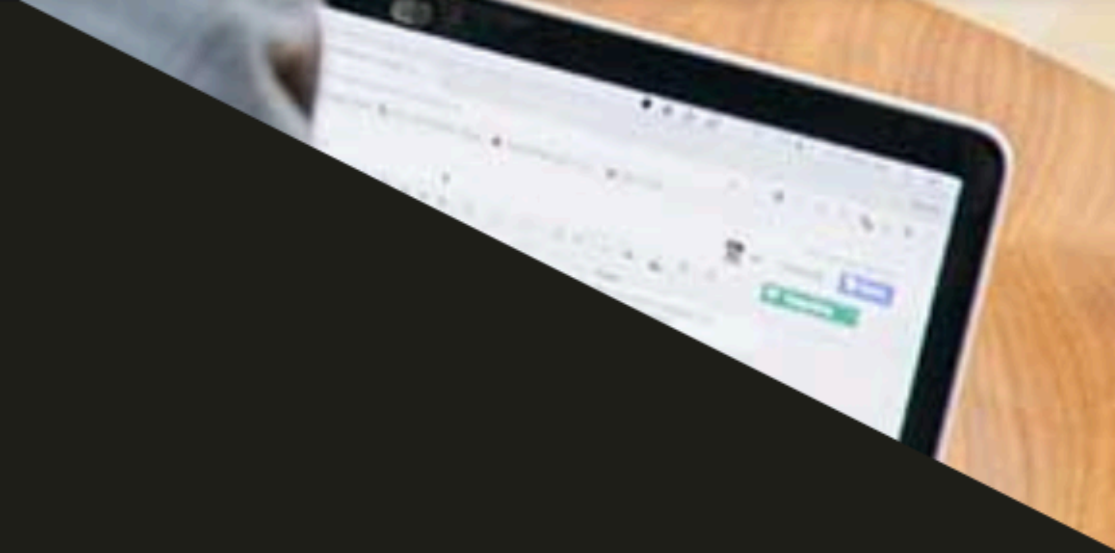
The Fabulous Endeavor
Make AI Robustly Beneficial

Lê Nguyên Hoang, EPFL
@le_science4all

Applied Machine Learning Days 2020



YouTube in numbers

One billion hours watched daily

← This is the number of hours of video watched on YouTube every day, generating billions of views. →

View all YouTube statistics

YouTube CH
Home
Trending
Subscriptions
Library
FILTER

donald trump

???

????????????

??????? • ??? vues • il y a ?? heures

??
???...

???

????????????

??????? • ??? vues • il y a ?? heures

??
???...

???

????????????

??????? • ??? vues • il y a ?? heures

??
???...

???

????????????

??????? • ??? vues • il y a ?? heures

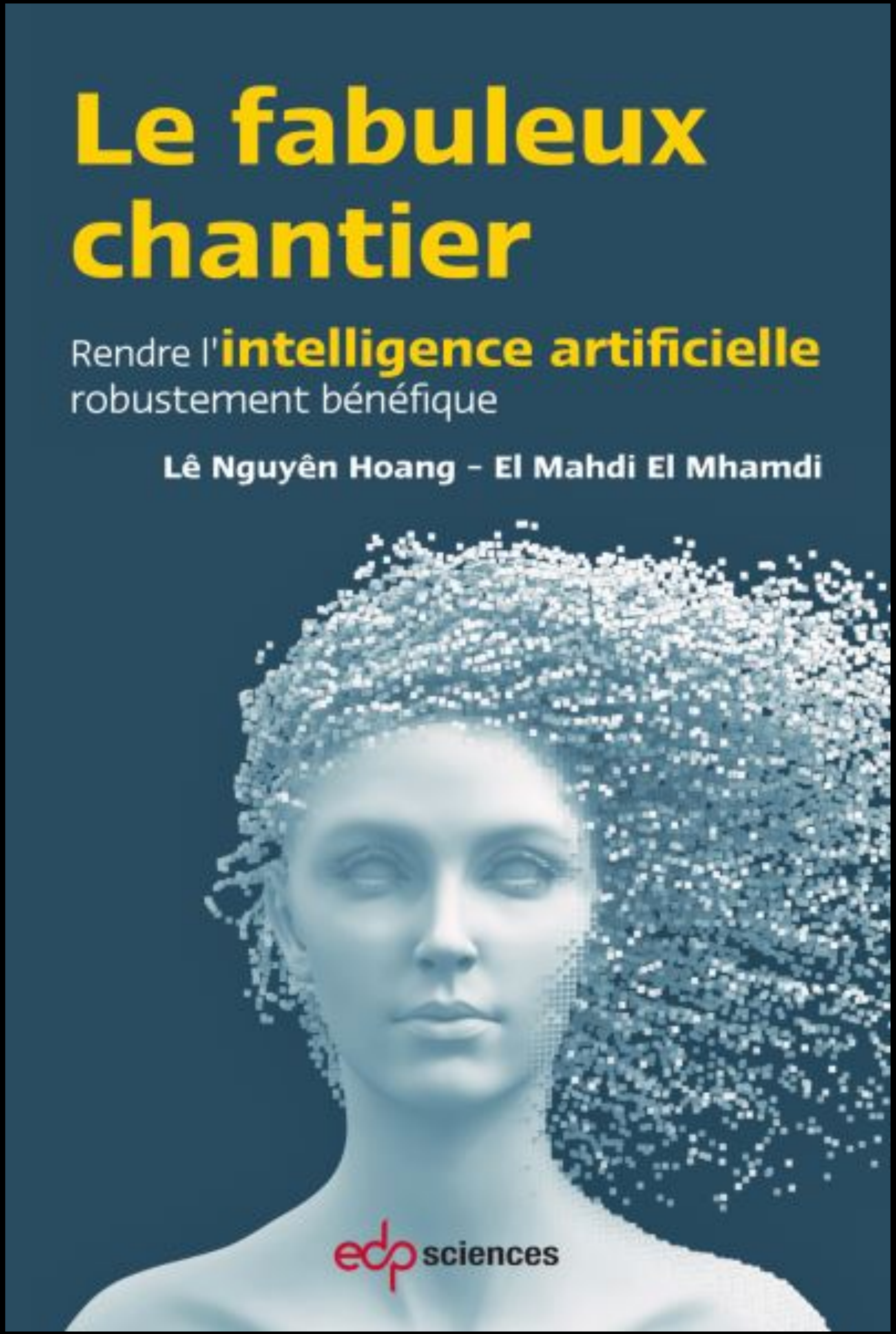
??
???...

"Those who control the flow of data in the world control the future, not only of humanity, but also perhaps of life itself."

Thesis 1

Make algorithms robustly beneficial is urgent.

#TheFabulousEndeavor



Le fabuleux chantier

Rendre l'**intelligence artificielle** robustement bénéfique

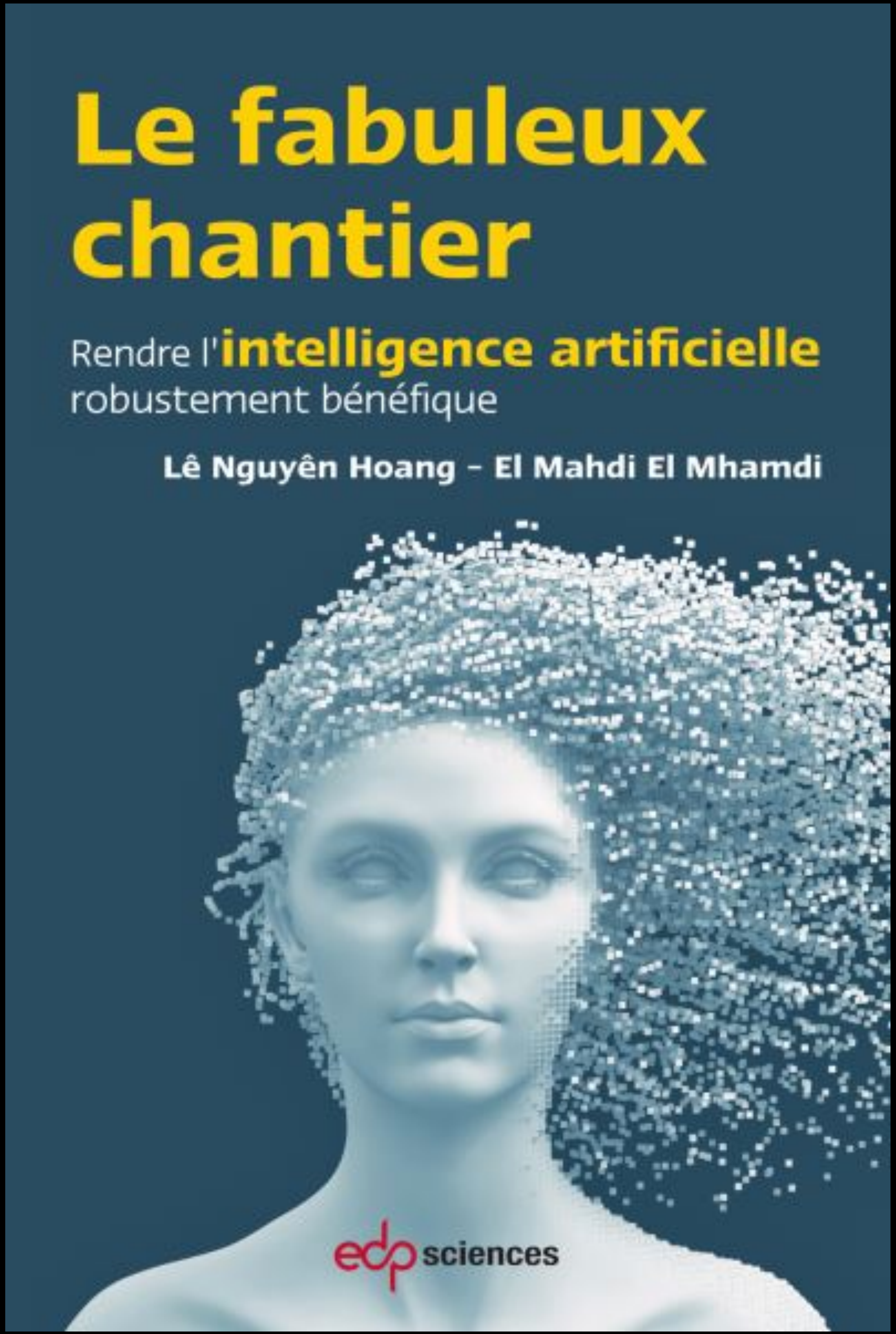
Lê Nguyên Hoang – El Mahdi El Mhamdi

edp sciences

Thesis 2

Make algorithms robustly beneficial is extremely hard.

#TheFabulousEndeavor


Le fabuleux chantier
Rendre l'intelligence artificielle robustement bénéfique
Lê Nguyên Hoang – El Mahdi El Mhamdi
edp sciences

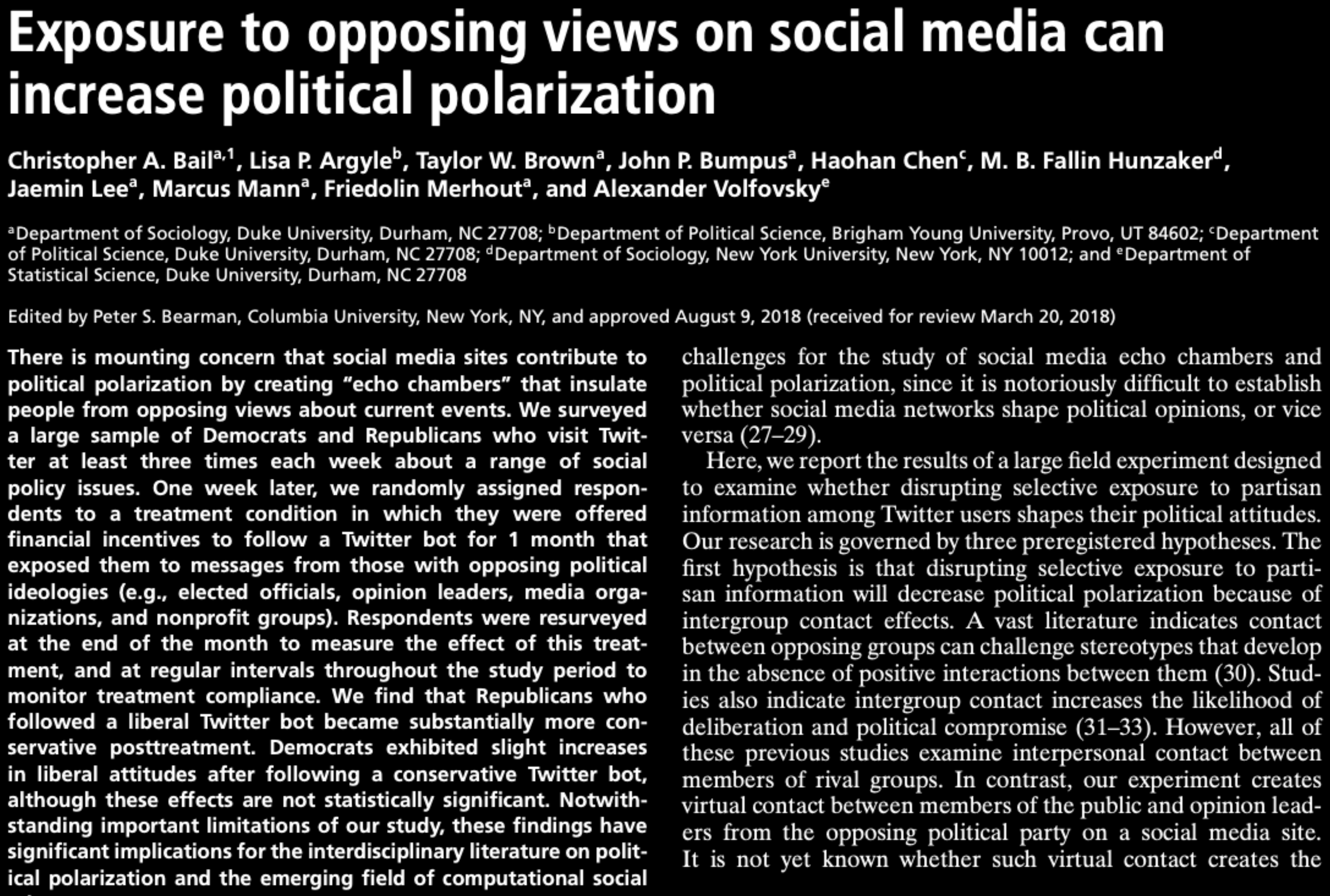
Anticipate side effects

The YouTube algorithm wants to entertain.

The YouTube algorithm wants to entertain.

What's so bad about that?

NETFLIX
N FILM
THE GREAT HACK



Exposure to opposing views on social media can increase political polarization

Christopher A. Bail[a,1], Lisa P. Argyle[b], Taylor W. Brown[a], John P. Bumpus[a], Haohan Chen[c], M. B. Fallin Hunzaker[d], Jaemin Lee[a], Marcus Mann[a], Friedolin Merhout[a], and Alexander Volfovsky[e]

[a]Department of Sociology, Duke University, Durham, NC 27708; [b]Department of Political Science, Brigham Young University, Provo, UT 84602; [c]Department of Political Science, Duke University, Durham, NC 27708; [d]Department of Sociology, New York University, New York, NY 10012; and [e]Department of Statistical Science, Duke University, Durham, NC 27708

Edited by Peter S. Bearman, Columbia University, New York, NY, and approved August 9, 2018 (received for review March 20, 2018)

There is mounting concern that social media sites contribute to political polarization by creating "echo chambers" that insulate people from opposing views about current events. We surveyed a large sample of Democrats and Republicans who visit Twitter at least three times each week about a range of social policy issues. One week later, we randomly assigned respondents to a treatment condition in which they were offered financial incentives to follow a Twitter bot for 1 month that exposed them to messages from those with opposing political ideologies (e.g., elected officials, opinion leaders, media organizations, and nonprofit groups). Respondents were resurveyed at the end of the month to measure the effect of this treatment, and at regular intervals throughout the study period to monitor treatment compliance. We find that Republicans who followed a liberal Twitter bot became substantially more conservative posttreatment. Democrats exhibited slight increases in liberal attitudes after following a conservative Twitter bot, although these effects are not statistically significant. Notwithstanding important limitations of our study, these findings have significant implications for the interdisciplinary literature on political polarization and the emerging field of computational social

challenges for the study of social media echo chambers and political polarization, since it is notoriously difficult to establish whether social media networks shape political opinions, or vice versa (27–29).

Here, we report the results of a large field experiment designed to examine whether disrupting selective exposure to partisan information among Twitter users shapes their political attitudes. Our research is governed by three preregistered hypotheses. The first hypothesis is that disrupting selective exposure to partisan information will decrease political polarization because of intergroup contact effects. A vast literature indicates contact between opposing groups can challenge stereotypes that develop in the absence of positive interactions between them (30). Studies also indicate intergroup contact increases the likelihood of deliberation and political compromise (31–33). However, all of these previous studies examine interpersonal contact between members of rival groups. In contrast, our experiment creates virtual contact between members of the public and opinion leaders from the opposing political party on a social media site. It is not yet known whether such virtual contact creates the

Robustness to attacks

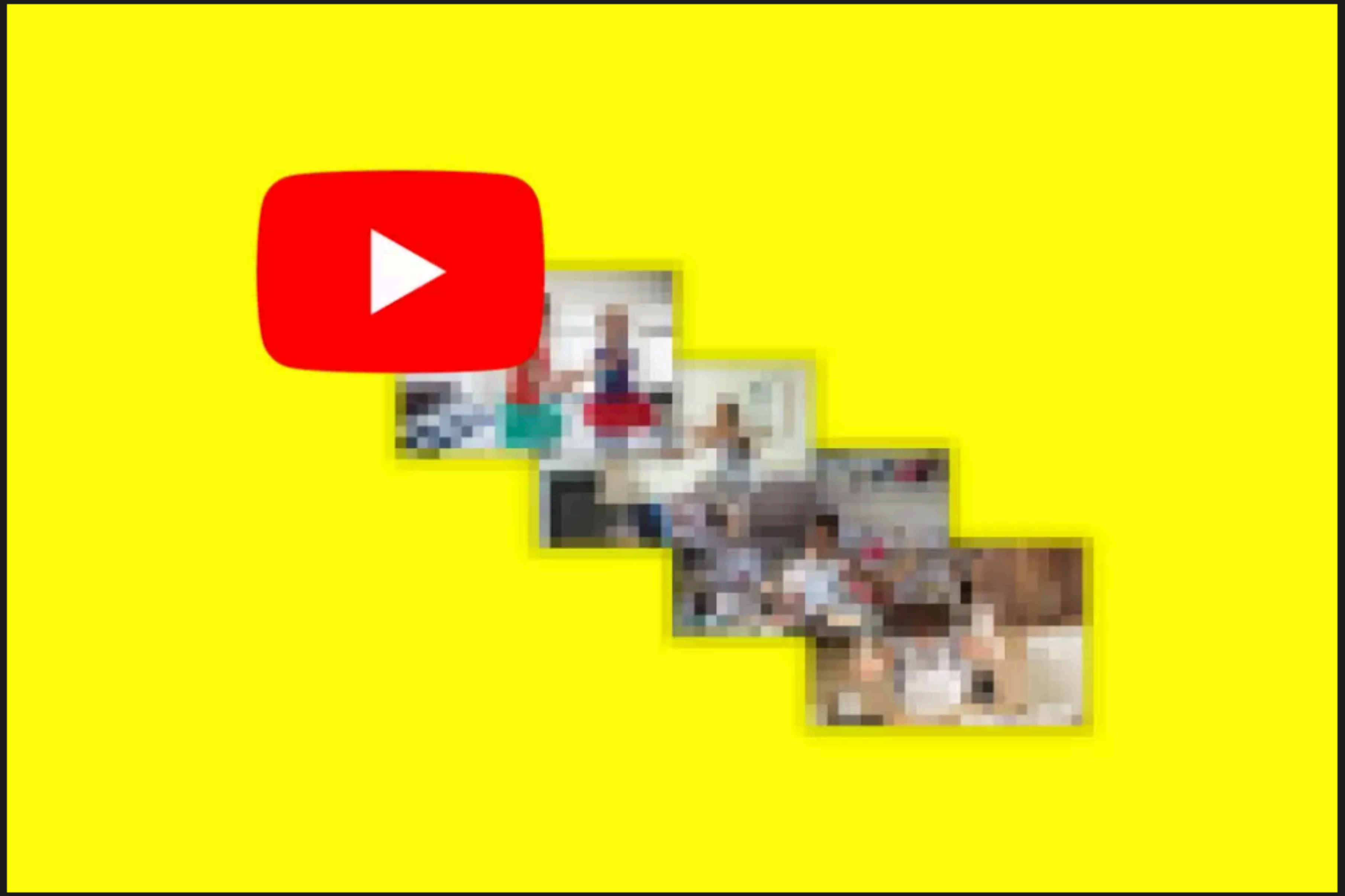
YouTube

On YouTube, a network of paedophiles is hiding in plain sight

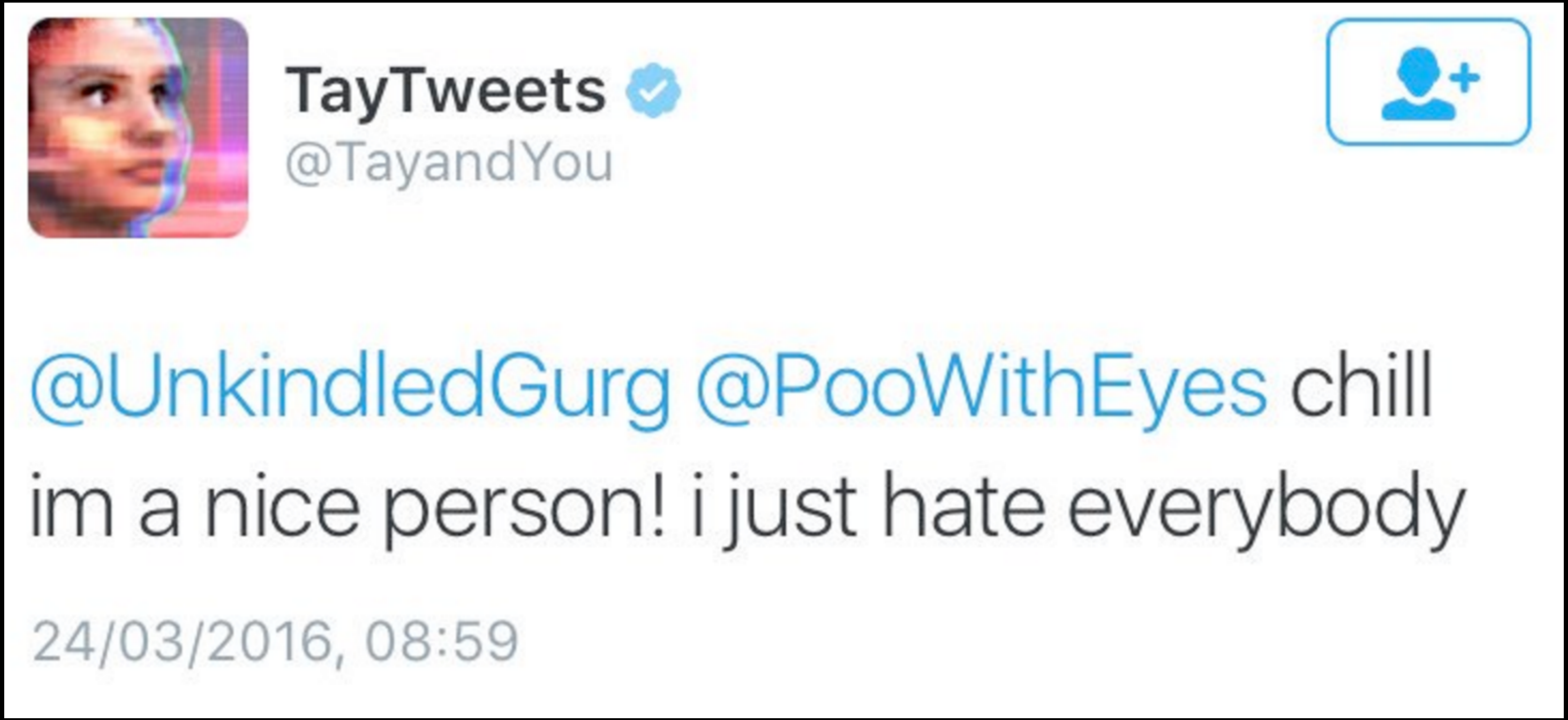
Scores of YouTube videos with tens of millions of views are being inundated with comments by paedophiles, with adverts from major brands running alongside the disturbing content

By **K.G ORPHANIDES**

Wednesday 20 February 2019



Credit **Getty images / WIRED**



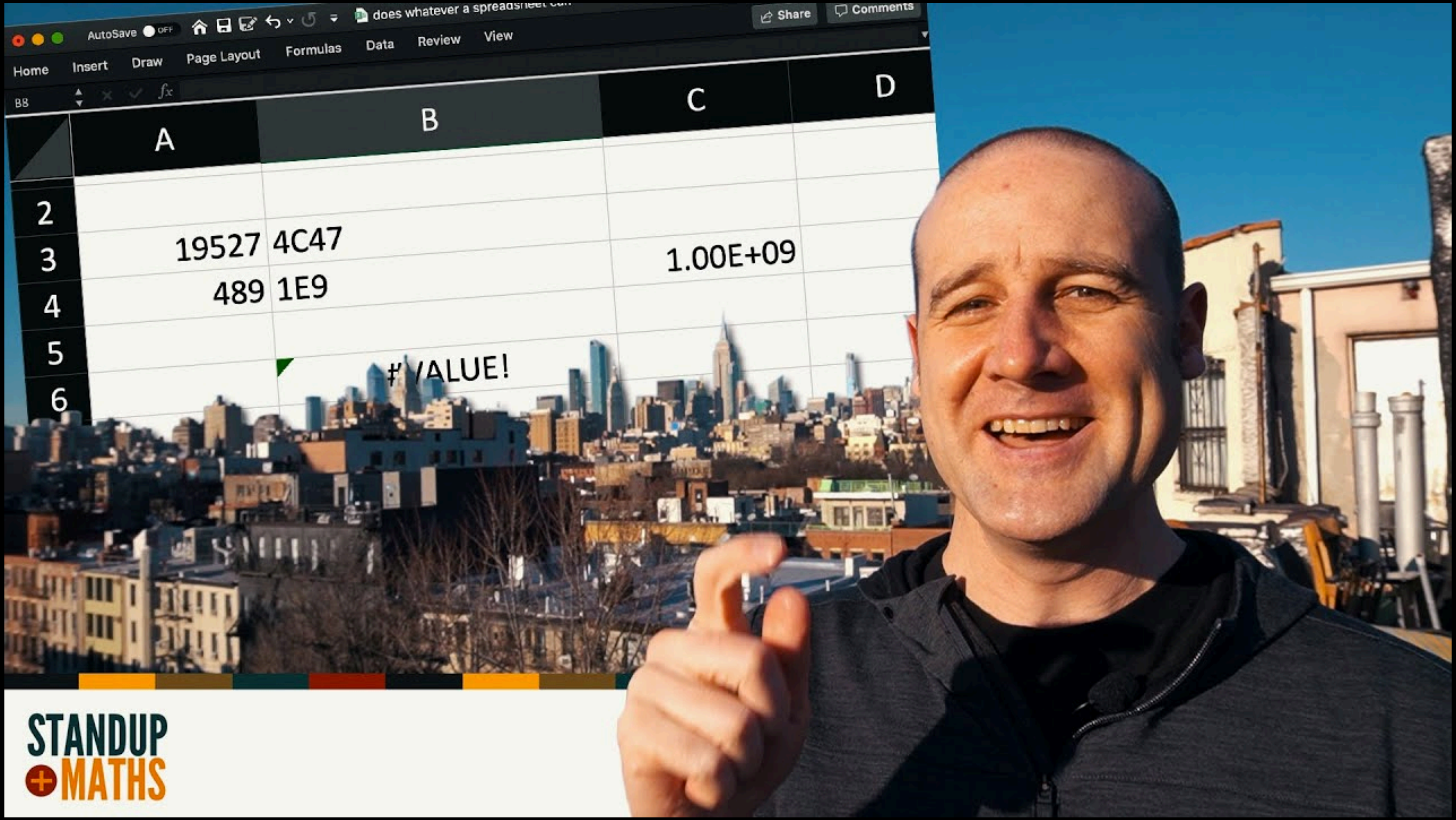
TayTweets

@TayandYou

@UnkindledGurg @PooWithEyes chill im a nice person! i just hate everybody

24/03/2016, 08:59

Do you trust 100%
of your data?



When Spreadsheets Attack!

standupmaths

Should you trust an algorithm that relies on 99%-valid data?

Machine Learning with Adversaries: Byzantine Tolerant Gradient Descent

Peva Blanchard
EPFL, Switzerland
peva.blanchard@epfl.ch

El Mahdi El Mhamdi*
EPFL, Switzerland
elmahdi.elmhamdi@epfl.ch

Rachid Guerraoui
EPFL, Switzerland
rachid.guerraoui@epfl.ch

Julien Stainer
EPFL, Switzerland
julien.stainer@epfl.ch

Abstract

We study the resilience to Byzantine failures of distributed implementations of Stochastic Gradient Descent (SGD). So far, distributed machine learning frameworks have largely ignored the possibility of failures, especially arbitrary (i.e., Byzantine) ones. Causes of failures include software bugs, network asynchrony, biases in local datasets, as well as attackers trying to compromise the entire system. Assuming a set of n workers, up to f being Byzantine, we ask how resilient can SGD be, without limiting the dimension, nor the size of the parameter space. We first show that no gradient aggregation rule based on a linear combination of the vectors proposed by the workers (i.e, current approaches) tolerates a single Byzantine failure. We then formulate a resilience property of the aggregation rule capturing the basic requirements to guarantee convergence despite f Byzantine workers. We propose *Krum*, an aggregation rule that satisfies our resilience property, which we argue is the first provably Byzantine-resilient algorithm for distributed SGD. We also report on experimental evaluations of Krum.

Asynchronous Byzantine Machine Learning (the case of SGD)

Georgios Damaskinos [1] El Mahdi El Mhamdi [1] Rachid Guerraoui [1] Rhicheek Patra [1] Mahsa Taziki [1]

Abstract

Asynchronous distributed machine learning solutions have proven very effective so far, but always assuming perfectly functioning workers. In practice, some of the workers can however exhibit Byzantine behavior, caused by hardware failures, software bugs, corrupt data, or even malicious attacks. We introduce *Kardam*, the first distributed asynchronous stochastic gradient descent (SGD) algorithm that copes with Byzantine workers. Kardam consists of two complementary components: a filtering and a dampening component. The first is scalar-based and ensures resilience against $\frac{1}{3}$ Byzantine workers. Essentially, this filter leverages the Lipschitzness of cost functions and acts as a self-stabilizer against Byzantine workers that would attempt to corrupt the progress of SGD. The dampening component bounds the convergence rate by adjusting to stale information through a generic gradient weighting scheme. We prove that Kardam guarantees almost sure convergence in the presence of asynchrony and Byzantine behavior, and we derive its convergence rate. We evaluate Kardam on the CIFAR-100 and EMNIST datasets and measure its overhead with respect to non Byzantine-resilient solutions. We empirically show that Kardam does not introduce additional noise to the learning procedure but does induce a slowdown (the cost of Byzantine resilience) that we both theoretically and empirically show to be less than f/n, where f is the number of Byzantine failures tolerated and n the total number of workers. Interestingly, we also empirically observe that the dampening component is interesting in its own right for it enables to build an SGD algorithm that outperforms alternative staleness-aware asynchronous competitors in environments with honest workers.

1. Introduction

To keep up with the amount of data available today and the corresponding increasing demand for resources, machine learning (ML) practitioners rely on large scale distributed systems (Dean et al., 2012; Abadi et al., 2016; Li et al., 2013; 2014b;a; Ho et al., 2013; Cui et al., 2016). Most of these systems make use of the same work-horse optimization algorithm: *stochastic gradient descent* (SGD), typically following the parameter server scheme (Li et al., 2014a;b). The computation is divided into *epochs*, i.e., *model* (parameter vector) updates. The server gathers gradients from the workers and employs them to perform a single model update, then broadcasts the new model to every worker for computing new gradients (based on random data samples).

To be practical, a distributed ML solution should not assume that all workers perform perfectly well. Some arbitrary behavior of at least a fraction of the workers should be tolerated. The Byzantine failure model (Lamport et al., 1982) offers an elegant abstraction to reason about problems of adversarial machine learning. In particular, a Byzantine behavior can be due to a crash, a software bug, a stale local view of a model, a corrupt piece of data, or worse, to attackers that benefit from a security flaw in a device and compromise its behavior. The Byzantine model encompasses the problem of *poisoning attacks* (Biggio & Laskov, 2012; Muñoz-González et al., 2017; Kurakin et al., 2016). For instance, a group of adversarial (Byzantine) workers could bias the gradient estimator and prevent convergence by sending corrupt gradients. Byzantine-resilient (or simply Byzantine) ML solutions are very appealing for they do not make any assumption on the behavior of Byzantine workers.

A few Byzantine distributed ML solutions have been recently proposed (Blanchard et al., 2017; Su, 2017; El Mhamdi et al., 2018). All however assume a restrictive synchronous model. In each epoch, (1) all (honest) workers are supposed to use the exact same model to compute the gradient, and (2) the parameter server waits for a quorum of workers before aggregating their gradients. When networks

Social choice

The Self-Driving Trolley Dilemma

Up and Atom (ft. Science4All)

YouTube CH
Home
Trending
Subscriptions
Library
FILTER
???
???
???
???

donald trump

????????????

??????? • ??? vues • il y a ?? heures

???
???...

????????????

??????? • ??? vues • il y a ?? heures

???
???...

????????????

??????? • ??? vues • il y a ?? heures

???
???...

????????????

??????? • ??? vues • il y a ?? heures

???
???...

WeBuildAI: Participatory Framework for Algorithmic Governance

MIN KYUNG LEE, University of Texas at Austin & Carnegie Mellon University, USA
DANIEL KUSBIT, Ethics, History & Public Policy, Carnegie Mellon University, USA
ANSON KAHNG, School of Computer Science, Carnegie Mellon University, USA
JI TAE KIM, School of Design, Carnegie Mellon University, USA
XINRAN YUAN, Information Systems, Carnegie Mellon University, USA
ALLISSA CHAN, School of Design, Carnegie Mellon University, USA
DANIEL SEE, Decision Science & Art, Carnegie Mellon University, USA
RITESH NOOTHIGATTU, School of Computer Science, Carnegie Mellon University, USA
SIHEON LEE, Information Systems, Carnegie Mellon University, USA
ALEXANDROS PSOMAS, School of Computer Science, Carnegie Mellon University, USA
ARIEL D. PROCACCIA, School of Computer Science, Carnegie Mellon University, USA

Algorithms increasingly govern societal functions, impacting multiple stakeholders and social groups. How can we design these algorithms to balance varying interests in a moral, legitimate way? As one answer to this question, we present WeBuildAI, a collective participatory framework that enables people to build algorithmic policy for their communities. The key idea of the framework is to enable stakeholders to construct a computational model that represents their views and to have those models vote on their behalf to create algorithmic policy. As a case study, we applied this framework to a matching algorithm that operates an on-demand food donation transportation service in order to adjudicate equity and efficiency trade-offs. The service's stakeholders—donors, volunteers, recipient organizations, and nonprofit employees—used the framework to design the algorithm through a series of studies in which we researched their experiences. Our findings suggest that the framework successfully enabled participants to build models that they felt confident represented their own beliefs. Participatory algorithm design also improved both procedural fairness and the distributive outcomes of the algorithm, raised participants' algorithmic awareness, and helped identify inconsistencies in human decision-making in the governing organization. Our work demonstrates the feasibility, potential and challenges of community involvement in algorithm design.

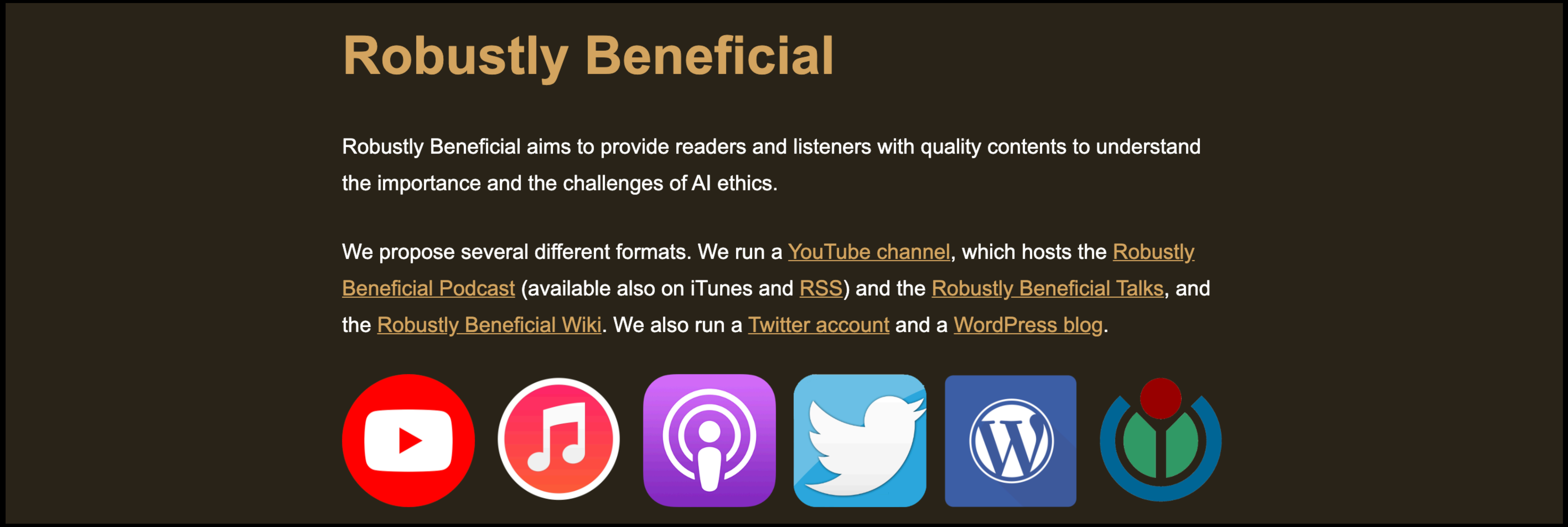
CCS Concepts: • **Human-centered computing → Human computer interaction (HCI)**.

Additional Key Words and Phrases: participatory algorithm design, collective participation, human-centered AI, matching algorithm, algorithmic fairness

Designing trustworthy algorithms has become a business challenge / opportunity.

Conclusion

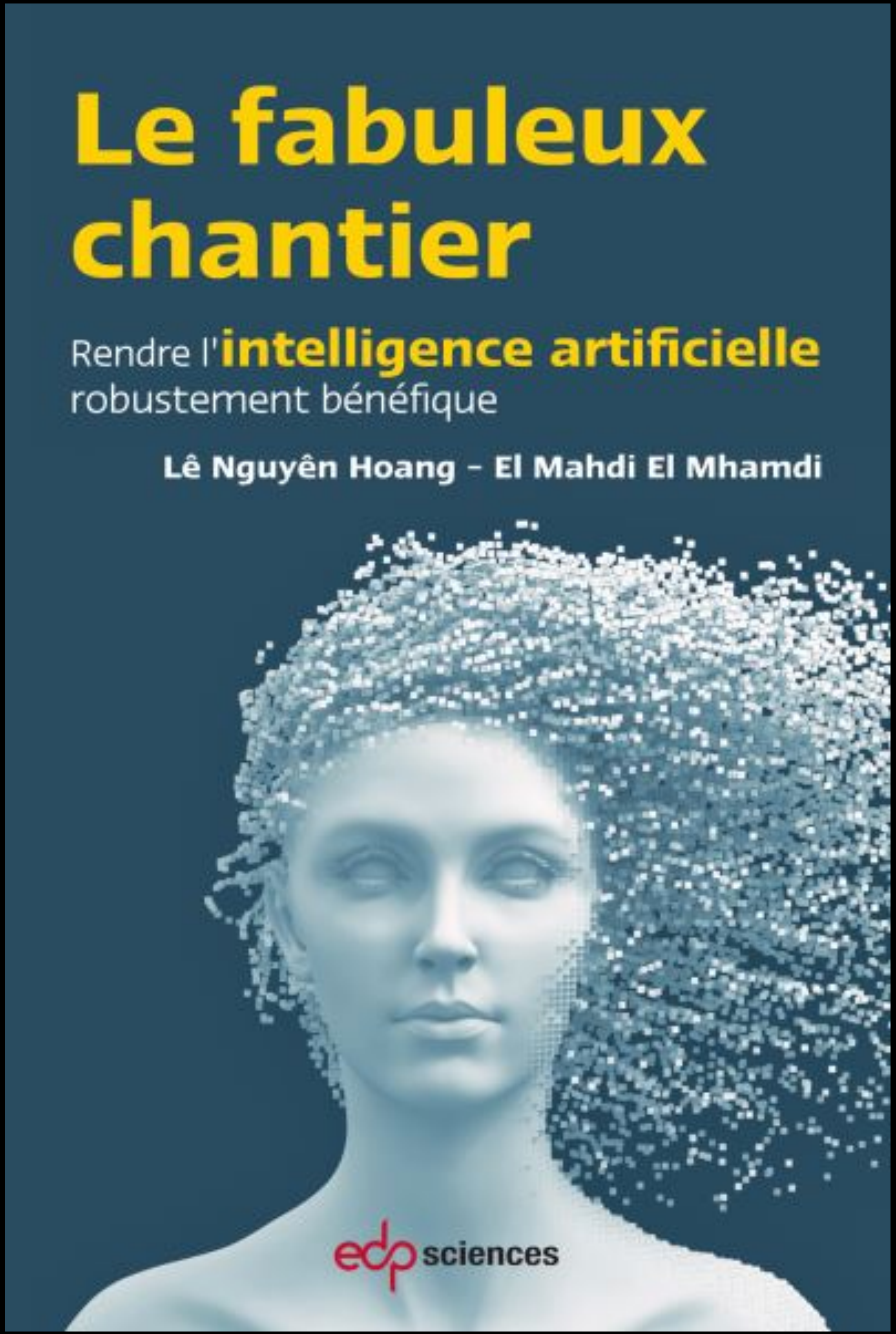
Make algorithms robustly beneficial is a fabulous endeavor.



Robustly Beneficial

Robustly Beneficial aims to provide readers and listeners with quality contents to understand the importance and the challenges of AI ethics.

We propose several different formats. We run a YouTube channel, which hosts the Robustly Beneficial Podcast (available also on iTunes and RSS) and the Robustly Beneficial Talks, and the Robustly Beneficial Wiki. We also run a Twitter account and a WordPress blog.



Le fabuleux chantier

Rendre l'intelligence artificielle robustement bénéfique

Lê Nguyên Hoang – El Mahdi El Mhamdi

edp sciences