



AI | BLINDSPOT

A discovery process for spotting unconscious biases
and structural inequalities in AI systems

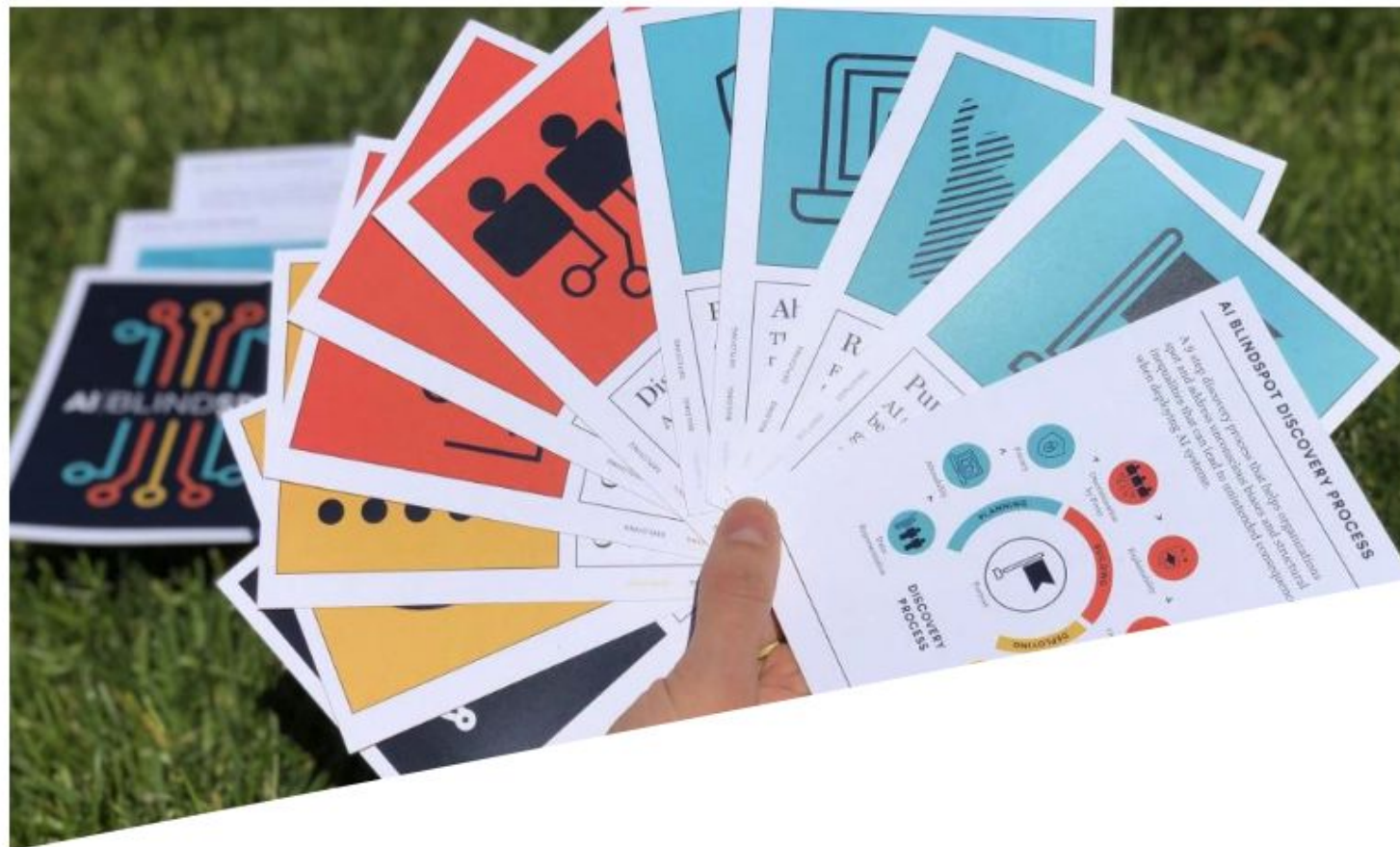
Mozfest 2019



Openness

Overview

- Introduction (5min)
- AI Blindspot discovery process (10min)
- Case study: YouTube video recommendations (10min)
- Simulation 1 (20min)
 - **Recruitment** - Hirevue
 - Report back
- Simulation 2 (20min)
 - **Facial recognition for border crossing**
 - Report back
- What's next (10min)



[HOME](#) > [NEWS](#) >

What We Created During Assembly 2019



Case study: YouTube recommendation algorithm

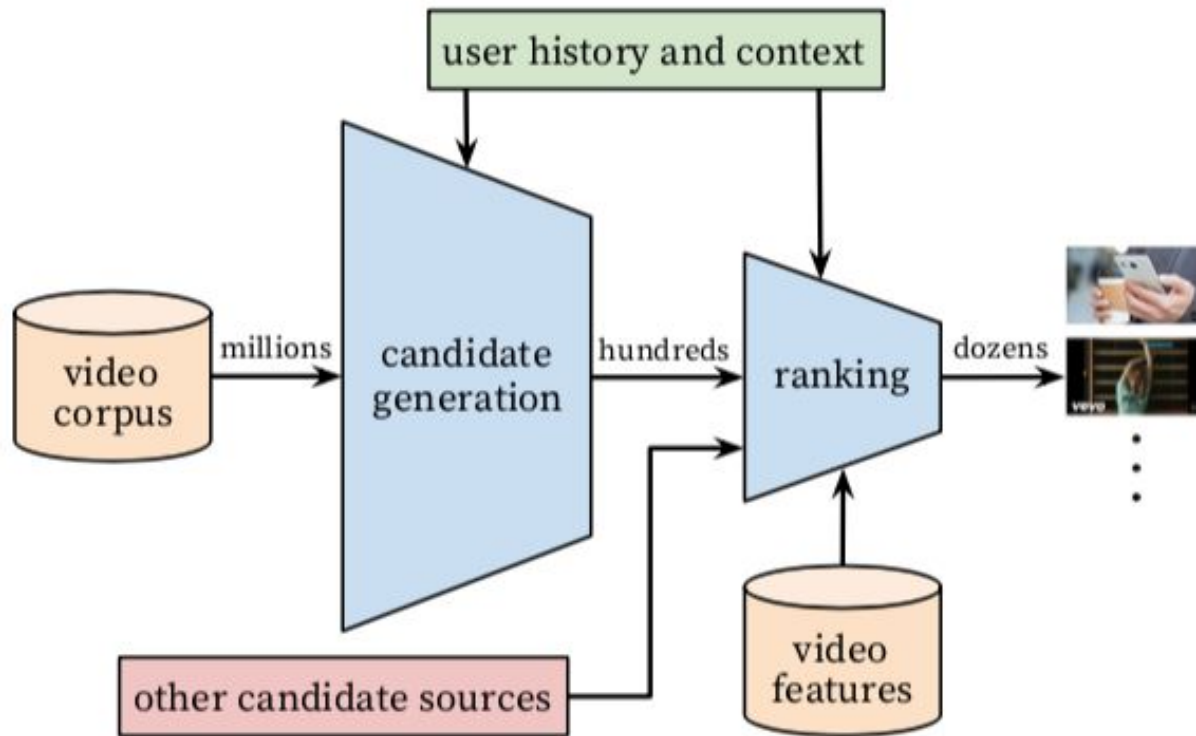


Figure 2: Recommendation system architecture demonstrating the “funnel” where candidate videos are retrieved and ranked before presenting only a few to the user.





Nudity or sexual content

YouTube is not for pornography or sexually explicit content. If this describes your video, even if it's a video of yourself, don't post it on YouTube. Also, be advised that we work closely with law enforcement and we report child exploitation. [Learn more](#)



Harmful or dangerous content

Don't post videos that encourage others to do things that might cause them to get badly hurt, especially kids. Videos showing such harmful or dangerous acts may get age-restricted or removed depending on their severity. [Learn more](#)



Hateful content

Our products are platforms for free expression. But we don't support content that promotes or condones violence against individuals or groups based on race or ethnic origin, religion, disability, gender, age, nationality, veteran status, or sexual orientation/gender identity, or whose primary purpose is inciting hatred on the basis of these core characteristics. This can be a delicate balancing act, but if the primary purpose is to attack a protected group, the content crosses the line. [Learn more](#)



Violent or graphic content

It's not okay to post violent or gory content that's primarily intended to be shocking, sensational, or gratuitous. If posting graphic content in a news or documentary context, please be mindful to provide enough information to help people understand what's going on in the video. Don't encourage others to commit specific acts of violence. [Learn more](#)



Harassment and cyberbullying

It's not ok to post abusive videos and comments on YouTube. If harassment crosses the line into a malicious attack it can be reported and may be removed. In other cases, users may be mildly annoying or petty and should be ignored. [Learn more](#)



Threats

Things like predatory behavior, stalking, threats, harassment, intimidation, invading privacy, revealing other people's personal information, and inciting others to commit violent acts or to violate the Terms of Use are taken very seriously. Anyone caught doing these things may be permanently banned from YouTube. [Learn more](#)



Privacy

If someone has posted your personal information or uploaded a video of you without your consent, you can request removal of content based on our Privacy Guidelines. [Learn more](#)



Spam, misleading metadata, and scams

Everyone hates spam. Don't create misleading descriptions, tags, titles, or thumbnails in order to increase views. It's not okay to post large amounts of untargeted, unwanted or repetitive content, including comments and private messages. [Learn more](#)



Copyright

Respect copyright. Only upload videos that you made or that you're authorized to use. This means don't upload videos you didn't make, or use content in your videos that someone else owns the copyright to, such as music tracks, snippets of copyrighted programs, or videos made by other users, without necessary authorizations. Visit our Copyright Center for more information. [Learn more](#)

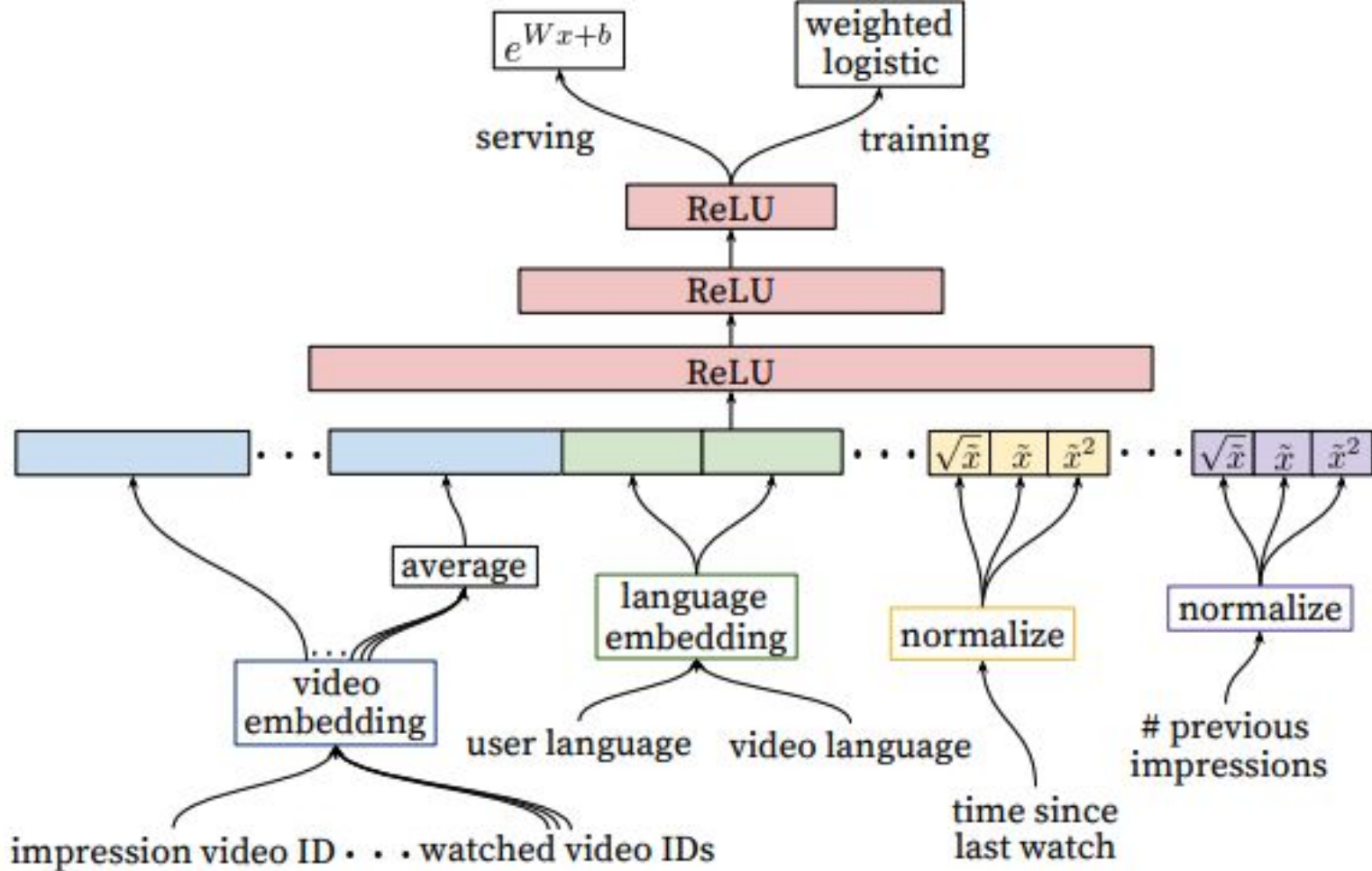


Impersonation

Accounts that are established to impersonate another channel or individual may be removed under our impersonation policy. [Learn more](#)

Deep Neural Networks for YouTube Recommendations

Paul Covington, Jay Adams, Emre Sargin
Google



Feature selection

Content signals

- Metadata
- Collaborative filtering
- Embedded video views
- Freshness
- Authoritative sources

User signals

- Video view history
- Search history
- Location and language
- Time since last watched
- Age

Abusability



UPLOADING 77%

★ Your video is still uploading. Please keep this page open until it's done.

Basic info Translations Advanced settings

|Title

Description

Tags (e.g., albert einstein, flying pig, mashup)

Upload status:
Uploading your video.
Your video will be live at:
<https://youtu.be/Qzh6iogcdcQ>

Video / Audio quality:
★ Your videos will process faster if you encode into a streamable file format. For more information, visit our [Help Center](#).

Auditing Radicalization Pathways on YouTube

Manoel Horta Ribeiro, Raphael Ottoni, Robert West, Virgílio A. F. Almeida, Wagner Meira

Optimization Criteria

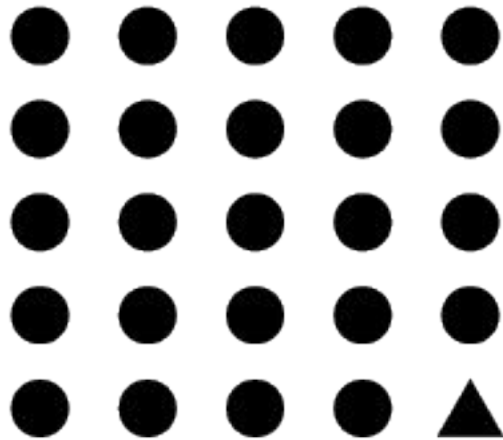


4.2 Modeling Expected Watch Time

Our goal is to predict expected watch time given training examples that are either positive (the video impression was clicked) or negative (the impression was not clicked). Positive examples are annotated with the amount of time the user spent watching the video. To predict expected watch time we use the technique of weighted logistic regression, which was developed for this purpose.

“Ranking by click-through rate often promotes deceptive videos that the user does not complete (“clickbait”) whereas watch time better captures engagement”

Generalization Error



Parents' Ultimate Guide to YouTube Kids

Is YouTube Kids really safe for kids? How does it work, and how do you set filters and parental controls? Learn more about this kid-targeted, but sometimes iffy, YouTube-lite app. By [Caroline Knorr](#) 6/17/2019

Topics: [Early Childhood](#), [Cellphone Parenting](#), [Healthy Media Habits](#), [Screen Time](#), [YouTube](#)

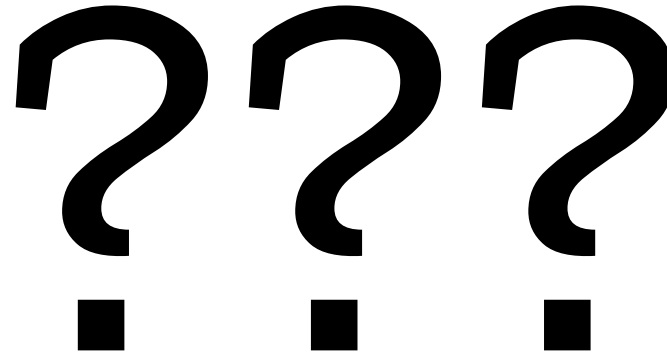
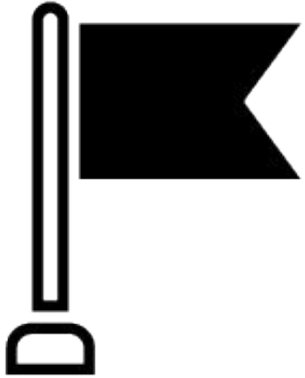
[Ver en español](#)

So many kids love watching videos on YouTube, it seemed like a slam dunk for Google to create a special app specifically for the online video service's youngest fans. And while [YouTube Kids](#) offers a colorful, easy-to-navigate environment, a wide range of high-quality videos, a few parental controls, and fun features for kids, it's been dogged by concerns over its advertising, branded content, and inappropriate clips slipping through the curation process. So is YouTube Kids right for kids -- or not?



Purpose

What is the purpose of the YouTube recommendation algorithm?



“Algorithms on sites like YouTube often push people towards extreme and misleading content. Overhauling these content recommendation systems could go a long way to curbing misinformation.”

Simulation 1

Applying for Your Next Job May Be an Automated Nightmare



Brian Merchant

4/24/19 1:50PM • Filed to: AUTOMATION



20.1K



76



2



If you think looking for a job is already daunting, anxiety-riddled, and unpleasant, just wait until the algorithms take over the hiring process. When they do, a newfangled “digital recruiter” like VCV, which [just received \\$1.7 million in early investment](#), hopes it will look something like this:

First, a search bot will be used to scan CVs by the thousands, yours presumably among them. If it’s picked out of the haystack, you will be contacted by a chatbot. Over SMS, the bot will set an appointment for a phone interview, which will be conducted by an automated system enabled by voice recognition AI. Next, the system will ask you, the applicant, to record video responses to a set of predetermined interview questions. Finally, the program can use facial recognition and predictive analytics to complete the screening, algorithmically determining whether [the nervousness, mood, and behavior patterns](#) you exhibit make you a fit for the company.

<https://gizmodo.com/applying-for-your-next-job-may-be-an-automated-nightmar-1834275825>

Simulation 1

Scenario

Recruitment:

Machine learning may offer significant potential to speed up and improve the quality of recruitment processes. However, historic recruiting patterns are often significantly biased and it is key that new tools are designed and implemented in a way which challenges rather than embeds these inequalities.

Roles

Executives

Developers

Affected communities

Civil society activists

In answering each question each person should take the perspective of the role they are playing.

Simulation 2

THE WALL STREET JOURNAL.

Subscribe | Sign In

Home World U.S. Politics Economy Business Tech Markets Opinion **Life & Arts** Real Estate WSJ. Magazine

Search 

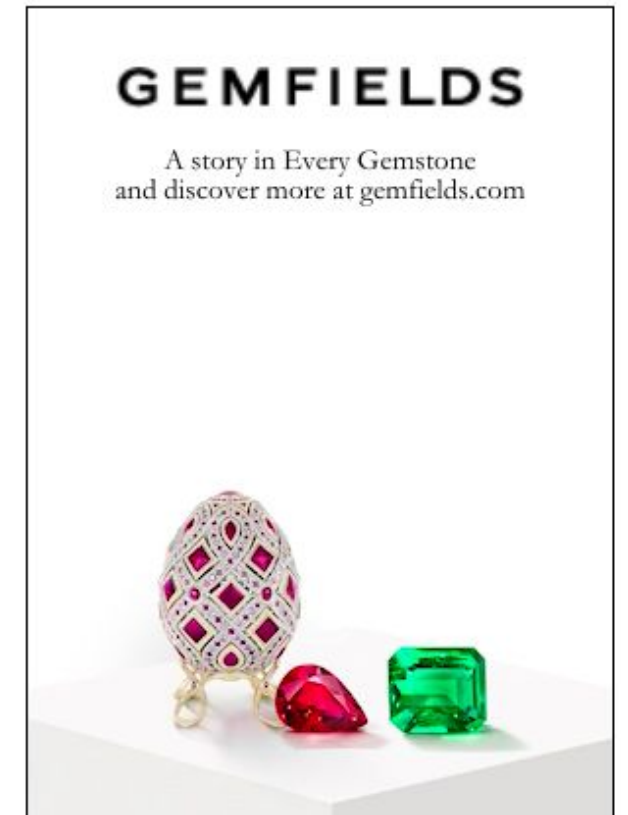
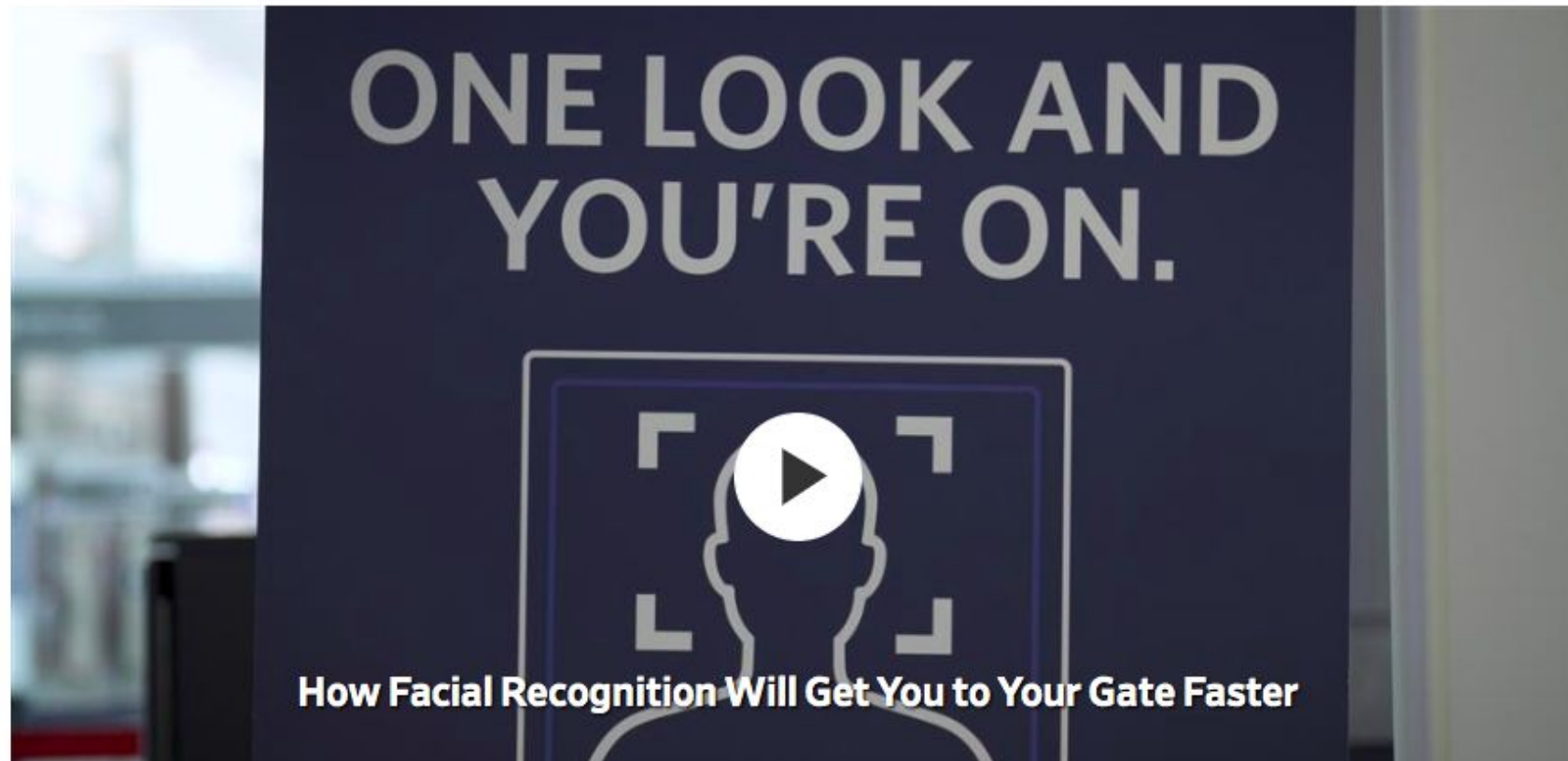
 SHARE

 AA
TEXT

 99

Are You Ready for Facial Recognition at the Airport?

Airlines and TSA are starting to scan faces to get people through security and boarding faster, as privacy advocates warn of unintended consequences



Simulation 2

Scenario

Crime and justice:

The use of predictive algorithms in policing is a growing field. Measures are being taken to introduce real time facial recognition tools at border crossings around the world, even as privacy concerns abound. Existing bias in policing and facial recognition is well documented, and inaccurate and unfair decision making in this sector can have enormous consequences for individuals and society.

Roles

- Executives
- Developers
- Affected communities
- Civil society activists

In answering each question each person should take the perspective of the role they are playing.

What's next?

- Fill a gap between ethical principles and technical processes and workflows on the ground
- Incubated by the Berkman Klein Center at Harvard
- Workshops to train data scientists, civil society activists, government officials
- Establish by-laws for collectively revising and expanding the blindspots
- Build coalitions and consensus for a shared vocabulary
- Stay in touch!
 - info@aiblindspot.com [@aiblindspot](https://twitter.com/aiblindspot)