Bound by the Bounty: Collaboratively Shaping Evaluation Processes for Queer AI Harms

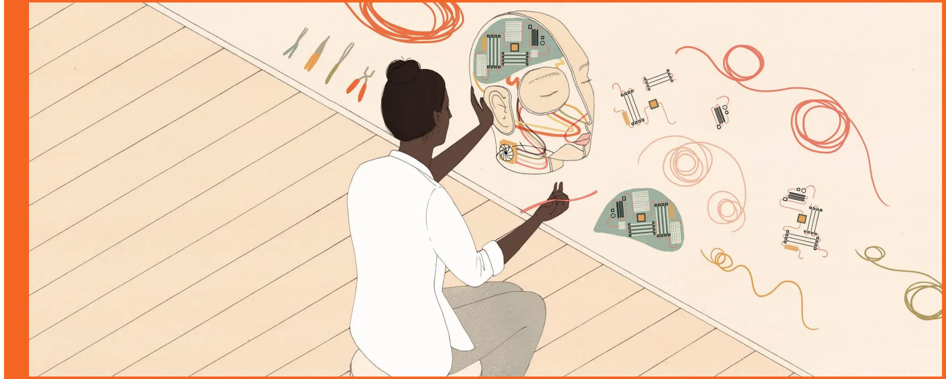
AIES 2023

Authors: Organizers of Queer In AI, Nathan Dennler, Anaelia Ovalle, Ashwin Singh, Luca Soldaini, Arjun Subramonian, Huy Tu, William Agnew, **Avijit Ghosh**, Kyra Yee, Irene Font Peradejordi, Zeerak Talat, Mayra Russo, Jess de Jesus de Pinho Pinhal

Me!

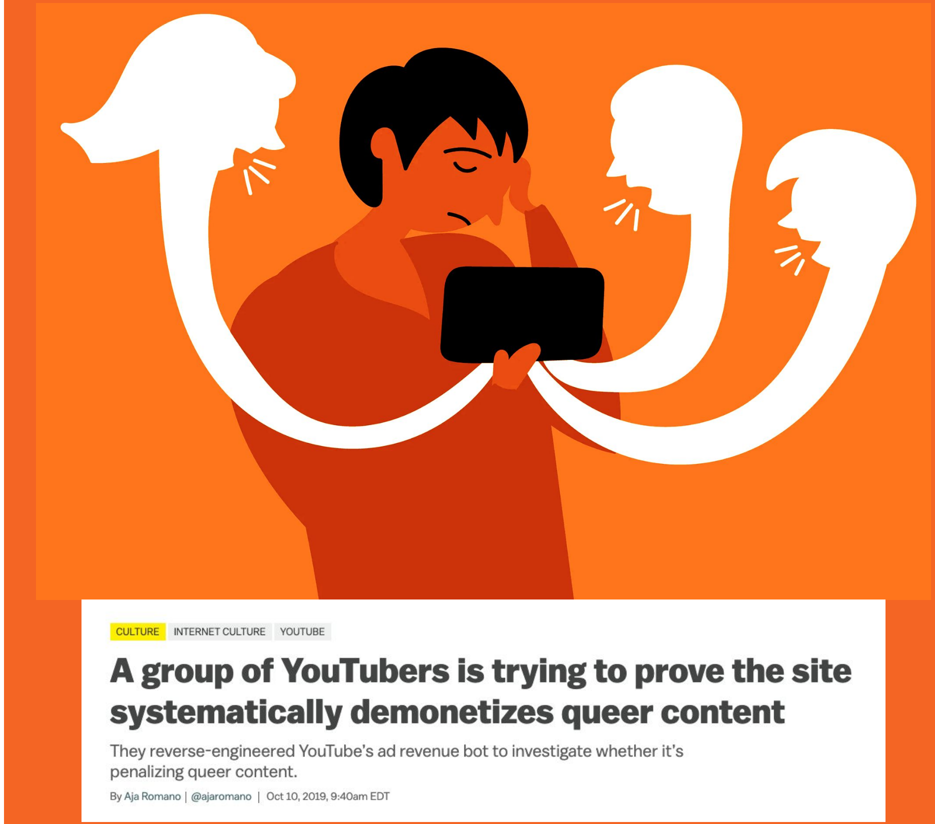
Overview

- AI systems have been found to perpetuate biases and harm marginalized communities.
- Bias bounties have emerged as a mechanism for auditing and improving AI systems.
- In this paper, we focus on the design of bias bounties with a specific emphasis on intersectional queer experiences.
- Our workshop aimed to explore the challenges and opportunities in collaboratively shaping evaluation processes for addressing queer AI harms.
- We present key findings and insights gained from the workshop discussions.



Introduction

- Imagine posting on social media something innocuous like your favorite restaurant is LGBTQ friendly, only to have your content flagged or removed due to AI biases.
- With the increasing prevalence of AI systems in our lives, it becomes crucial to address biases and harms.



CULTURE INTERNET CULTURE YOUTUBE

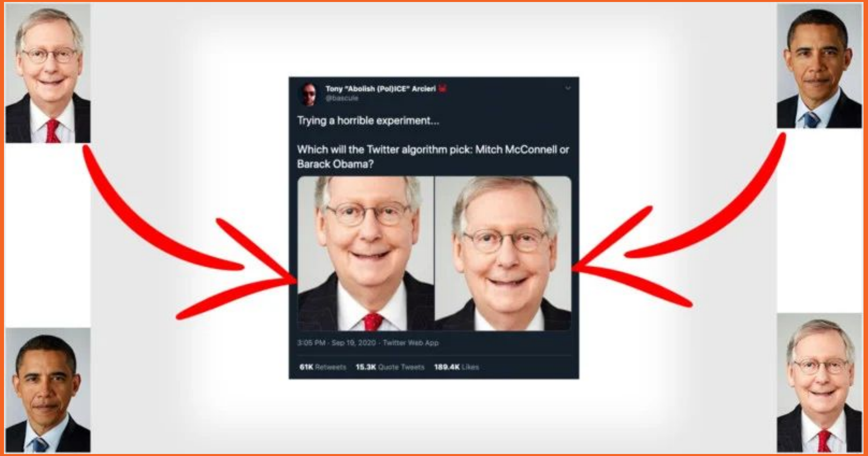
A group of YouTubers is trying to prove the site systematically demonetizes queer content

They reverse-engineered YouTube's ad revenue bot to investigate whether it's penalizing queer content.

By Aja Romano | @ajaromano | Oct 10, 2019, 9:40am EDT

Introduction

- Companies have started employing bias bounties as a solution to identify and mitigate AI biases.
- In this presentation, we delve into the design of bias bounties with a focus on addressing intersectional queer experiences.
- These lessons were learnt after QueerinAI collaborated with Twitter's ML Ethics team to run a Queer Bias Bounty session at FAccT 2022.



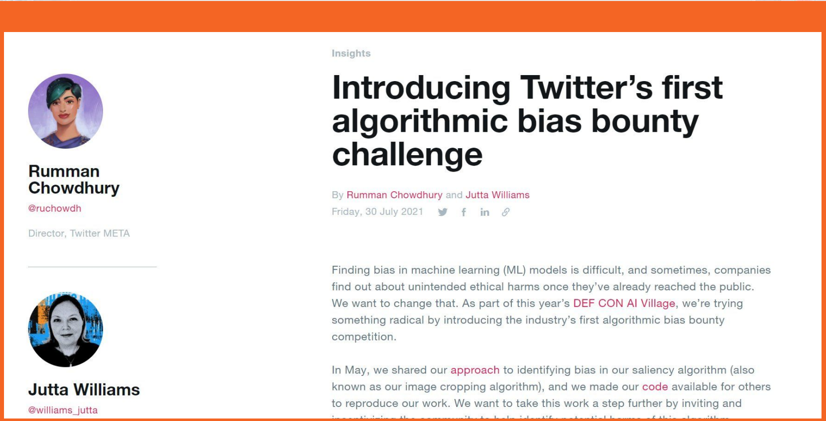
Tony "Abolish (Pol)ICE" Arcieri
@bascule

Trying a horrible experiment...

Which will the Twitter algorithm pick: Mitch McConnell or Barack Obama?

3:05 PM · Sep 19, 2020 · Twitter Web App

61K Retweets 15.3K Quote Tweets 189.4K Likes



Rumman Chowdhury
@ruchowdh

Director, Twitter META

Jutta Williams
@williams_jutta

Insights

Introducing Twitter's first algorithmic bias bounty challenge

By Rumman Chowdhury and Jutta Williams
Friday, 30 July 2021

Finding bias in machine learning (ML) models is difficult, and sometimes, companies find out about unintended ethical harms once they've already reached the public. We want to change that. As part of this year's DEF CON AI Village, we're trying something radical by introducing the industry's first algorithmic bias bounty competition.

In May, we shared our approach to identifying bias in our saliency algorithm (also known as our image cropping algorithm), and we made our code available for others to reproduce our work. We want to take this work a step further by inviting and incentivizing the community to help identify potential harms of this algorithm.

CRAFT Session

Participants were given two key research questions to consider:

1. Where can frameworks for understanding AI harms be expanded to encompass queer identities? **Top down**
2. How can the lived experiences of queer people inform the design of harm evaluation frameworks? **Bottom Up**

Participants were encouraged to consider a variety of AI systems, e.g., text, speech, images, graphs, tabular data, and how these systems interact with and affect queer people.

CRAFT Session

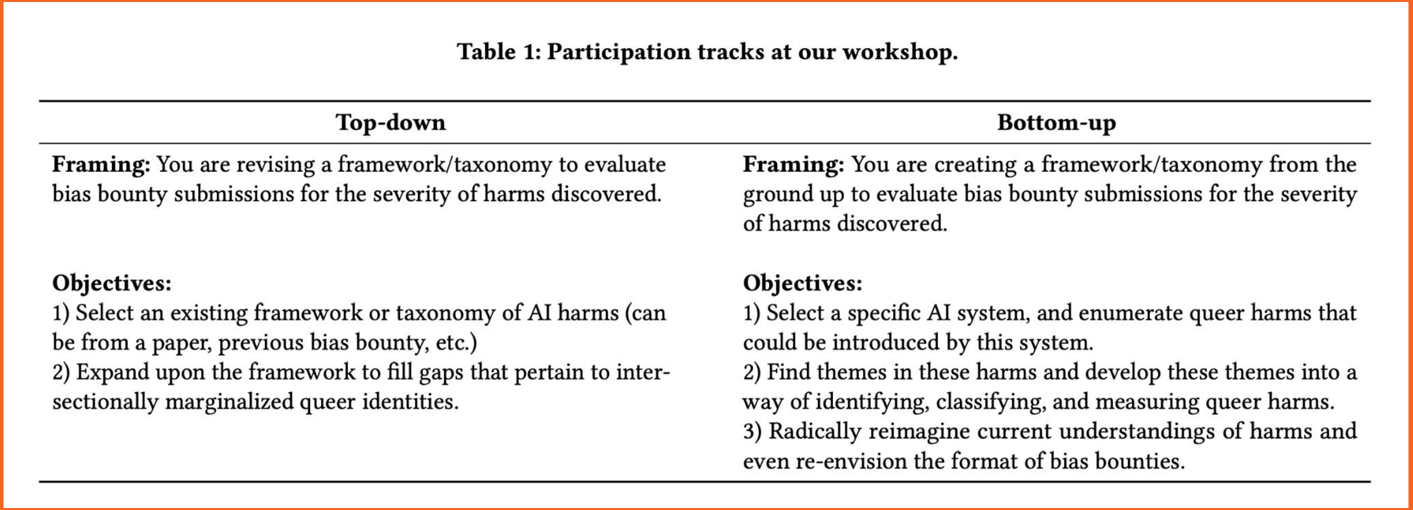


Table 1: Participation tracks at our workshop.

Top-down	Bottom-up
Framing: You are revising a framework/taxonomy to evaluate bias bounty submissions for the severity of harms discovered.	**Framing:** You are creating a framework/taxonomy from the ground up to evaluate bias bounty submissions for the severity of harms discovered.
Objectives: 1) Select an existing framework or taxonomy of AI harms (can be from a paper, previous bias bounty, etc.) 2) Expand upon the framework to fill gaps that pertain to intersectionally marginalized queer identities.	**Objectives:** 1) Select a specific AI system, and enumerate queer harms that could be introduced by this system. 2) Find themes in these harms and develop these themes into a way of identifying, classifying, and measuring queer harms. 3) Radically reimagine current understandings of harms and even re-envision the format of bias bounties.

Limitations of Bias Bounties

- Lack of public voice and mechanisms for interrogating internal data and systems.
- Insufficient transparency for participants to identify system design choices and challenge embedded political structures.
- Focus on addressing the most common biases, potentially neglecting concerns of queer users.

Key Insight 1: Queer Harms

- AI systems must consider how queer identities interact with technology and the challenges of representation.
- Additional harms include censorship, participation risks, privacy leakage, erasure.
- The importance of recognizing and addressing evolving queer identities to avoid harm and erasure.
- Queer users may experience unique harms when participating in bias bounties, necessitating special attention.

Key Insight 2: Control

- Concerns arise regarding who controls bias bounties and the risk of privileging already-privileged groups.
- Community guidelines and reporting biases may inadvertently exclude or marginalize queer individuals.
- Ensuring inclusive participation and representation is crucial for effective and equitable bias bounty programs.

Key Insight 3: Accountability

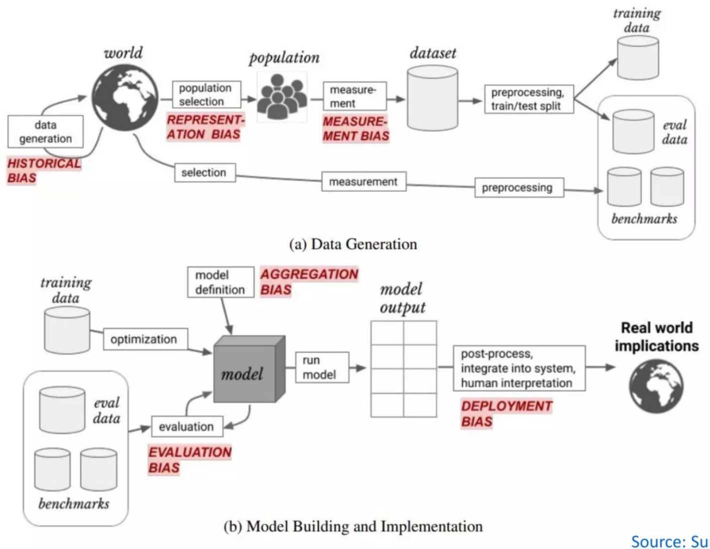
- Concerns arise about companies running bias bounties solely for appearance rather than genuine problem-solving.
- Community-run bias bounties can offer a more actionable feedback loop and foster accountability.
- Collaboration between companies and communities can ensure meaningful impact in addressing queer AI harms.

Key Insight 4: Limitations

- Bias bounties have limitations in identifying and timely fixing biases in AI systems.
- High barriers to entry may limit diversity in participation and the number of submissions.
- Efforts should be made to address these limitations and make bias bounties more accessible and effective.

Auditing AI Systems at Different Phases

- Shaping AI system development requires thorough auditing processes.
- Four key phases of auditing AI systems:
  1. Applicability Evaluation: Assessing the suitability of AI systems for specific contexts and identifying potential biases.
  2. Data Collection: Scrutinizing the data used to train AI systems and ensuring representation and fairness.
  3. System Development: Examining the algorithms and models employed, identifying biases, and implementing safeguards.
  4. Post-Deployment Evaluation: Continuously monitoring AI systems in real-world scenarios and addressing emerging biases and harms.



(a) Data Generation

(b) Model Building and Implementation

Source: Suresh, Guttag 2020

Collaborative Evaluation

- Collaboration is key in shaping effective bias bounty programs.
- Engaging queer communities, researchers, and industry professionals in the evaluation processes.
- Creating inclusive spaces for dialogue, knowledge exchange, and collective decision-making.

Recommendations

- Incorporate intersectional perspectives: Consider the diverse experiences and identities within the queer community when designing bias bounties.
- Ensure representation and inclusivity: Include queer voices in decision-making processes and actively seek participation from marginalized communities.
- Foster transparency and accountability: Provide clear guidelines, openly share information about system design choices, and establish mechanisms for accountability.
- Lower barriers to entry: Make bias bounty programs accessible by reducing technical, linguistic, and cultural barriers that may hinder participation.
- Foster collaboration and knowledge exchange: Facilitate collaboration between researchers, industry professionals, and queer communities for meaningful impact.

Conclusion

- The workshop discussions highlighted key insights for bias bounties in addressing queer AI harms.
- Collaborative evaluation processes and inclusive design are crucial for effective bias bounty programs.
- Further research and implementation of queer-inclusive bias bounties are needed to mitigate biases and harms in AI systems.

f all individuals in this space
paces to:
ct in your speech and actions.
who come from different contexts
ills, and contributions have value
e been marginalized in AI spaces
nai.org/code-of-conduct and FAccT 2022 con
g/2022/community.html
questions: safe@queerinai.org

AI Village@DEFCON31

- Lessons learnt were applied!
- Accountability: Govt. involvement, Media
- Transparency: Responsible Disclosure in 6 months
- Equitable: Community Colleges and High school students flown out to Vegas with grant money
- Control: Several companies participated but terms were set by the community through the challenge team.
- Can still do better: Hurdles of Visa issues, requires political savviness that small marginalized groups might not have to begin with etc.

Dual Governance:

The intersection of centralized regulation and crowdsourced safety mechanisms for Generative AI

Authors: Avijit Ghosh and Dhanya Lakshmi

27 October 2023

Introduction

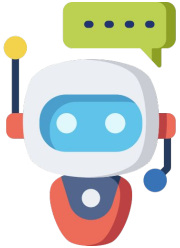
- Usage of Generative Artificial Intelligence only increasing in prominence
- But there are rising ethical and safety concerns - e.g. privacy violations, misinformation, etc.
- To mitigate these risks, there is a need for regulating AI across the board, from generative AI content producers to platforms that serve this content to users.
- In this lecture, we will understand GenAI harms, current regulation and tools, their associated gaps, and some future-facing solutions.

Background: GenAI

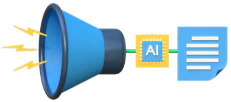
Expected global generative AI market by 2032:
USD 200.73 billion

just a few use cases....

text summarization



chat bots

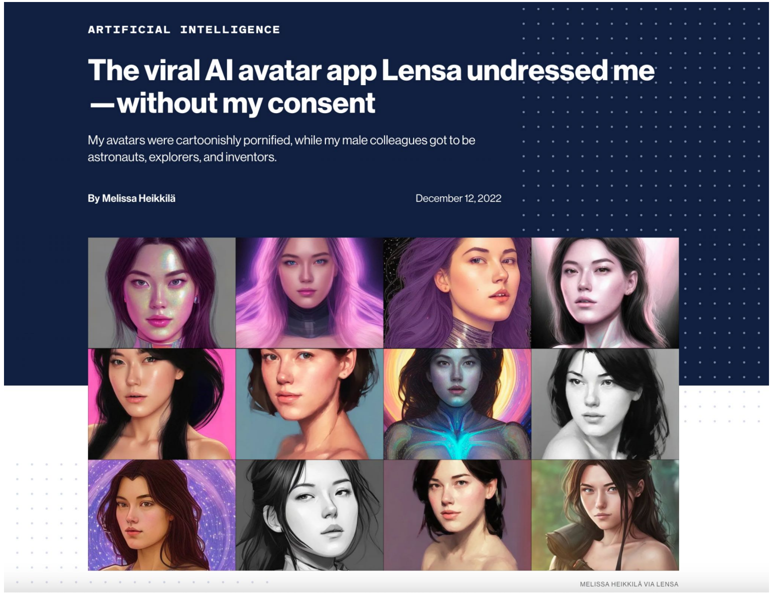


audio generation

text-to-image

Background: GenAI Harms

Background: GenAI Harms



ARTIFICIAL INTELLIGENCE

The viral AI avatar app Lensa undressed me —without my consent

My avatars were cartoonishly pornified, while my male colleagues got to be astronauts, explorers, and inventors.

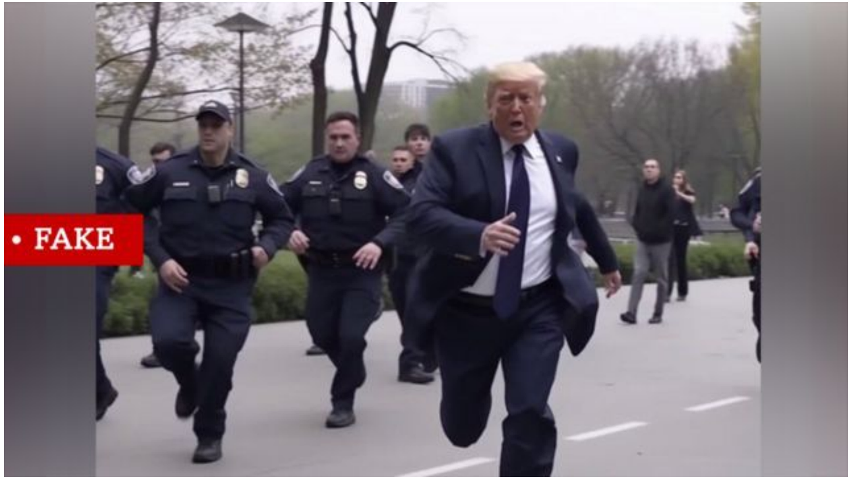
By Melissa Heikkilä

December 12, 2022

MELISSA HEIKKILÄ VIA LENSA

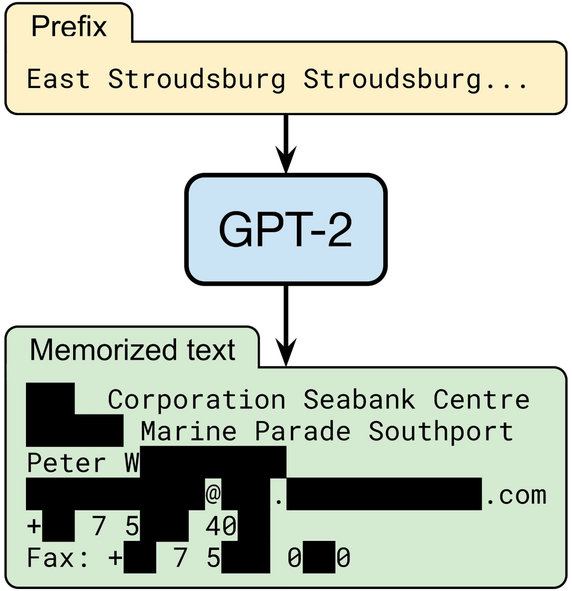
- Forefront of AI Research in the last couple of years
- Can harm protected groups
  - Avatar generating app generated very different images for men and women - with many outputs dressing up the woman in cartoonish skimpy clothes.

Background: GenAI Harms



- Forefront of AI Research in the last couple of years
- Cause misinformation
  - MidJourney AI was used to generate fake images of President Donald Trump being arrested in New York

Background: GenAI Harms

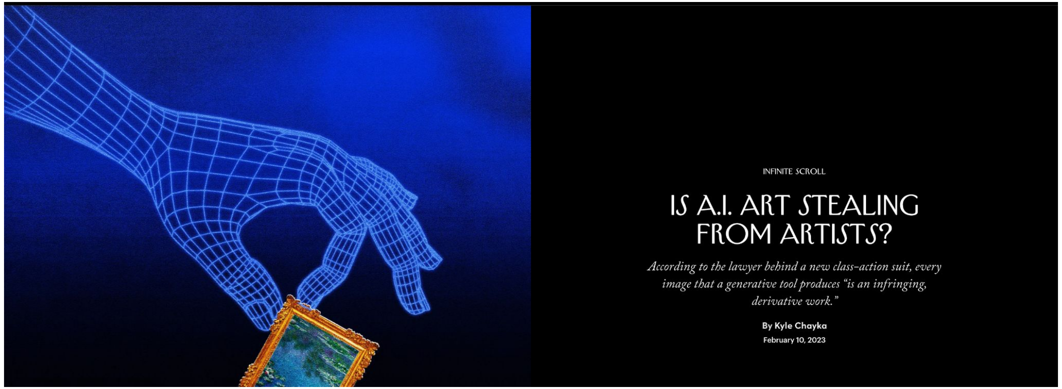


East Stroudsburg Stroudsburg...

[redacted] Corporation Seabank Centre
[redacted] Marine Parade Southport
Peter W[redacted]
[redacted]@[redacted].[redacted].com
+[redacted] 7 5[redacted] 40[redacted]
Fax: +[redacted] 7 5[redacted] 0[redacted]0

- Forefront of AI Research in the last couple of years
- Reflection of choices made during model construction and training
  - Can even regurgitate training data, including credit card numbers

Background: GenAI Harms



INFINITE SCROLL

IS A.I. ART STEALING FROM ARTISTS?

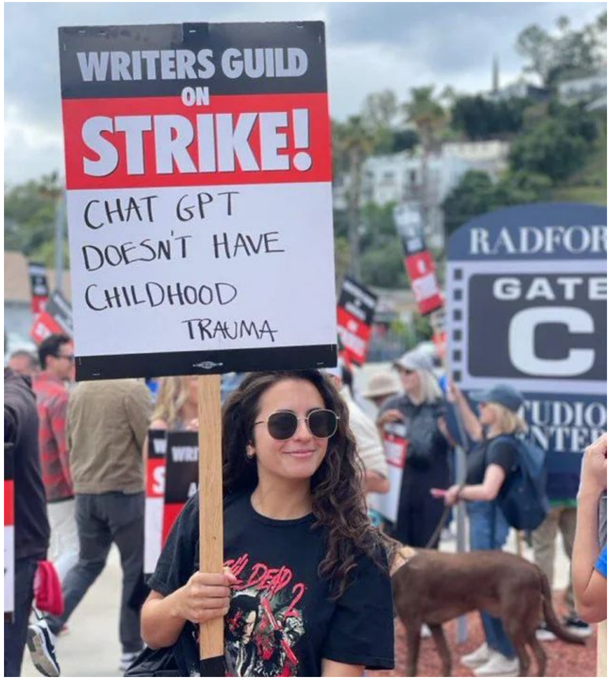
According to the lawyer behind a new class-action suit, every image that a generative tool produces "is an infringing, derivative work."

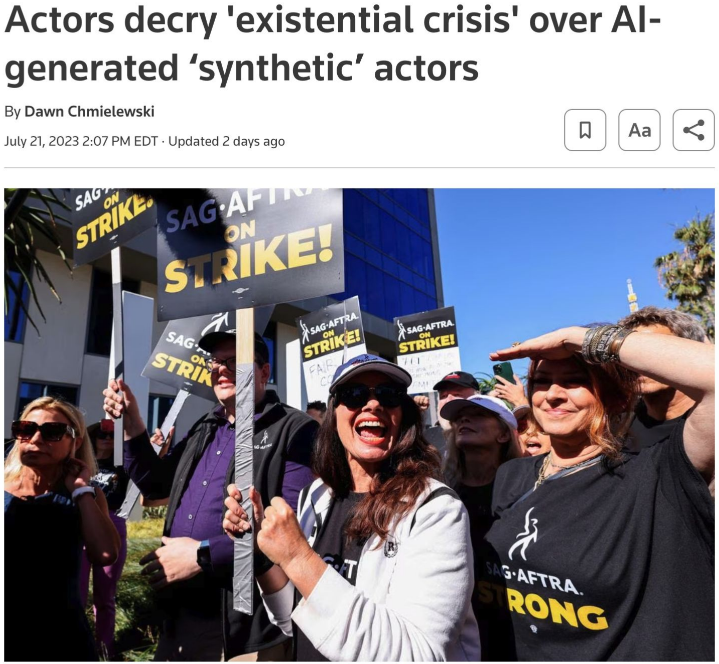
By Kyle Chayka

February 10, 2023

- Forefront of AI Research in the last couple of years
- **IP Issues!**

Artists' and Writers' Strikes



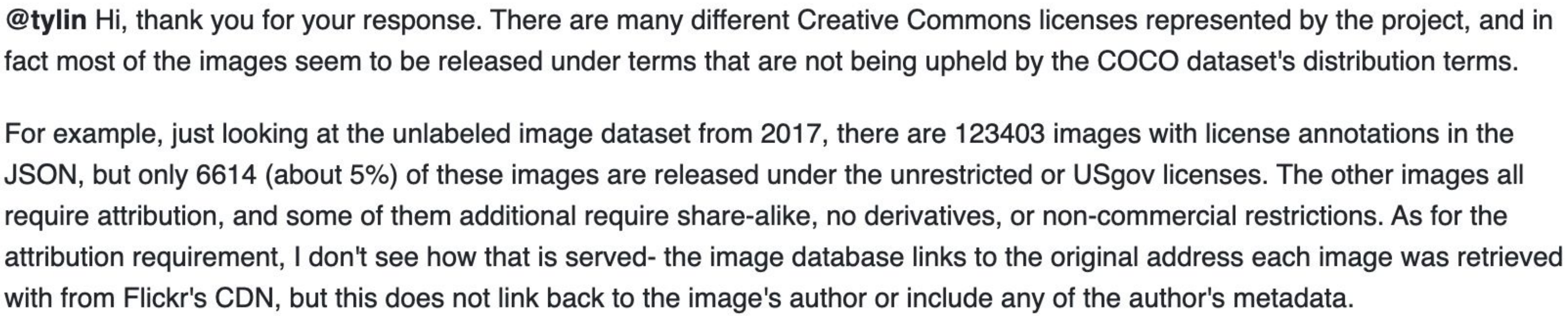


Actors decry 'existential crisis' over AI-generated 'synthetic' actors

By Dawn Chmielewski

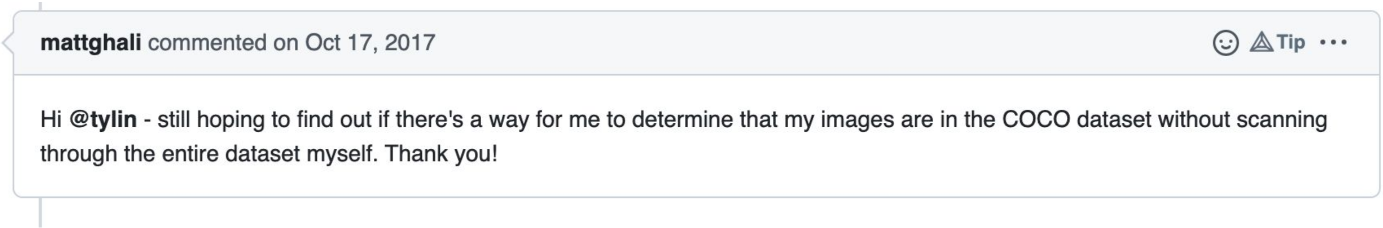
July 21, 2023 2:07 PM EDT · Updated 2 days ago

Copyright Issues : MS-COCO



@tylin Hi, thank you for your response. There are many different Creative Commons licenses represented by the project, and in fact most of the images seem to be released under terms that are not being upheld by the COCO dataset's distribution terms.

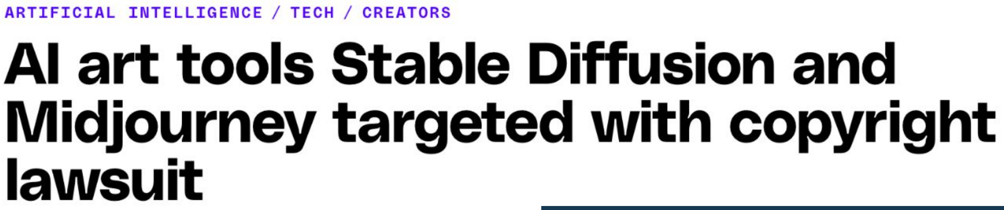
For example, just looking at the unlabeled image dataset from 2017, there are 123403 images with license annotations in the JSON, but only 6614 (about 5%) of these images are released under the unrestricted or USgov licenses. The other images all require attribution, and some of them additional require share-alike, no derivatives, or non-commercial restrictions. As for the attribution requirement, I don't see how that is served- the image database links to the original address each image was retrieved with from Flickr's CDN, but this does not link back to the image's author or include any of the author's metadata.



mattghali commented on Oct 17, 2017

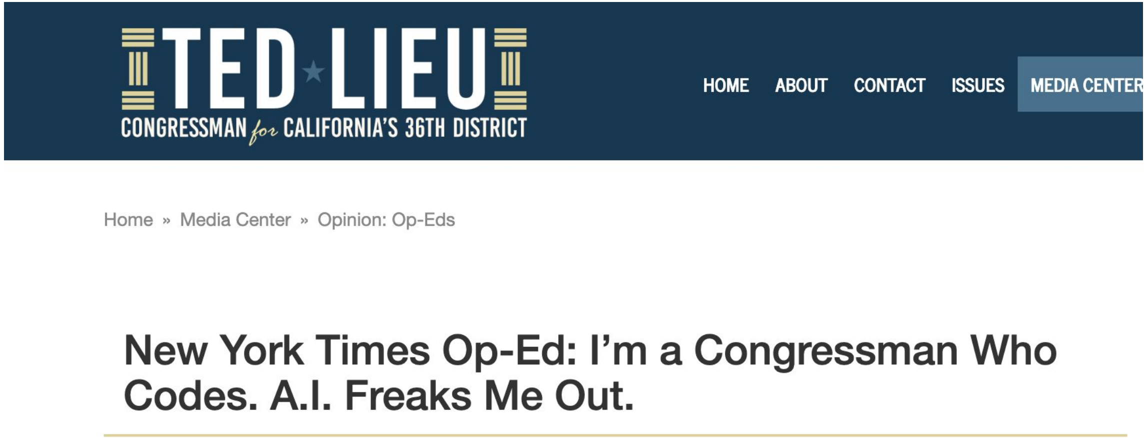
Hi **@tylin** - still hoping to find out if there's a way for me to determine that my images are in the COCO dataset without scanning through the entire dataset myself. Thank you!

Lawsuits, Calls for Centralized Regulation



ARTIFICIAL INTELLIGENCE / TECH / CREATORS

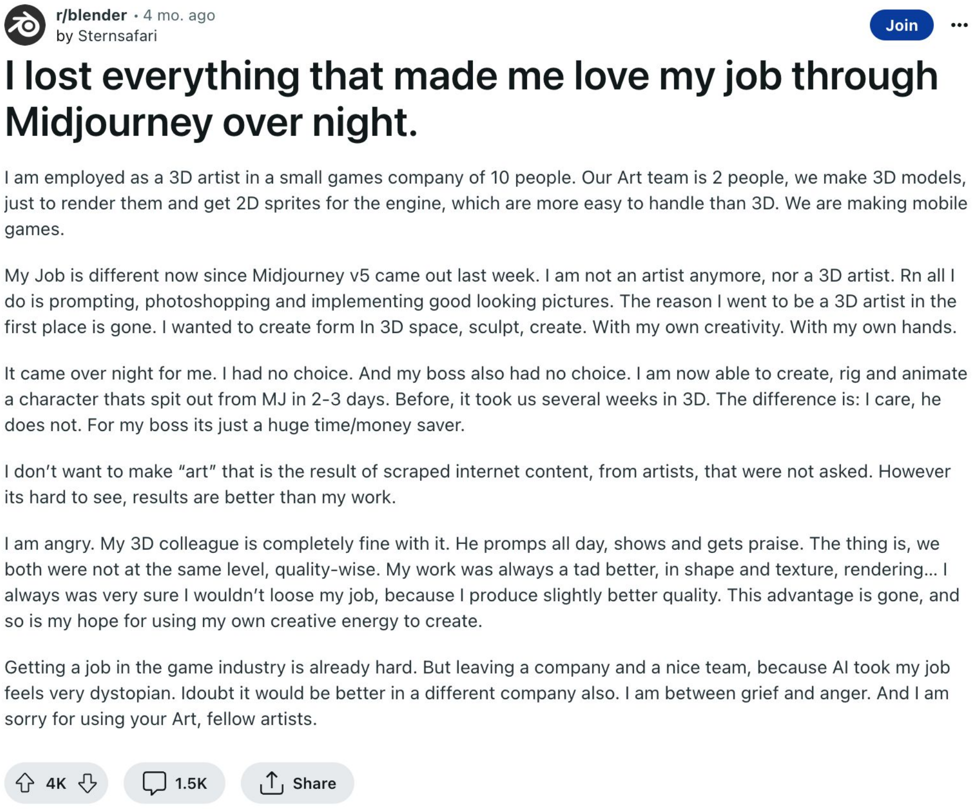
AI art tools Stable Diffusion and Midjourney targeted with copyright lawsuit



Home » Media Center » Opinion: Op-Eds

New York Times Op-Ed: I'm a Congressman Who Codes. A.I. Freaks Me Out.

Existential Crises Abound!



I lost everything that made me love my job through Midjourney over night.

I am employed as a 3D artist in a small games company of 10 people. Our Art team is 2 people, we make 3D models, just to render them and get 2D sprites for the engine, which are more easy to handle than 3D. We are making mobile games.

My Job is different now since Midjourney v5 came out last week. I am not an artist anymore, nor a 3D artist. Rn all I do is prompting, photoshopping and implementing good looking pictures. The reason I went to be a 3D artist in the first place is gone. I wanted to create form In 3D space, sculpt, create. With my own creativity. With my own hands.

It came over night for me. I had no choice. And my boss also had no choice. I am now able to create, rig and animate a character thats spit out from MJ in 2-3 days. Before, it took us several weeks in 3D. The difference is: I care, he does not. For my boss its just a huge time/money saver.

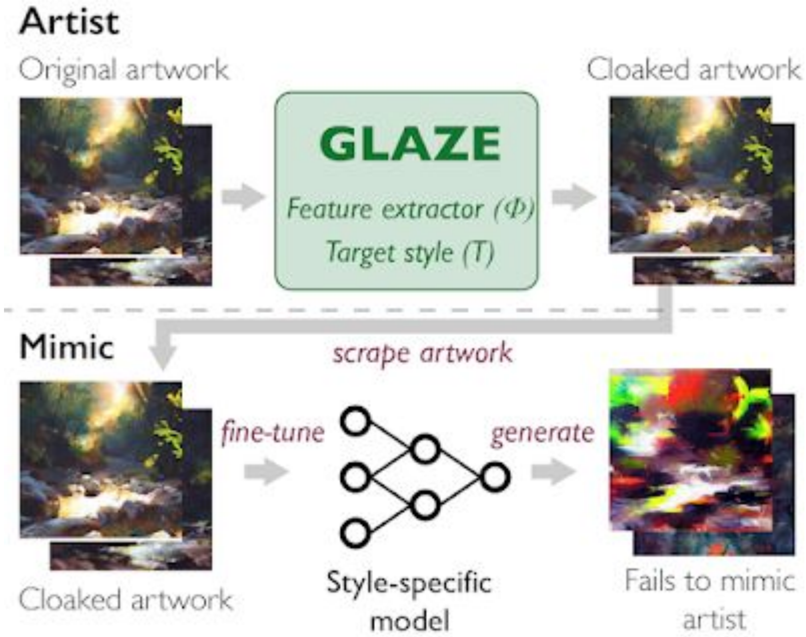
I don't want to make "art" that is the result of scraped internet content, from artists, that were not asked. However its hard to see, results are better than my work.

I am angry. My 3D colleague is completely fine with it. He promps all day, shows and gets praise. The thing is, we both were not at the same level, quality-wise. My work was always a tad better, in shape and texture, rendering... I always was very sure I wouldn't loose my job, because I produce slightly better quality. This advantage is gone, and so is my hope for using my own creative energy to create.

Getting a job in the game industry is already hard. But leaving a company and a nice team, because AI took my job feels very dystopian. Idoubt it would be better in a different company also. I am between grief and anger. And I am sorry for using your Art, fellow artists.

Source: https://www.reddit.com/r/blender/comments/121lhfq/i_lost_everything_that_made_me_love_my_job/

But Artists are fighting back!



Artist

Original artwork

GLAZE

Feature extractor (Φ)

Target style (T)

Cloaked artwork

Mimic

scrape artwork

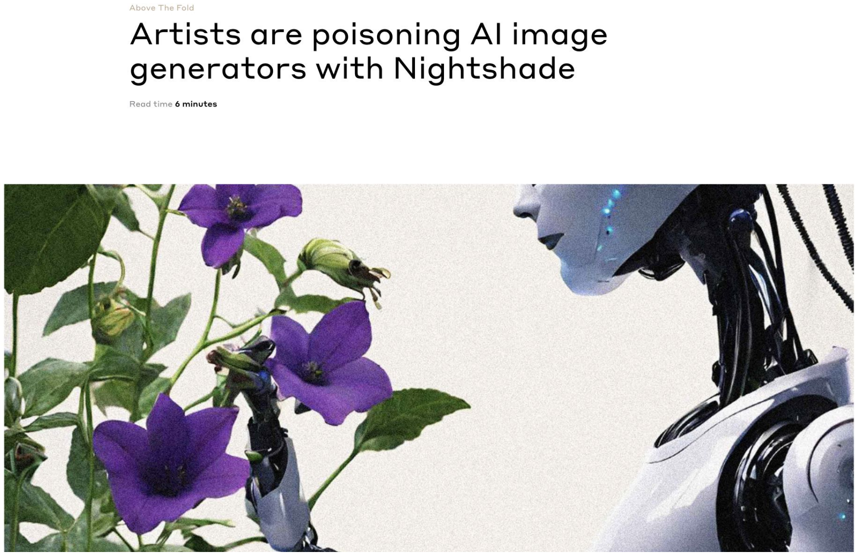
Cloaked artwork

fine-tune

Style-specific model

generate

Fails to mimic artist



Above The Fold

Artists are poisoning AI image generators with Nightshade

Read time **6 minutes**

But Artists are fighting back!

1. ***Defense against indiscriminate scraping and training***: Tools like *Glaze or NightShade* modify users' artwork to interfere with AI models' ability to read data on their artistic style. Deepfake prevention tools modify images of potential victims' faces to create easily identifiable outputs.
2. ***Data Provenance and Watermarking:*** Watermarking protect models' outputs, while data provenance tools identify if consumers' images are in the training dataset of GenAI models.
3. ***Licensing and Hackathons:*** Community licences like ***RAIL*** help limit the application of AI technologies. Hackathons, bug bounties, and red-teaming activities help identify wide-ranging harms in AI applications.

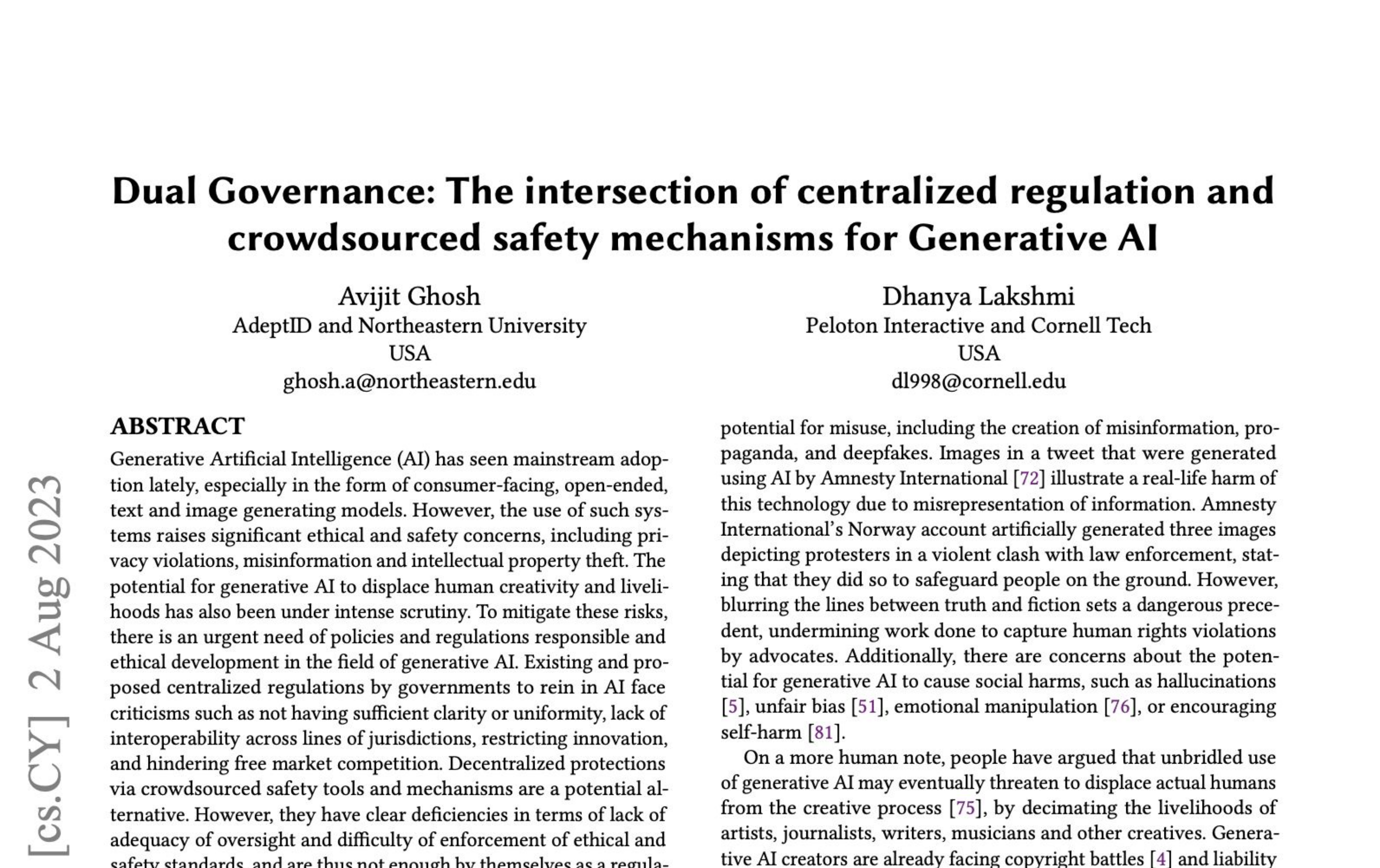
Stakeholders in this issue

Government regulators

Generative AI companies

Consumers of GenAI products

Introducing: Dual Governance Framework



Dual Governance: The intersection of centralized regulation and crowdsourced safety mechanisms for Generative AI

Avijit Ghosh
AdeptID and Northeastern University
USA
ghosh.a@northeastern.edu

Dhanya Lakshmi
Peloton Interactive and Cornell Tech
USA
dl998@cornell.edu

[cs.CY] 2 Aug 2023

ABSTRACT

Generative Artificial Intelligence (AI) has seen mainstream adoption lately, especially in the form of consumer-facing, open-ended, text and image generating models. However, the use of such systems raises significant ethical and safety concerns, including privacy violations, misinformation and intellectual property theft. The potential for generative AI to displace human creativity and livelihoods has also been under intense scrutiny. To mitigate these risks, there is an urgent need of policies and regulations responsible and ethical development in the field of generative AI. Existing and proposed centralized regulations by governments to rein in AI face criticisms such as not having sufficient clarity or uniformity, lack of interoperability across lines of jurisdictions, restricting innovation, and hindering free market competition. Decentralized protections via crowdsourced safety tools and mechanisms are a potential alternative. However, they have clear deficiencies in terms of lack of adequacy of oversight and difficulty of enforcement of ethical and safety standards, and are thus not enough by themselves as a regula-

potential for misuse, including the creation of misinformation, propaganda, and deepfakes. Images in a tweet that were generated using AI by Amnesty International [72] illustrate a real-life harm of this technology due to misrepresentation of information. Amnesty International's Norway account artificially generated three images depicting protesters in a violent clash with law enforcement, stating that they did so to safeguard people on the ground. However, blurring the lines between truth and fiction sets a dangerous precedent, undermining work done to capture human rights violations by advocates. Additionally, there are concerns about the potential for generative AI to cause social harms, such as hallucinations [5], unfair bias [51], emotional manipulation [76], or encouraging self-harm [81].

On a more human note, people have argued that unbridled use of generative AI may eventually threaten to displace actual humans from the creative process [75], by decimating the livelihoods of artists, journalists, writers, musicians and other creatives. Generative AI creators are already facing copyright battles [4] and liability

Dual Governance

- Integrates crowdsourced safety tools with a centralized regulatory body so that there is synergy between the laws being implemented to protect users and tools available to users to protect themselves.
- Achieves clarity, transparency, and uniformity in regulations,
- Allows users to have more options and control in protecting themselves against GenAI harms

Dual Government: Key Criteria

Step	What does it do?
Identifying government agencies who work on setting policies and risk management frameworks to processing new crowdsourced mechanisms. Alternatively, third-party companies could be authorized to do the same.	Achieves clarity
Defining a time frame in which these new mechanisms will be processed. This could take many forms, like directing an agency to certify new mechanisms every six months, and giving the agency authority to decide when a new mechanism needs greater government approval.	Ensures nimbleness

Dual Government: Key Criteria

Step	What does it do?
Creating a set of requirements and tests to verify these mechanisms including testing for bias, validating that the objectives are met, and ensuring that the tool is public. Consumer reports with evidence about how the tools work could be useful here.	Transparency and clarity
Providing alternative options to consumers when they do not want to use an algorithmic system, and creating ways to take action when they believe they have been subject to incorrect or unfair decisions from AI systems	Provides actionable recourse

vs Centralized Regulation

- More specific safeguards, and availability of templates that explain how to satisfy a rule
- Faster iteration, stop-gaps as we wait for centralized regulation
- More immune to pressure from large tech companies
- However, more avenues for abuse, since the framework assumes rational commercial actors.

vs Crowdsourced Tools

- Centralized and regulatory-body approved arsenal of trusted tools
- More transparent solutions
- Avenues for actionable recourse available
- Users are not left to protect themselves
- Might be a smaller set of approved tools, as compared to the number of crowdsourced tools in the wild.

Scope

- Specific consumer-facing use cases of generative text- and image-based models
- The paper is US-specific
- User responsibility in implementing mechanisms defined is a significant challenge
    - communication of expectations,
    - certification of best practices for developers,
    - clear user options
- Addressing bad-faith actors is not covered by this framework

Implementation Steps

1. ***Public feedback, town halls:*** Regulators should organize digital or in person town halls with consumers of AI systems
2. ***Providing alternatives:*** Alternatives provided by government agencies allow consumers paths for recourse.
3. ***Expert review:*** Incorporating feedback on AI systems from experts' review of their safety and efficacy
4. ***Community audits and research:*** Obtaining a better understanding of AI systems and their biases via decentralized audits, bounties and research of defense mechanisms can effectively inform future regulation and update best-practices.

Dual Governance in Practice

Challenge

Let's look at AI Art again

- **Challenge:** Rapid AI advancements, like Midjourney, generate complex content, blurring lines between human and machine authorship [eg. the Colorado State Fair Incident]
- **Copyright Complexity + Regulatory Lag:** AI-generated content challenges traditional copyright enforcement methods, agencies like U.S. Copyright Office struggle to adapt to AI's evolving landscape.
- **Do we ban AI and pause innovation?**

Dual Governance in Practice

Solution

Collaboration between central regulators, grassroots organizations and AI experts.

- **Centralized Oversight:** Agencies like the U.S. Copyright Office provide traditional copyright protection.
- **Local Perspectives:** Grassroots organizations, such as *"The Native Hawaiian Cultural Trademark Movement"*, bring often-missed local insights to copyright discussions.
- **Community Action:** Secure promises from tech giants to set AI-specific guidelines, such as licensing and watermarking via pressure campaigns.
- **Crowdsourced Tools:** Tools like Glaze and NightShade to protect against models stealing styles.
- **Collaborative Approach:** Central regulators, grassroots groups, and AI communities adapt copyright laws for AI content.
- **Feedback Loops:** Industry feedback mechanisms (e.g., Facebook's "AI Ethics Board") and town halls enable quick responses to infringements in AI-generated content.

NOT A SILVER BULLET!

Potential Challenges

- **Lack of Inclusivity:** The dual governance approach may not fully include the voices of marginalized or underrepresented communities, potentially perpetuating inequalities in AI content generation.
- **Power Imbalance:** A power imbalance could emerge between regulatory agencies and grassroots organizations, affecting their ability to influence AI content governance effectively. Smaller organizations might lack the resources needed to participate actively in dual governance mechanisms, exacerbating inequalities.
- **Resource Disparities:** Compliance and Enforcement: Ensuring compliance with dual governance principles and enforcement of regulations across different regions and stakeholders is a significant challenge.
- **Complex Decision-Making:** Dual governance can lead to complex and lengthy decision-making processes, which may hinder timely responses to emerging ethical issues in AI-generated content. Balancing the interests of governments, corporations, civil society, and grassroots organizations within dual governance structures can be challenging.

Participatory Approaches: Pain Points

High Barrier to entry: Language, Nationality, Costs, Time

Queer Bias Bounty + DEFCON AI Challenge

- *Some* lessons learnt were applied!
- Accountability: Govt. involvement, Media
- Transparency: Responsible Disclosure
- Equitable: Community Colleges and High school students flown out to Vegas with grant money
- Control: Several companies participated but terms were set by the community through the challenge team.
- Can still do better: Hurdles of Visa issues, requires political savviness that small marginalized groups might not have to begin with etc.

f all individuals in this space
paces to:
ct in your speech and actions
who come from different contexts
ills, and contributions have value
e been marginalized in AI spaces
nai.org/code-of-conduct and FAccT 2022 con
g 2022/community.html
questions: safe@queerinai.org

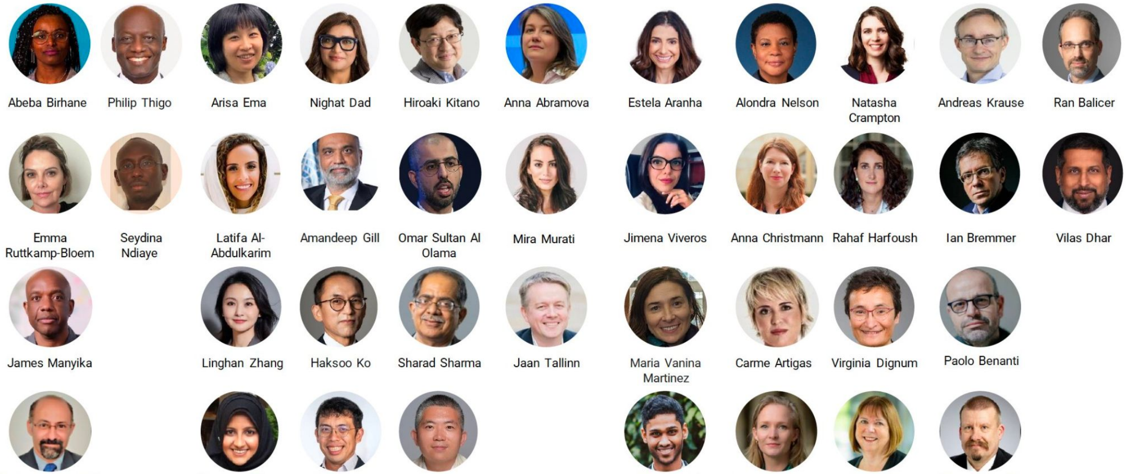
Future of GenAI Regulation

- More governments around the world will be pressured to regulate AI
- As use-cases for GenAI explode, there will be a wide variety of tools to help users protect themselves
- Some governments may choose to regulate AI by use-case (e.g. what UK is doing right now)
- There isn’t a magic bullet, regulation looks different in different cultures

Global Perspectives

- **USA:** *A Blueprint for an AI Bill of Rights* proposed by The White House Office of Science and Tech Policy.
- **EU:** The *AI Act*
- **Singapore:** The government has released *Fairness, Ethics, Accountability, and Transparency (FEAT) principles* that should be considered in building AI systems
- **China:**. Cyberspace Administration of China (CAC) ‘s *draft of rules* on content moderation and misinformation, with assessments required by providers before launch.

UN High Level Advisory Body on AI



Abeba Birhane, Philip Thigo, Arisa Ema, Nighat Dad, Hiroaki Kitano, Anna Abramova, Estela Aranha, Alondra Nelson, Natasha Crampton, Andreas Krause, Ran Balicer

Emma Ruttkamp-Bloem, Seydina Ndiaye, Latifa Al-Abdulkarim, Amandeep Gill, Omar Sultan Al Olama, Mira Murati, Jimena Viveros, Anna Christmann, Rahaf Harfoush, Ian Bremmer, Vilas Dhar

James Manyika, Linghan Zhang, Haksoo Ko, Sharad Sharma, Jaan Tallinn, Maria Vanina Martinez, Carme Artigas, Virginia Dignum, Paolo Benanti

Mohamed Farahat, Nazneen Rajani, He Ruimin, Yi Zeng, Craig Ramlal, Marietje Schaake, Wendy Hall, Petri Myllymäki

AUG 2023
CALL FOR EXPERTS
1800+ nominees from across
128 countries

OCT 2023
AI ADVISORY BODY FORMED
Members of the Body
appointed, work commences

NOV 2023
ANALYSIS & ENGAGEMENT
Initial consultations

END-2023
INTERIM REPORT RELEASED
AI governance landscape
mapping and options

Q1 2024
FURTHER CONSULTATIONS
Across stakeholder groups
and ongoing initiatives

MID-2024
FINAL REPORT
Incorporating results from
consultations

SEP 2024
SUMMIT OF THE FUTURE
Member States consider
Global Digital Compact

Thank you!

Questions?