème du juge marseillais

Certains contenus sont disproportionnellement plus jugés par des juges qui souffrent d'une exagération compulsive.

Une formalisation des différences de style d'expression

Fonctions d'utilité Von Neumann - Morgenstern

Les préférences cardinales sont définies à une transformation affine positive près.

Une formalisation des différences de style d'expression

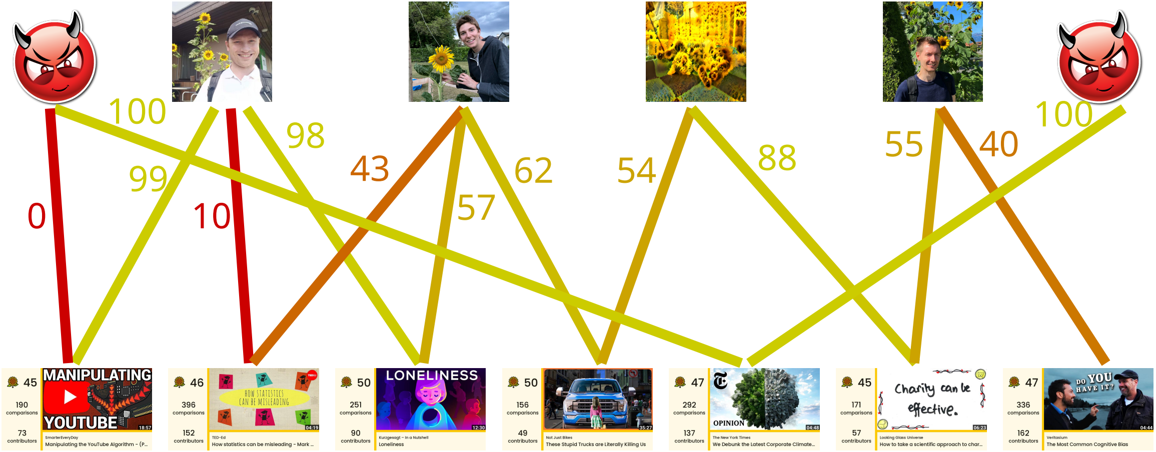
Fonctions d'utilité Von Neumann - Morgenstern

Les préférences cardinales sont définies à une transformation affine positive près.

Unanimité creuse (version informelle)

Si tous les électeurs ont une même préférence (à transformation affine positive près), et si chaque contenu est suffisamment évaluée, alors le scrutin doit retourner cette préférence consensuelle.

Robustesse du scrutin



Notre garantie de sécurité : la résilience Lipschitz

Résilience Lipschitz (informelle)

Un scrutin est Lipschitz-resilient avec une constante L, si les votes d'un contributor affectent les scores d'au plus une quantité L.

Notre garantie de sécurité : la résilience Lipschitz

Résilience Lipschitz (informelle)

Un scrutin est Lipschitz-resilient avec une constante L, si les votes d'un contributor affectent les scores d'au plus une quantité L.

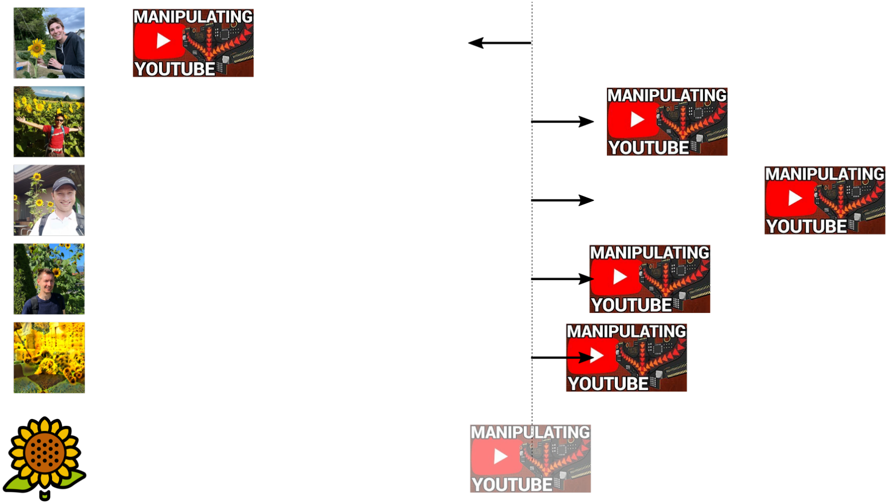
Generalization to the case of continuous voting rights

Un scrutin est Lipschitz-resilient avec une constante L ssi le scrutin est L-lipschitzienne en ses droits de vote (en considérant la norme ℓ_1 pour le vecteur des droits de vote, et la norme ℓ_∞ pour les scores).

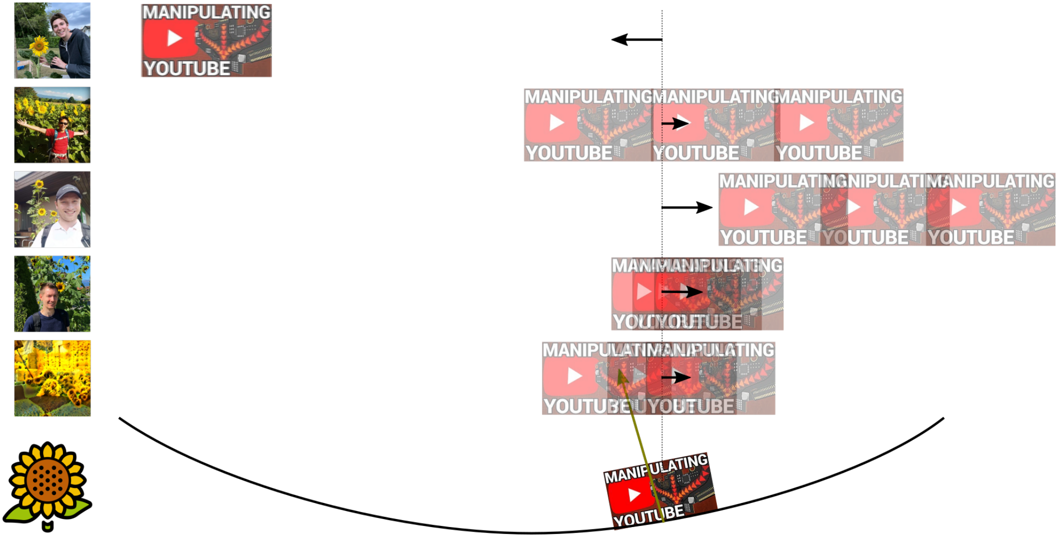
Question

Y a-t-il un algorithme qui satisfait
l'unanimité creuse et la résilience Lipschitz ?

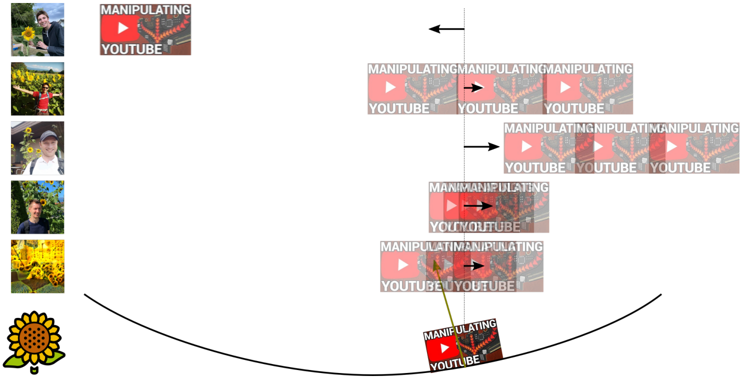
La primitive clé: la médiane quadratiquement régularisée



La primitive clé: la médiane quadratiquement régularisée



La primitive clé: la médiane quadratiquement régularisée



Theorem

$QrMed_L(\mathbf{x}) \triangleq \arg\min_{z \in \mathbb{R}} \left\{ \frac{1}{2L} z^2 + \sum_{i=1}^{n} |x_i - z| \right\}$ *is L-Lipschitz resilient.*

Notre solution : Mehestan

Definition (informelle)

1. Min-max-normaliser le vecteur de chaque électeur.
2. Pour chaque pair d'électeurs, comparer leurs scalings relatifs.
3. Pour chaque électeur i, agréger les scalings relatifs comparés aux autres, en utilisant la primitive $LrMean$, pour obtenir le scaling de l'électeur I.
4. Aggregate rescaled scores with $QrMed$.

Notre solution : Mehestan

Definition (informelle)

1. Min-max-normaliser le vecteur de chaque électeur.
2. Pour chaque pair d'électeurs, comparer leurs scalings relatifs.
3. Pour chaque électeur i, agréger les scalings relatifs comparés aux autres, en utilisant la primitive *LrMean*, pour obtenir le scaling de l'électeur I.
4. Aggregate rescaled scores with *QrMed*.

Theorem (Allouah, Guerraoui, H, Villemaud (AISTATS'24))

Mehestan is sparsely unanimous and Lipschitz resilience.

Section 6

Solidago : les fondements de la gouvernance algorithmique

Solidago: A Modular Pipeline for Collaborative Scoring

Anonymous Authors[1]

Abstract

This paper presents SOLIDAGO, an end-to-end modular pipeline to allow any number of users to collaboratively score any number of entities. SOLIDAGO decomposes the problem in six modules. First, we use pretrust and peer-to-peer vouches to assign trust scores to users, with a novel secure trust propagation algorithm. Second, based on user participation, trust scores are turned into voting rights per user per entity. Third, for each user, a user model is inferred from the user's evaluation data, which we do using a generalized Bradley-Terry model. Fourth, users' models are scaled, using MEHESTAN and other solutions. Fifth, these models are securely aggregated, using QRMED among other algorithms. Sixth and finally, models are post-processed to yield human-readable global scores for the evaluated entities. Our pipeline has been successfully deployed on the open-source platform `tournesol.app`. We believe that it lays an appealing reusable foundation for the collaborative, effective, scalable, fair, interpretable and secure scoring of any set of entities.

disinformation groups, whose coordinated attacks are endangering the value of the system (Elliott & Gilbert, 2023).

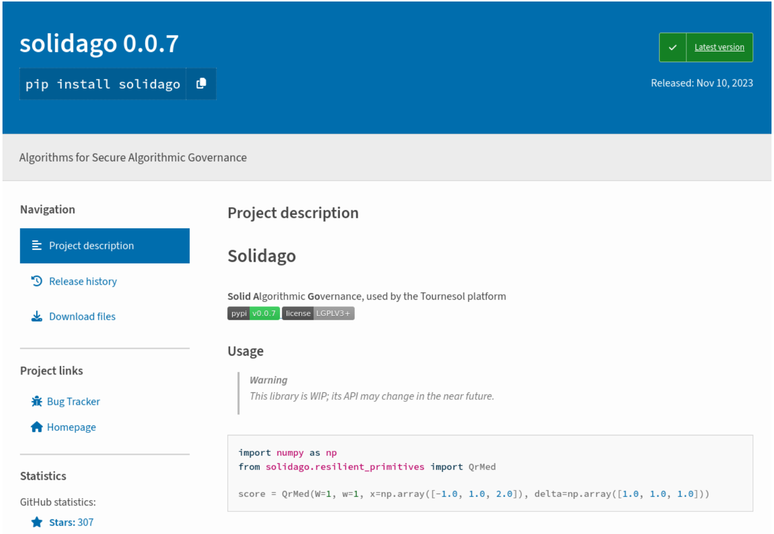
Unfortunately, building information systems that appropriately prioritize information (and its societal implications) is arguably under-researched, and currently lacks satisfactory solutions. As a result, perhaps unsurprisingly, today's algorithms are mostly designed, managed and governed in a relatively unilateral and opaque manner. As exposed by the Facebook Files (Hagey & Horwitz, 2023), while these algorithms shape narratives and public and geopolitical attention, they are benefiting from an alarming lack of accountability.

Our paper presents a contribution to the algorithmic toolbox for collaborative governance, and to the understanding of its challenges. More precisely, our goal is to provide an end-to-end modular pipeline, which we instantiate with state-of-the-art algorithms, to allow any community of non-experts to securely and collaboratively score any number of entities. More specifically, we make the following contributions.

Contributions. Our main contribution is to introduce a modular end-to-end pipeline called SOLIDAGO[1]. SOLIDAGO's six modules are (1) *trust propagation*, (2) *voting rights assignment*, (3) *user model inference*, (4) *model*

```python
@dataclass
class DefaultPipeline:
    """ Instantiates the default pipeline described in
    "Solidago: A Modular Pipeline for Collaborative Scoring".
    """
    trust_propagation: TrustPropagation = LipschiTrust(
        pretrust_value=0.8,
        decay=0.8,
        sink_vouch=5.0,
        error=1e-8
    )
    voting_rights: VotingRights = AffineOvertrust(
        privacy_penalty=0.5,
        min_overtrust=2.0,
        overtrust_ratio=0.1,
    )
    preference_learning: PreferenceLearning = UniformGBT(
        prior_std_dev=7,
        convergence_error=1e-5,
        cumulant_generating_function_error=1e-5,
    )
    scaling: Scaling = ScalingCompose(
        Mehestan(
            lipschitz=0.1,
            min_activity=10,
            n_scalers_max=100,
            privacy_penalty=0.5,
            p_norm_for_multiplicative_resilience=4.0,
            error=1e-5
        ),
        QuantileZeroShift(
            zero_quantile=0.15,
            lipschitz=0.1,
            error=1e-5
        )
    )
    aggregation: Aggregation = QuantileStandardizedQrMedian(
        dev_quantile=0.9,
        lipschitz=0.1,
        error=1e-5
    )
    post_process: PostProcess = Squash(
        score_max=100
    )
```

pip install solidago



solidago 0.0.7

`pip install solidago`

Latest version

Released: Nov 10, 2023

Algorithms for Secure Algorithmic Governance

Navigation

- Project description
- Release history
- Download files

Project links

- Bug Tracker
- Homepage

Statistics

GitHub statistics:

- **Stars:** 307

Project description

Solidago

Solid Algorithmic **Go**vernance, used by the Tournesol platform

pypi v0.0.7 license LGPLV3+

Usage

> *Warning*
> *This library is WIP; its API may change in the near future.*

```python
import numpy as np
from solidago.resilient_primitives import QrMed

score = QrMed(W=1, w=1, x=np.array([-1.0, 1.0, 2.0]), delta=np.array([1.0, 1.0, 1.0]))
```

Plein de collaborations possiblse

Research

"We seek to support research on the ethics of algorithms by providing a large and reliable database of human judgments."

Our data are open

We hope that other projects can benefit from the efforts of the Tournesol community. To this end we are making available a database made up of all public contributions that anyone can use.

These data are published under the terms of the Open Data Commons Attribution License (ODC-BY 1.0).

DOWNLOAD THE DATABASE

Our algorithms are Free/Libre

In a perspective of transparency and knowledge sharing, the algorithms and all source code we created are Free Software.

ACCESS THE CODE ON GITHUB

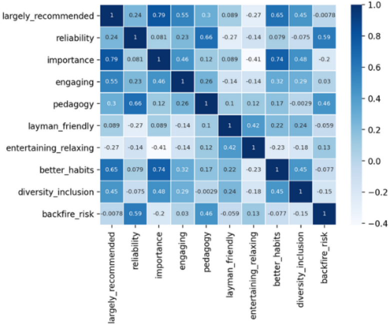
Discover the Tournesol Talks

We organize public online events with experts to talk about their works related to ethics, algorithms, information and more.

SEE THE TOURNESOL TALKS

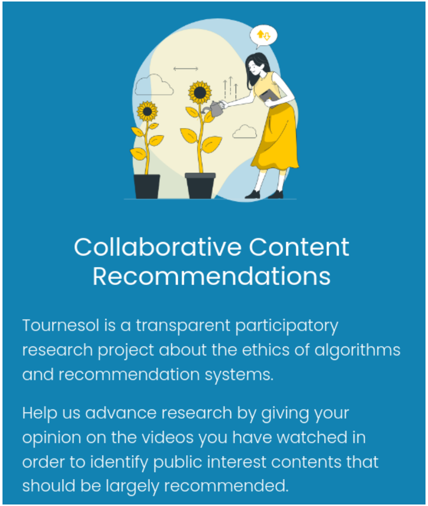
Visualize the data

You can quickly explore our public database with our Tournesol Data Visualization application made with Streamlit.



Pearson correlation coefficients matrix of comparison criteria scores (2023/06/13).

Un torrent de questions de recherche fascinantes !



- Proof of Personhood.
- Démocratie liquide.
- Réseau de confiance.
- Filtrage collaboratif Lipschitz-résilient.
- Vote bayésien Lipschitz-résilient.
- Apprentissage actif.
- Diversité et équité des recommendations.
- Interface humain-machine et ludification.
- Impacts cognitifs sur les consommateurs.
- Apprentissage de la volition.
- Présomption de non-recommandabilité.
- `tournesol.app/#research`

Section 7

Conclusion

“On est face à un abime.”

Nathalie Riché (notre éditrice)

“On est face à un abime.”

Nathalie Riché (notre éditrice)

Syllogisme du politicien

- Il faut faire quelque chose.
- X est quelque chose.
- Donc il faut faire X.

"On est face à un abime."

Nathalie Riché (notre éditrice)

Syllogisme du politicien

- Il faut faire quelque chose.
- X est quelque chose.
- Donc il faut faire X.

Cynisme éclairé

- Seul l'impact compte vraiment.
- Mon impact est négligeable.
- Donc je n'ai pas à agir.

Trois raisons (plus rationnelles) d'agir

Trois raisons (plus rationnelles) d'agir

1. Chaque fraction de degré compte.

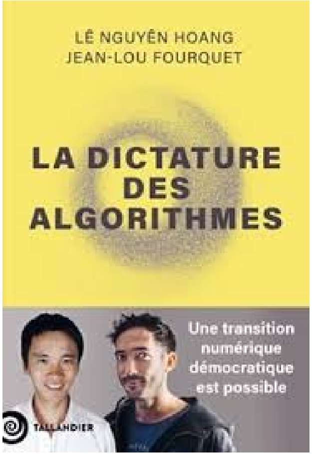
Trois raisons (plus rationnelles) d'agir

1. Chaque fraction de degré compte.
2. Créer des adjacents possibles.

Trois raisons (plus rationnelles) d'agir

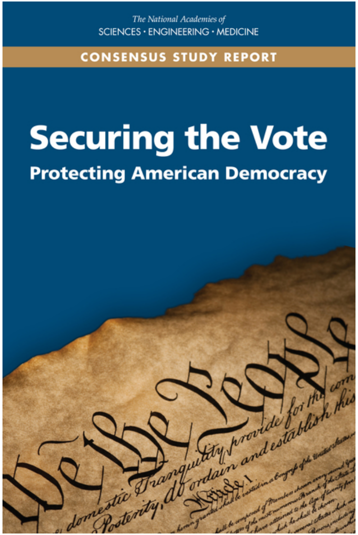
1. Chaque fraction de degré compte.
2. Créer des adjacents possibles.
3. Le plus beau des hobbys/métiers.

Notre ambition profonde : rendre le web démocratique



Photographie extraite du journal *Le Figaro*.

Pas de cybersécurité. Pas de démocratie numérique.



"À ce jour, Internet (ou n'importe quel réseau connecté à Internet) ne devrait pas être utilisé pour l'envoi de bulletins de vote remplis. De plus, le vote par Internet ne devrait pas être utilisé dans le futur, jusqu'au jour où, si ce jour arrive, des garanties très robustes de sécurité et de vérifiabilité sont développées et mises en place, sachant qu'aucune technologie connue ne garantit le secret, la sécurité et la vérifiabilité d'un bulletin rempli transmis via Internet."