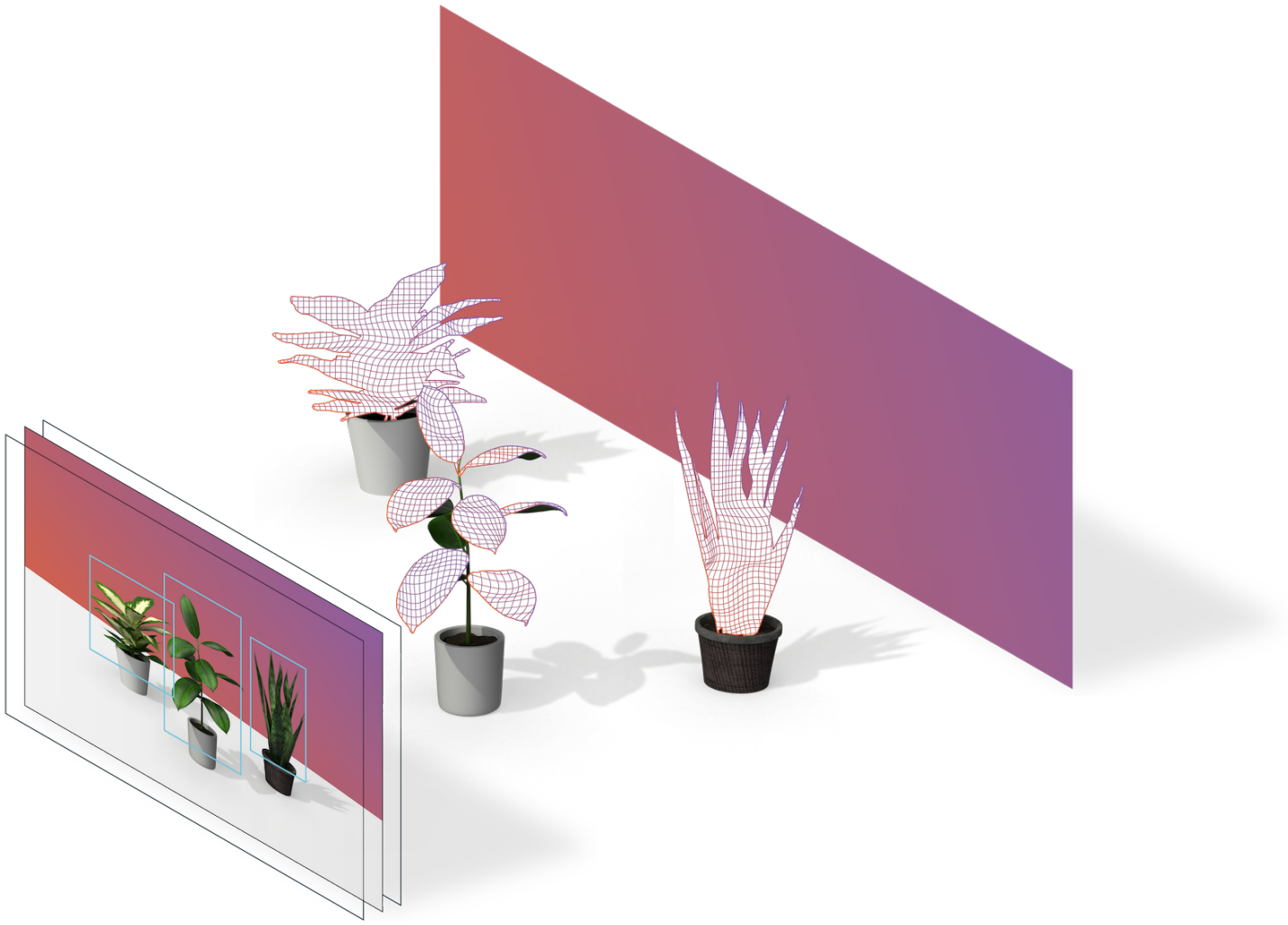
A report from

cdt | Research



EXECUTIVE SUMMARY

# Do You See What I See?

## Capabilities and Limits of Automated Multimedia Content Analysis

Carey Shenkman
Dhanaraj Thakur
Emma Llansó

May 2021


The Center for Democracy & Technology (CDT) is a 25-year-old 501(c)3 nonpartisan nonprofit organization working to promote democratic values by shaping technology policy and architecture. The organisation is headquartered in Washington, D.C. and has a Europe Office in Brussels, Belgium.

---


### CAREY SHENKMAN

Carey Shenkman is an independent consultant and human rights attorney.

### DHANARAJ THAKUR

Dhanaraj Thakur is the Research Director at CDT, where he leads research that advances human rights and civil liberties online.

### EMMA LLANSÓ

Emma Llansó is the Director of CDT's Free Expression Project, where she leads CDT's work to promote laws and policies that support Internet users' free expression rights in the United States, Europe, and around the world.

A report from

cdt | Research

## EXECUTIVE SUMMARY

# Do You See What I See?

## Capabilities and Limits of Automated Multimedia Content Analysis

**Carey Shenkman**
**Dhanaraj Thakur**
**Emma Llansó**

### WITH CONTRIBUTIONS BY

DeVan Hankerson, Hannah Quay-de la Vallee, Samir Jain, and Tim Hoagland.

### ACKNOWLEDGEMENTS

We thank Robin Burke for his feedback on sections of this paper. We also thank the various experts from academia, industry, and civil society that we interviewed and who helped inform the analysis in this paper.

This work is made possible through a grant from the John S. and James L. Knight Foundation.

**Suggested Citation:** Shenkman, C., Thakur, D., Llansó, E. (2021) Do You See What I See? Capabilities and Limits of Automated Multimedia Content Analysis. Center for Democracy & Technology.

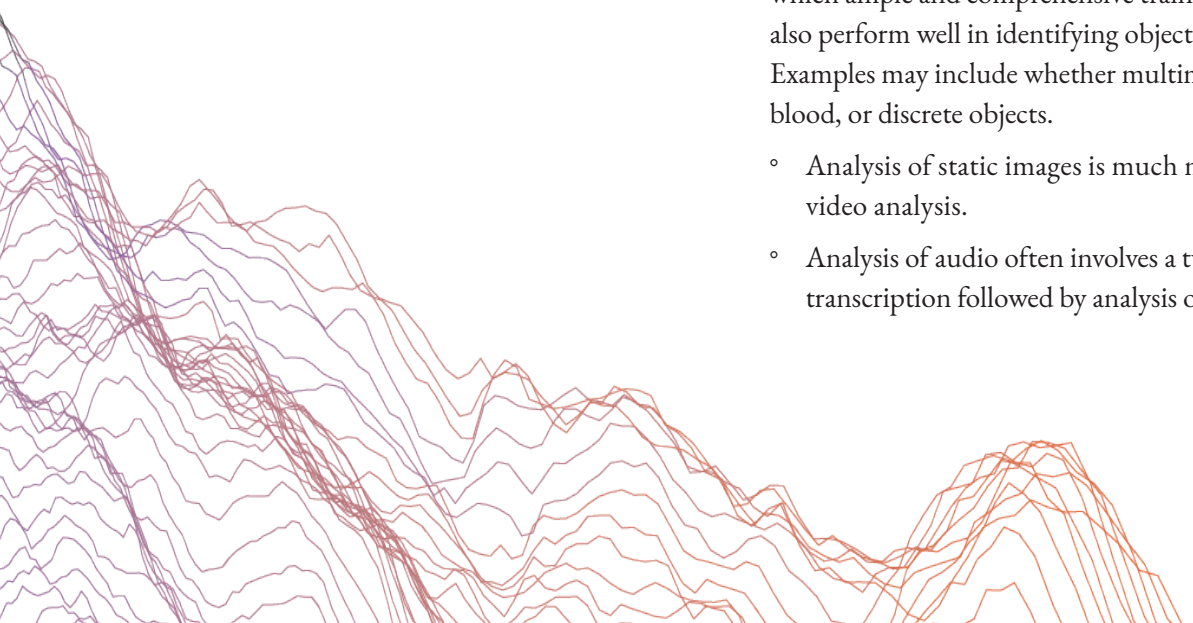
# Executive Summary

The ever-increasing amount of user-generated content online has led, in recent years, to an expansion in research and investment in automated content analysis tools. Scrutiny of automated content analysis has accelerated during the COVID-19 pandemic, as social networking services have placed a greater reliance on these tools due to concerns about health risks to their moderation staff from in-person work. At the same time, there are important policy debates around the world about how to improve content moderation while protecting free expression and privacy. In order to advance these debates, we need to understand the potential role of automated content analysis tools.

This paper explains the capabilities and limitations of tools for analyzing online multimedia content and highlights the potential risks of using these tools at scale without accounting for their limitations. It focuses on two main categories of tools: matching models and computer prediction models. Matching models include cryptographic and perceptual hashing, which compare user-generated content with existing and known content. Predictive models (including computer vision and computer audition) are machine learning techniques that aim to identify characteristics of new or previously unknown content.

These tools are most useful under certain conditions:

- **Matching models** are generally well-suited for analyzing known, existing images, audio, and video, particularly where the same content tends to be circulated repeatedly.
    - Perceptual hashing is almost always better-suited to matching items that feature slight variations, which may occur either naturally or from attempts to circumvent detection.
- **Predictive models** can be well-suited to analyzing content for which ample and comprehensive training data is available. They may also perform well in identifying objective features in multimedia. Examples may include whether multimedia contains clear nudity, blood, or discrete objects.
    - Analysis of static images is much more straightforward than video analysis.
    - Analysis of audio often involves a two-step process of transcription followed by analysis of the transcribed text.



Even in these scenarios, automated multimedia content analysis tools have many limitations. And those limitations become even more evident when the tools are used in more challenging settings. Any applications of these tools should consider at least five potential limitations:

## 1. Robustness

**State-of-the-art automated analysis tools that perform well in controlled settings struggle to analyze new, previously unseen types of multimedia.**

Automated models are repeatedly shown to fail in situations they have never encountered in their design or training. *Robustness* of the tools underlying automated content analysis—or the ability to not be fooled by minor distortions in data—is a constant and unsolved problem. Some challenges for automated analysis are due to natural occurrences (such as a photograph taken at a slightly different angle from a reference photo). But in a social media analysis setting, many challenges are *deliberately* caused by efforts to slip past detection. These can include anything from watermarks, to subtle rotations or blurs, to sophisticated methods such as deepfakes which create synthetic, realistic-seeming videos to harass or spread disinformation. Machine learning models struggle with these cases because circumvention efforts are constantly evolving, and models may be over-optimized for the examples with which they are created or trained. They may not generalize performance well to novel data. This is akin to memorizing answers to specific questions before a test without actually understanding the underlying concepts.

## 2. Data Quality

**Decisions based on automated content analysis risk amplifying biases present in the real world.**

Machine learning algorithms rely on enormous amounts of training data, which can include large databases of photos, audio, and videos. It is well documented that datasets are susceptible to both intended and unintended biases. How specific concepts are represented in images, videos, and audio may be prone to biases on the basis of race, gender, culture, ability, and more. Multimedia sampled randomly from real-world data can likewise propagate real-world biases. For example, existing news coverage of "terrorist propaganda" often perpetuates racial and religious biases. This can lead to problematic asymmetries as to what automated models identify as "terrorist" images. While some methods exist for attempting to mitigate these biases at the machine learning level, they are far from sufficient. Moreover, efforts to "clean" datasets to address some kinds of risks can actually introduce other forms of bias into the training data.

## 3. Lack of Context

### Automated tools perform poorly when tasked with decisions requiring appreciation of context.

While some types of content analysis may be relatively straightforward, the task of understanding user-generated content is typically rife with ambiguity and subjective judgment calls. Certain types of content are easier to classify without *context*—i.e. there may be wider consensus on what constitutes gore, violence, and nudity versus what is sexually suggestive or hateful. And even then, for instance, artistic representations and commentary may contain nudity or violence but be permitted on a given service when depicted in those contexts. The same content shared by one person in a particular setting, such as photos of baked goods, may have entirely different implications in another where those baked goods are a photo selling illicit drugs. Machines are ill-suited to make contextual assessments or apply the nuanced ethical standards that may be necessary for any given decision.

## 4. Measurability

### Generalized claims of accuracy typically do not represent the actual multitude of metrics for model performance.

Real-world impacts of automated analysis decisions may be difficult or impossible to measure without knowing all the content a system fails to properly analyze. For this and other reasons, metrics that convey reliability in the content analysis space, such as "99.9% accuracy," are typically practically meaningless. For example, some forms of harmful content, such as terrorist propaganda, can comprise a very small percentage of multimedia content. An algorithm that merely labels every piece of content "not extreme" could technically be "accurate" at least 99.9% of the time. But it would be right *for entirely the wrong reasons*. Moreover, even if a model predicted the right result 999 out of 1000 times, the one wrong result might have extremely harmful impacts at a scale of millions or billions of pieces of content. Metrics of positive model performance may also be self-selective. They may result from optimization to a specific dataset that is not generalizable to real-world problems. Better measures than "accuracy" are metrics such as *precision* and *recall*, which capture false negative and false positive rates.

## 5. Explainability

### It is difficult to understand the steps automated tools take in reaching conclusions, although there is no "one-size-fits-all" approach to explainability.

State-of-the-art machine learning tools, by default, cannot be "opened up" to get a plain-spoken explanation of why they reached a decision they did. These tools utilize large *neural networks* which may have up to millions or billions of interrelated parameters involved in learning and producing outputs. While the inputs and outputs of these systems may be understood by humans, comprehending the intermediate steps, including how an automated analysis system makes decisions or weighs various features, is a daunting technical task, and these intermediate steps typically do not translate into the kinds of judgments a human would make. Research efforts are being made to promote *explainability*, the

ability to map the operations of machine judgment onto concepts that can be understood by humans. Explainability has important implications for developing trust in these systems and for preventing disparate impacts across various groups, as well as identifying opportunities for redress. At the same time, explainability may vary depending on whether what needs to be known involves the factors in a singular decision, or the structural characteristics of a network as a whole.

While there are many important and useful advances being made in the capabilities of machine learning techniques to analyze content, policymakers, technology companies, journalists, advocates, and other stakeholders need to understand the limitations of these tools. A failure to account for these limitations in the design and implementation of these techniques will lead to detrimental impacts on the rights of people affected by automated analysis and decision making. For example, a tool with limited robustness can be circumvented and fail to identify abusive content. Poor data quality can lead to machine learning models that perpetuate or even exacerbate existing biases in society, and can yield outputs with a disparate impact across different demographics. Insufficient understanding of context can lead to overbroad limits on speech and inaccurate labeling of speakers as violent, criminal, and abusive. Poor measures of the accuracy of automated techniques can lead to a flawed understanding of their effectiveness and use, which can lead to an over-reliance on automation and inhibit the introduction of necessary safeguards. Finally, limited explainability can restrict the options for remedying both individual errors and systematic issues, which is particularly important where these tools are part of key decision-making systems.

Large scale use of the types of automated content analysis tools described in this paper will only amplify their limitations and associated risks. As a result, such tools should seldom be used in isolation; if they are used, it should only be as part of more comprehensive systems that incorporate human review and other opportunities for intervention. Design of such systems requires an accurate understanding of the underlying tools being used and their limitations.

Policymakers must also be versed in the limitations of automated analysis tools to avoid promulgating statutes or regulations based on incorrect assumptions about their capabilities. For example, legislators should not pass laws about content moderation that are premised on the ability of automated analysis tools to perform moderation tasks at scale, and automated content filtering should never be required by law. More generally, policies that do not account for the limitations discussed here risk normalizing an uncritical view of the efficacy of these tools. This can undermine important and needed public dialogue about what problems machine learning or “artificial intelligence” can – and cannot – help us solve.

cdt.org

cdt.org/contact

Center for Democracy & Technology
1401 K Street NW, Suite 200
Washington, D.C. 20005

202-637-9800

@CenDemTech

CENTER FOR
DEMOCRACY
& TECHNOLOGY