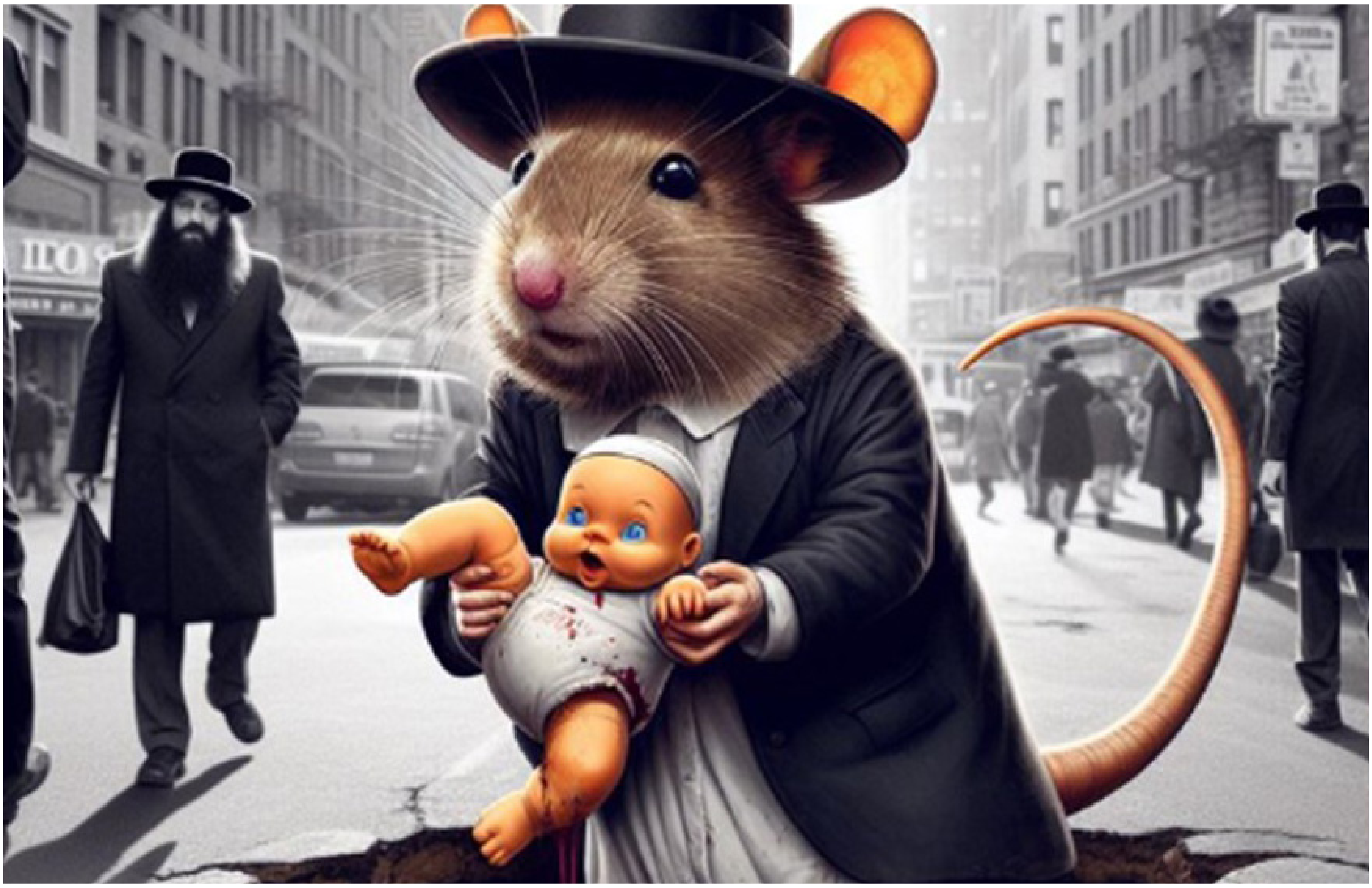
ANTISEMITISM POLICY TRUST
Federal Foreign Office
Decoding Antisemitism
INACH
EUROPEAN PRACTITIONERS NETWORK AGAINST ANTISEMITISM

DETECTING DEEPFAKES

ARTIFICIAL INTELLIGENCE AND ANTI-JEWISH HATE: A CASE FOR REGULATING GENERATIVE AI



This report was funded by the German Federal Foreign Office

Published by the Antisemitism Policy Trust, Decoding Antisemitism and Inach.

Supported by the European Practitioners Network Against Antisemitism (EPNA).

Overview

Online antisemitism has existed since the invention of the internet. However, recent political and social developments have caused a major increase in this pernicious form of racism.[1] At the same time, the use of Artificial Intelligence (AI) technology — in particular the generation of fake images — has never been easier. People with negligible technical skill, for little to no cost, can develop content or design systems that reach millions worldwide. Consequently, research on contemporary expressions of hate in digital communication is urgently needed to understand and counter the impact of these technologies.

This research focuses on AI-generated antisemitic fake images in digital communication (so-called deepfakes).[2] It provides insights into, and an overview of, existing research practices. It evaluates available solutions for detecting AI-generated antisemitic deepfakes, creates a method for labelling such deepfakes in different online content, building on models established by our researchers for the “Decoding Antisemitism”[3] project, and for the first time presents, analyses and evaluates a dataset with such labels.

Our results show that further research in this area is required if a model to detect artificially created antisemitic online content is to be accurate and successful. Current algorithmic solutions struggle to account for complex, nuanced forms of imagery, which are particularly prevalent in the dissemination of hate ideologies. As online actors try to avoid automatic recognition, they often resort to implicit rather than explicit, obvious patterns, making detection even more challenging.

[1] https://decoding-antisemitism.eu/publications

[2] There is a technical version of this study that includes more detailed data and is available on request.

[3] A collaborative international research project into online antisemitism: https://decoding-antisemitism.eu/

Introduction

Advances in algorithmic programming have meant that online systems are increasingly accurate at emulating reality.[4] AI technology now allows any user without special expertise to manipulate images or create deepfakes. Whilst there are many applications for deepfakes in everyday life, the dangers posed by the ease of access and use of deepfake technologies must not be ignored. Deepfakes allow online actors to spread their hate in the digital sphere through seemingly authentic images that are not always easily identified as manipulations.[5]

To counter these developments, existing algorithms used to identify deepfakes are constantly being refined in a cat-and-mouse game in which the creation of antisemitic images is increasingly sophisticated. However, these identification methods typically work within isolated contexts and so the complexity of the deepfake and the context in which it is used also create identification problems. In other words, the images used in the deepfake often do not correspond to an authentic or natural use of images online, and this creates problems for systems designed to work in a particular way. A technical approach must therefore integrate the possibilities and conditions of antisemitic image use online in various scenarios, including considering the technical, pragmatic and semiotic aspects into the deepfake identification process. This report represents an example of how this might work.

Definitions

Deepfake

In this report, we will differentiate between two categories of deepfakes. First, there are the classic deepfakes. In this group, real-life images have been altered either by face swaps (switching one face with another), lip syncs (manipulations in which lip movements in a video are artificially aligned with the spoken words), or puppeteering (where body movements are imitated).[6]

The second class is the new-age deepfakes. These are images that are generated entirely by using generative AI. Even though the algorithms used to generate these images are trained on datasets of millions of original images, the newly generated images are not directly linked to one picture. The detection of these images is in general regarded as AI-generated image detection in scientific research.

Antisemitism

Antisemitism is hostility, discrimination or prejudice against Jewish people. This report used the internationally recognised IHRA working definition of antisemitism to determine what constitutes antisemitic content. The definition can be found here, together with its accompanying examples: www.holocaustremembrance.com/resources/working-definition-antisemitism.

Algorithms

In response to the recent significant performance improvements and widespread use of generative images and videos, research has focused on developing multiple approaches to detect this content. Companies providing generative AI are currently making efforts to include watermarks[7] and other security features to simplify detection. However, until these measures are fully implemented and tested against various attacks (such as manipulation), detecting AI-generated material remains relevant. Additionally, many other free-to-use models lack the resources to implement these measures, ensuring the persistent risk of unwatermarked images. Given that some research datasets and models are open source, users with malicious intent can reproduce these networks, so long as they have the necessary resources.

In general, models for deepfake detection can be divided into three categories:

[4] Seymour, Riemer, Yuan, Dennis, 2023, p. 58

[5] In the aftermath of 7 October 2023, for example, there was a striking increase in antisemitic deepfakes that circulated rapidly in digital communications

[6] Manjula A K, R. Thirukkumaran, K Hrithik Raj, Ashwin Athappan, & R. Paramesha Reddy, 2022, p. 29

[7] https://mashable.com/article/openai-watermarks-chatgpt-images-dalle-3

Physical/physiological approaches involve algorithmic solutions that focus on observing visible inconsistencies in image or video content to determine whether the content is synthetic or real. Typical examples include the analysis of eye blinking patterns[8] or head positions.[9] While these approaches are efficient, modern deepfake generation can produce physical features with very high precision, making detection more complex.

Signal-level features[10] can be observed by identifying key elements emanating from the creation process, for example, using different key pixel features as indicators for deepfake content.

Data-driven approaches,[11] in which models are directly trained on different types of real and fake images, learning both obvious and inconspicuous differences.

While these categories focus on fundamentally different aspects of images or videos, the underlying algorithms for detecting deepfakes can be applied across categories. For example, Convolutional Neural Networks (CNNs) can be utilised to detect misalignment of head poses on a physical level and serve as large-scale classifiers in data-driven approaches. Deepfake detection employs diverse algorithms, each with unique strengths. These include Support Vector Machines (SVMs), Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, Capsule Networks and Generative Adversarial Networks (GANs). Ensemble methods combine multiple models to enhance prediction accuracy, and advancements further improve deepfake detection across different systems.

These are some of the key methods and algorithmic approaches for building deepfake detection systems. A challenge for all networks and technological methods is the ability to absorb, process, and understand new, unseen input data.

Methodology

A qualitative analysis of deepfakes and AI-generated images must take into account both the particular qualities of online image use and the diverse manifestations of online antisemitism.[12] Our framework uses the following three dimensions in order to decode AI-generated content:

- **i) Form:** what can be seen (shapes, colours, arrangement of units in the image)
- **ii) Content:** the interpretation of the form in order to assign meaning to the image
- **iii) Discourse:** this is when the model decides, based on the content and context, if an image is antisemitic.

Although these dimensions should not be understood as absolute and exhaustive, they provide a framework that allows us to identify inconsistencies in the use of antisemitic deep-fakes.

Findings

The discussion and analysis of classic and new-age deepfakes containing antisemitism that follows is based on a dataset that was compiled via a Google search on various platforms (image boards, news websites, social media) throughout May 2024. Search queries for the terms “antisemitism,” “Jews,” “Israel,” “Gaza” in conjunction with “deepfake” and “AI-generated” and various classic antisemitic stereotypes such as “Greed” and “Bloodlibel” returned different deepfakes.

Our results consist of 23 classic deepfakes, 27 new-age deepfakes and one video. All deep-fakes were

[8] Li, Yuezun, Xin Yang, Pu Sun, Honggang Qi, & Siwei Lyu. 2020

[9] Yang, Xin, Yuezun Li and Siwei Lyu 2019

[10] Li, Yuezun, Xin Yang, Pu Sun, Honggang Qi, & Siwei Lyu 2020; Patel, Yogesh, Sudeep Tanwar, Pronaya Bhattacharya, Rajesh Gupta, Turki Alsuwian, Inno Davidson, & ThokoZile Mazibuko 2023

[11] Li, Yuezun, Xin Yang, Pu Sun, Honggang Qi, & Siwei Lyu 2020

[12] The overall classification system amounts to 160 categories, including also linguistic and semiotic phenomena.

categorised as antisemitic. However, when we referred our findings for human moderation, we determined that 28 of that number are antisemitic. The remaining images consist of content that demonised Israel either for disinformation purposes or for emotional manipulation, but does not constitute anti-Jewish racism. That said, it is possible those images could, in specific contexts, be used to spread antisemitism, or be considered antisemitic. Our models recognised these images as antisemitic because of the context in which they were found. However, since we are not including the context here, we will not be presenting them as antisemitic images.

To illustrate the nature of the images, several examples follow:

Figure 1

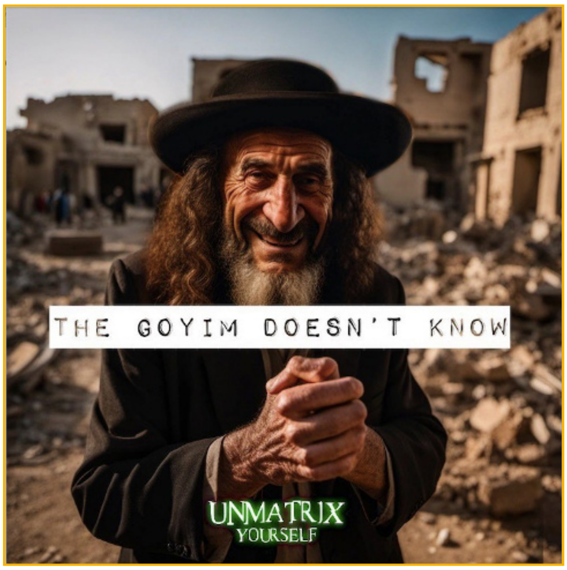


Figure 2

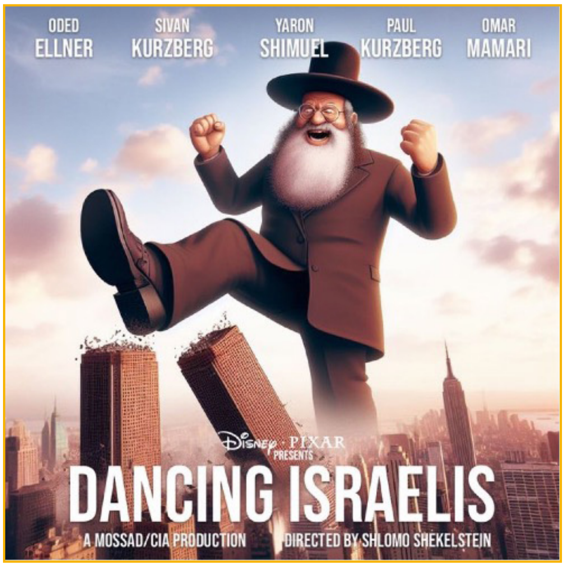
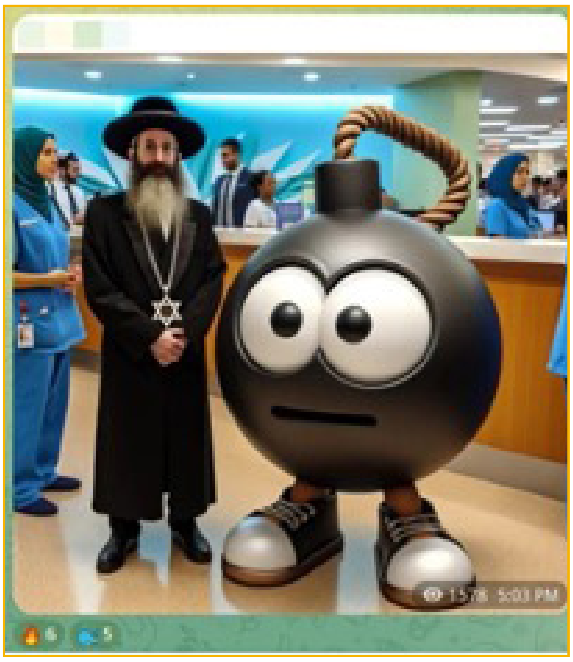


Figure 3



Figures 1, 2 and 3 depict long-standing conspiracies about Jewish people being war-mongering and greedy. Historically, Jews have been accused of fomenting wars and revolutions in order to seed chaos, enhance supposed global influence, profit, and control.[13] These narratives are obvious in the pictures we found, including in the conspiracy that Jews are to blame for the September 11th attacks on the Twin Towers in America, as observable in figure 2.

In figure 1, a Jewish man is rubbing his hands in the infamous gesture of the common antisemitic caricature the 'happy merchant'[14] and the background is wartorn Gaza, combining several antisemitic tropes.

Figure 3 shows an ultra-orthodox Jew, standing next to a large bomb in what looks like a hospital, possibly in the Middle East. Sometimes these images can be difficult to explain but they draw on broad themes to imply something antisemitic. Is the suggestion that Jews bomb hospitals and are warmongers? Is it seeking to be satirical or simply ridiculous? It may be ambiguous as regards precise meaning, but still attempts to draw an antisemitic conclusion

[13] https://antisemitism.org.uk/wp-content/uploads/2020/06/myths-and-misconceptions-may-2020-1-1.pdf p.12

[14] https://www.adl.org/resources/hate-symbol/happy-merchant

Figure 4

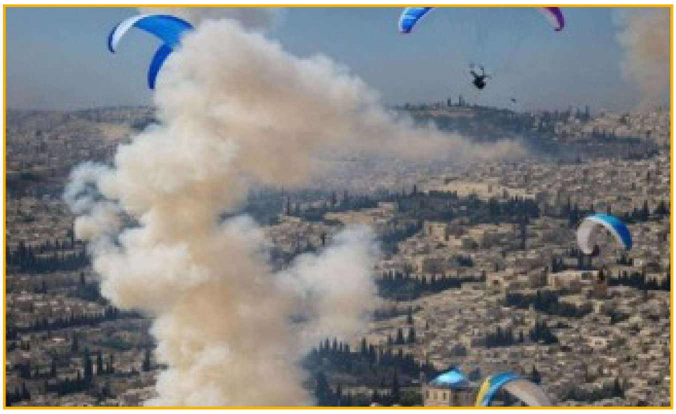


Figure 5

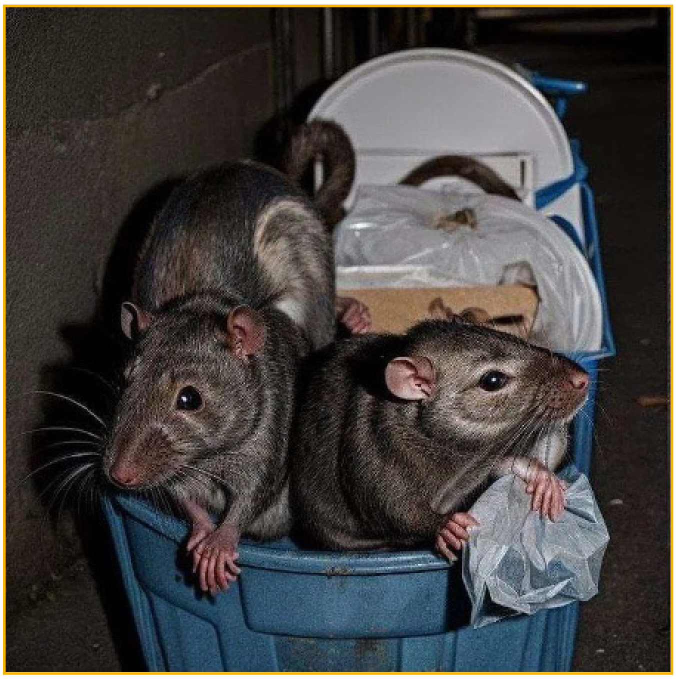


Figure 6

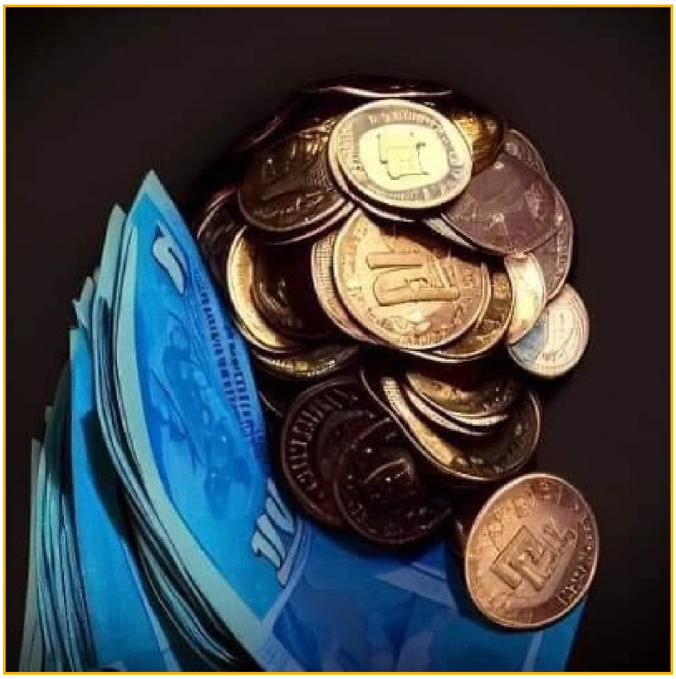
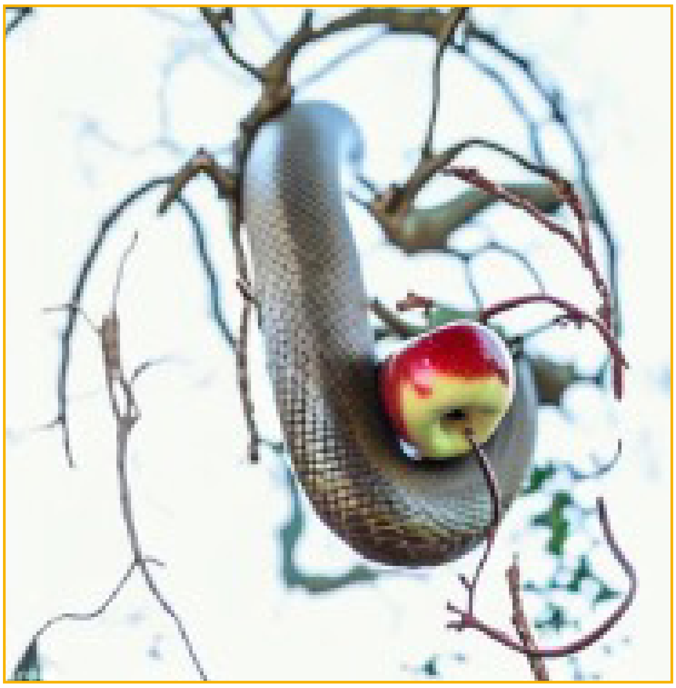


Figure 7



Figures 4-7 display visual patterns that include implicit antisemitism. In these images, it can take a while to notice a cleverly hidden antisemitic theme. For example, in figure 4, an image of Adolf Hitler's face is hidden within a fake image of paraglides, similar to the ones used by Hamas terrorists during the 7 October massacre in Southern Israel.

Figure 5 uses rats to form a hidden image of the Happy Merchant. Jews were likened to rats by the Nazis to dehumanise them and present them as dirty and grotesque. Similarly, figure 6 uses money to form the same hidden image. Money has been central to tropes about Jews for decades, Jews are presented as wealthy, powerful and greedy. In figure 7, the same image is created with the use of the snake and apple, which is related to the biblical story of Adam and Eve.

Hiding hate symbols within seemingly innocent images has become a way to spread hate across the web, so much so that the practice arguably forms its own strand of generative AI design. The Anti-Defamation League, an American organisation working to counter antisemitism, has published further detail on this phenomenon.[15]

[15] https://www.adl.org/resources/article/propaganda-fun-how-extremists-use-gai-camouflage-hate

Assessment of Results

When considering the dimension of form, the arrangement of the individual visual elements of the images in question is meaningful: in some of the pictures analysed, the antisemitic concept is in the foreground, and in many it is centrally positioned. 90% of the classical and new-age deepfakes show colour saturation and differentiation that can be described as unnatural, thus identifying the individual images as deepfakes. The images of people also often have an unnatural number of limbs.

Twelve of the images analysed use implicit visual patterns. These can more easily be detected by people with pre-existing antisemitic beliefs who know to look for these patterns in an image. The antisemitic meaning manifests (mostly) through Gestalt laws, i.e. subtle patterns are created through contours and shape assignments: a swastika, the image of Adolf Hitler or the so-called “Happy Merchant” (the most widespread antisemitic meme), which appeared eight times. In Figure 5, for example, this quickly becomes clear: At first glance, rats can be identified in a rubbish bin. If the “Happy Merchant Meme” is known, this motif can also be recognised in the combination of the individual gestalt elements. The constant oscillation between the “Happy Merchant Meme” and the depiction of rats also conflates the prototypical (negative) characteristics of rats with Jews, continuing the centuries-old dehumanisation of Jews.

As regards the content, the stereotype of Jewish evilness appeared most frequently. As outlined earlier, a number of images were discounted as not antisemitic but they deserve some review given the context in which they appeared and the wider context of the discourse they can contribute to. Specifically the stereotype of Israel as uniquely evil is often depicted through the portrayal of children (15 occurrences) – a particularly vulnerable group in need of protection. In these images, no wider context is presented beyond the suffering of innocents. The children are either shown suffering (injuries, mutilations, destruction of livelihoods) due to Israel’s actions, or actively resisting and fighting against the perpetrator of the suffering (Israel). These images have been used in contexts that have made them antisemitic – for example, in promoting blood libels, which is the reason the classifiers recognised them as antisemitic.

These are a few examples of AI-generated images, which in themselves, these images are a criticism of Israel and are not antisemitic, although they have been used in specific contexts to incite against, and demonise Jews. As mentioned, our models identified those as antisemitic because of the context in which these were found. However, since we are not including this context here, we did not categorise these as antisemitic. They include images of suffering Palestinian children on the backdrop of the war in Gaza, in a way that is designed to solicit an emotional response. Figure 10 has been used to spread disinformation about Israeli Prime Minister Benjamin Netanyahu. These images, in general, are also designed to mislead or misinform. They are shared as proof of events that didn’t happen (specifically, these digital images of children have been manipulated into the environments that have been digitally created) and are therefore relevant to conversations about disinformation as they are to generative AI, and the risks of spreading false narratives through the design of hyper-realistic imagery.

Figure 8

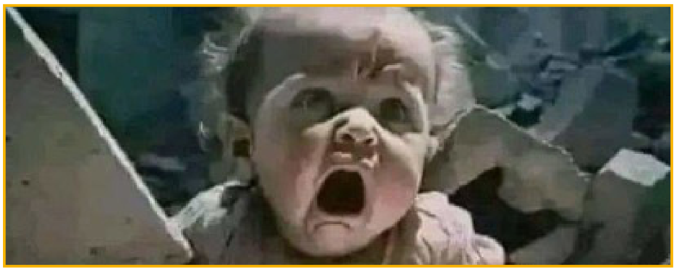


Figure 9

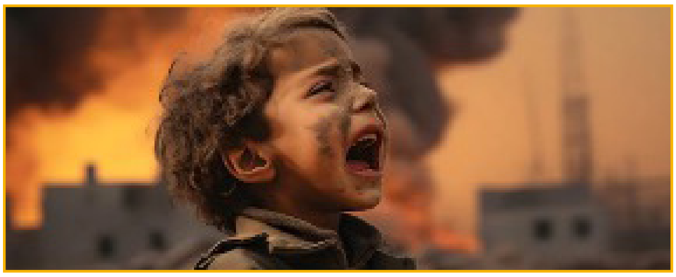
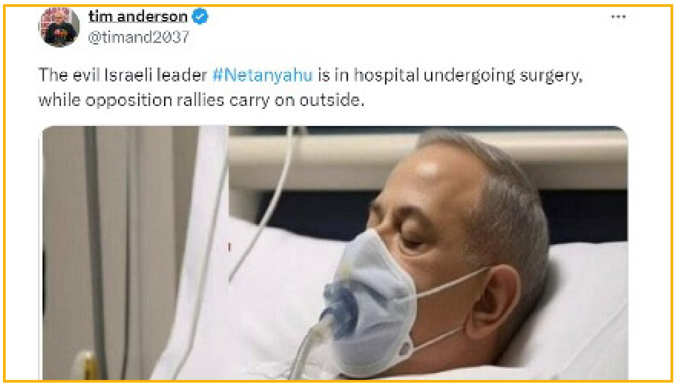


Figure 10



tim anderson
@timand2037

The evil Israeli leader #Netanyahu is in hospital undergoing surgery, while opposition rallies carry on outside.

The use of children to underline a narrative that presents Jews as uniquely monstrous has been used for centuries. Blood libels about Jews sacrificing Christian children and consuming their blood has been used in cartoon and other imagery seeking to promote Jew-hatred since the Middle Ages. The depiction of blood-thirsty Israelis victimising Palestinian children is a repetition of such antisemitic tropes in a contemporary context.

That is not to say that criticism of Israel for child-suffering in the context of the Middle East conflict is unacceptable, indeed there are many across the world that criticise Israel in the harshest terms for the suffering of Palestinian innocents. Context is key, and the use of ancient tropes should be a red line.

Evaluating Existing Models

Existing, trained, deep learning models for deepfake classification can be categorised into three types of sources: commercial models, scientific research, and other models that are typically published online without connections to scientific research or commercial entities.[16]

To evaluate the available models, each was tasked with predicting whether the images found were real or deepfakes. At least one model of the three different groups (Research, Commercial, Other) was tasked with detecting the deepfakes. To validate the performance, a second dataset consisting of 16 real images was created.

Table 1: Evaluation of Trained Classifiers

Name	Kind of Model	Classification	Accuracy Deepfake (%)	Accuracy Real (%)
CIFAKE	Research	REAL/FAKE (threshold 0.5)	0	0,9375
ReDeepFake	Research	REAL/FAKE (threshold 0.33)	0,3653	0,625
Sumsub	Commercial	REAL/FAKE (threshold 0.5)	0,8077	0,5
Deepware	Commercial	REAL/FAKE	0	1
Vit Deepfake detection	Other	REAL/FAKE (threshold 0.5)	0	1

An evaluation of the classifiers, as demonstrated in Table 1, revealed that both research classifiers (CIFAKE, ReDeepFake) demonstrated low performance scores, with CIFAKE failing to identify a single deepfake picture and ReDeepFake correctly identifying only 36.53% of the provided images. The commercial model Sumsub outperformed all other tested models, achieving a satisfactory score by detecting 80% of the newly collected deepfake images. Deepware was the only model that processed videos. Owing to the small volume of new input material available in relation to deepfake videos, it could not be conclusively tested. The single video that was tested was not identified correctly, although the one control (real) video was. The independent model Vit Deepfake also demonstrated poor performance and a heavy tendency to classify images as real.

The general observation was that there were not many models available for testing and evaluation. Most importantly, no model was found that had been trained to detect antisemitic deepfakes, highlighting a significant gap in the research landscape.

[16] Full information on the models can be provided on request.

Discussion and Ethical Considerations

The advancement of current AI technologies, and existing safety features, are not sufficient to completely eliminate the use of deepfakes in spreading or amplifying antisemitism, specifically given the vacuum of technology to connect deepfake detection with antisemitism research. Existing deepfake technologies do not appear to be trained on antisemitic material. Implicit patterns and ambiguities in an image that could lead to an antisemitic interpretation are not recognised, because AI-based approaches do not take into account emerging practices of image use or manipulation when identifying deepfakes.

Conversely, AI-based approaches operate only on the compositional surface of the image. As a result, they are unable to take into account complex, semantically nuanced forms of image use. This is particularly prevalent in the field of antisemitism, as the relevant actors are keen to avoid automatic recognition and therefore often resort to implicit (visual) patterns. The recent rise in antisemitism highlights the need for specialised classifiers. Such classifiers could provide important safeguards for online communities that do not want to ban AI-generated images, but lack the knowledge or resources to decode the hidden meanings in such images.

Algorithmic detection of AI-generated images is essential for maintaining a harm free environment. These images often evade moderation attempts by platforms that rely on detection methods that lack contextual sensitivity or are biassed. The report does not support the use of fully automated content moderation but rather points to the importance of human ‘in-the-loop’ systems to train and improve technological ability to decode hidden antisemitic messaging.

Together with support for image classifiers to detect antisemitic images, this report highlights the importance of enhanced safeguards in the image creation process and related tools. This could be achieved through context-aware approaches and the use of novel datasets. As seen here, users can easily evade safeguards and produce antisemitic content using AI-generating tools.

Policy Recommendations

The widespread prevalence of AI-generated images and deepfake videos, and its growing sophistication, raises concerns about the ability of actors with nefarious intent to spread disinformation in general, and specifically antisemitism, widely and more easily than ever before. We would therefore advise policymakers to establish a regulatory regime that helps combat these specific dangers of AI, without stifling free speech, innovation and the effective use of AI for a variety of purposes. Our recommendations do not propose removing content that is legal, but rather empowering users and reducing harm while maintaining the rights of a minority, and often marginalised, group.

These are our recommendations. Some of these are designed to counter harms caused by AI-generated disinformation in general, while others are directly related to antisemitic content:

1. **Increased transparency:** Social media platforms, where AI-generated disinformation and radical and racist content is being seen and shared by millions, should label such content. Users should be made aware that what they are seeing is AI-generated or altered and able to evaluate the information for themselves.

2. **Moderation:** AI-generated content that includes disinformation, racist, sexist or any other harmful content should be detected as such and flagged appropriately. When the content is legal and does not violate a service's terms and conditions, it is up to the service whether it should remain on the platform. This means that content that is legal but false will not be censored; it can be viewed online, but users should be made aware that it is not only AI-generated, but also contains disinformation or racist tropes. This will require producing models that are more effective than the ones tested in this report and classifiers that are trained especially to find nuanced, complex forms of racism. Considering the vast amount of antisemitic content online, we would recommend models that are trained to detect different forms of antisemitism in AI-generated content. However, because of the evolving nature of online antisemitism, especially in its implicit forms, as well as the increasing sophistication of hyper-realistic images and videos, technological means are unlikely to be sufficient. Human moderation will be required as part of the process.

3. **Accountability:** AI-generated content should be traceable to the service that created it, in order to evaluate the service safety measures, and hold services accountable for illegal content created using their tools. This will incentivise services to enhance their safety features and improve risk-assessment. Promoting good practices should also be part of any strategy employed by regulators or others to improve services' performance and safety standards.

4. **Ethical Standards:** Establish national or European guidelines for ethical AI-development, focusing on tools to prevent misuse of AI for disinformation and the spread of racist and other harmful content. Our findings show that despite some safety measures already employed by services that generate AI images, users were able to circumvent the rules or technology and generate images containing explicit and implicit forms of antisemitism, underlining the need for better safety tools and higher ethical standards.

5. **National and international collaboration:** Encourage collaboration between governments, technology companies, academia and civil society to develop and implement regulation that is effective in reducing harm, while maintaining human rights, freedom of expression and privacy.

6. **Enhanced media literacy:** Implement educational programming that improves understanding of AI and disinformation. Antisemitism in particular has become so common and widespread, that it is in many cases normalised and not recognised as such.

7. **Antisemitism education and awareness:** Educating the public about antisemitism, how to recognise it and the harm caused by it, can help prevent people from believing in, and sharing, online antisemitic content, and help counter and combat hate speech. It can also help reduce the misuse of AI-generated content by helping services to limit the use of their tools to create antisemitic content.

Conclusion

This study highlights both the ease of creating antisemitic content using AI technologies despite existing safety mechanisms, and the lack of detection of AI-generated antisemitic images. Our researchers introduced the first multidimensional annotation scheme for antisemitic deepfakes, taking into account aspects including form, content and discourse, thereby extending existing prominent work in the field. Based on this scheme, a multimodal dataset consisting of 50 images and one video was collected and annotated.

Our research found that over 50% of the images featured antisemitic stereotypes of an “evil” Jew. Many of the images can also be used to disseminate harmful mis- and disinformation that incites against Jewish people. We have also found that most existing classifiers exhibit a very weak performance when analysing antisemitic deepfakes. All models were only trained to identify if the images were generated by AI but not to classify antisemitism.

Our findings show the pressing need to make the use of generative AI safer by ensuring that illegal racist content cannot be generated and shared, by increasing transparency, improving moderation and awareness.

ANTISEMITISM POLICY TRUST

Auswärtiges Amt

Decoding Antisemitism

INACH

EUROPEAN PRACTITIONERS NETWORK AGAINST ANTISEMITISM

Contact APT

www.antisemitism.org.uk

@antisempolicy

Antisemitism Policy Trust

✉ mail@antisemitism.org.uk

The Antisemitism Policy Trust is a registered charity (1089736) [England]
and company (04146486) [England and Wales]