PARTNERSHIP ON AI

# The Deepfake Detection Dilemma: A Multistakeholder Exploration of Adversarial Dynamics in Synthetic Media

Claire Leibowicz,[1,*] Sean McGregor,[2,*] Aviv Ovadya[3,*]
Artificial Intelligence, Ethics, and Society (AIES21)
* All authors contributed equally
1: The Partnership on AI, 2: XPRIZE Foundation, Syntiant Corp, 3: Thoughtful Technology Project

## Abstract
Synthetic media detection technologies label media as either synthetic or non-synthetic and are increasingly used by journalists, web platforms, and the general public to identify misinformation and other forms of problematic content. As both well-resourced organizations and the non-technical general public generate more sophisticated synthetic media, the capacity for purveyors of problematic content to adapt induces a \newterm{detection dilemma}: as detection practices become more accessible, they become more easily circumvented. This paper describes how a multistakeholder cohort from academia, technology platforms, media entities, and civil society organizations active in synthetic media detection and its socio-technical implications evaluates the detection dilemma. Specifically, we offer an assessment of detection contexts and adversary capacities sourced from the broader, global AI and media integrity community concerned with mitigating the spread of harmful synthetic media. A collection of personas illustrates the intersection between unsophisticated and highly-resourced sponsors of misinformation in the context of their technical capacities. This work concludes that there is no ``best'' approach to navigating the detector dilemma, but derives a set of implications from multistakeholder input to better inform detection process decisions and policies, in practice.

## Problem
Synthetic media is increasingly common and there are now hundreds of artifact detectors with varied capabilities

## What is an "Artifact"?
Inconsistencies with the physical world (e.g., a hat that is part liquid in the picture to the right) or statistical abnormalities.

## Questions
1. When can you trust the detectors?
2. Are detectors reliable over time?
3. Who gets access to the detectors and resultant content evaluation?

## Multistakeholder Process
PAI's AI and Media Integrity Steering Committee is a formal body of stakeholders from journalism, media, civil society, and industry that informed our understanding of the dilemma.

FIRST DRAFT · WITNESS SEE IT FILM IT CHANGE IT · XPRIZE · CBC Radio-Canada · BBC · The New York Times · facebook · Microsoft · aws · Adobe

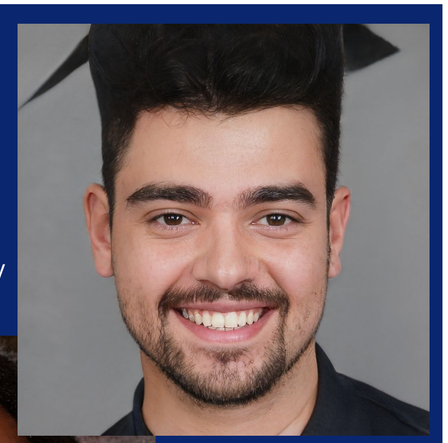
## Play "Artifact Detector"
Can you spot which photos are real and which are synthetic?

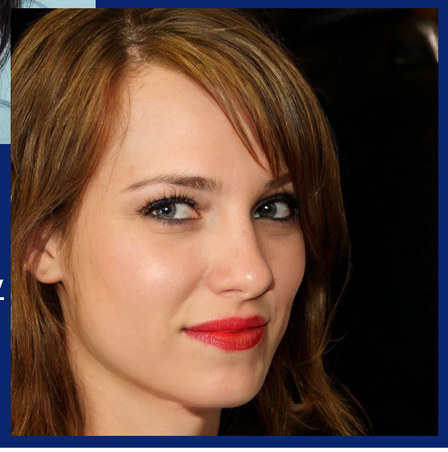
Answer: Aviv is the upper left, Sean is the upper right, Claire is the lower left. The rest are synthetic.

## Personas for Common Ground: Open Access Detector Example

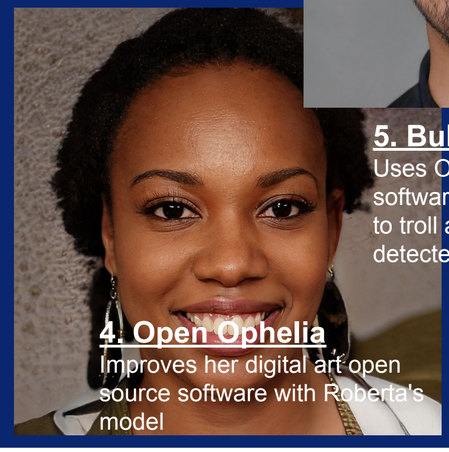
### 1. Researcher Roberta
Creates a new detection model and posts it to GitHub where a social network downloads it and uses it in production

### 2. Bully Bob
Creates a deepfake to troll a classmate and is detected by the social network





### 3. Nation-State Nancy
Retrains her generator to defeat the detector with near certainty



### 4. Open Ophelia
Improves her digital art open source software with Roberta's model

### 5. Bully Bob
Uses Ophelia's digital art software to create a deepfake to troll a classmate and it is not detected by the social network

More on the personas in the paper!

## *The Deepfake Detection Dilemma*
The more accessible detection technology becomes, the more easily it can be circumvented

Many images created with *thispersondoesnotexist.com*