ANTISEMITISM
POLICY TRUST

ONLINE ANTISEMITISM AND THE RISKS OF ARTIFICIAL INTELLIGENCE

Table of Contents

Special thanks to Rupert Small for advising on this briefing.
This guide was produced by the Antisemitism Policy Trust

Introduction

Developments in Artificial Intelligence (AI) are prompting governments across the globe, and experts from across multiple sectors, to future proof society. In the UK, Ministers have published a discussion paper on the capabilities, opportunities and risks presented by frontier artificial intelligence. The document outlines that whilst AI has many benefits, it can act as a simple, accessible and cheap tool for the dissemination of disinformation, and could be misused by terrorists to enhance their capabilities. The document warns that AI technology will become so advanced and realistic, that it will be nearly impossible to distinguish deep fakes and other fake content from real content. AI could also be used to incite violence and reduce people's trust in true information.

It is clear that mitigating risks from AI will become the next great challenge for governments, and for society. Of all the possible risks, the Antisemitism Policy Trust is focused on the development of systems that facilitate the promotion, amplification and sophistication of discriminatory and racist content, that is material that can incite hatred of and harm to Jewish people. This briefing explores how AI can be used to spread antisemitism. It also shows that AI can offer benefits in combating antisemitism online and discusses ways to mitigate the risks of AI in relation to anti-Jewish racism. We set out our recommendations for action, including the development of system risk assessments, transparency and penalties for any failure to act.

What is Artificial Intelligence?

Artificial Intelligence is the science of making intelligent machines.[1] It involves teaching a machine to perform cognitive actions at a much faster rate than the human mind. Actions include reasoning and rationalising, learning, problem solving, calculating and exercising creativity. AI can help businesses and individuals work more efficiently and increase profitability.[2] Among its many functions, it can write codes, read and generate stories and articles, translate, analyse data and summarise long documents.

AI is becoming increasingly widespread and is already being used by hundreds of millions around the world. Voice assistants like Apple's Siri and Amazon's Alexa use AI technology. Siri is used by some 85 million people and Alexa by about 73 million according to estimates.[3]

AI is based on machine learning; the use of algorithms in order to detect patterns and learn to make predictions by processing very large amounts of data – too large for humans to consume and analyse. AI needs training in order to form predictions. As it processes data, it learns from its mistakes and improves its performance.[4] After the training stage, AI goes through fine tuning in order to be able to perform specialised tasks with accuracy.[5]

Developing highly capable AI can cost tens of millions of pounds – and this is expected to increase significantly in the coming years.[6] AI's advance capabilities offer countless opportunities for uses in nearly every industry, making it a potentially profitable field despite the high development costs. It can, for example both generate content that will be disseminated through social media platforms to millions around the word, and at the same time it can also help social platforms moderate the enormous volume of user-generated information, including large amounts of antisemitic content.

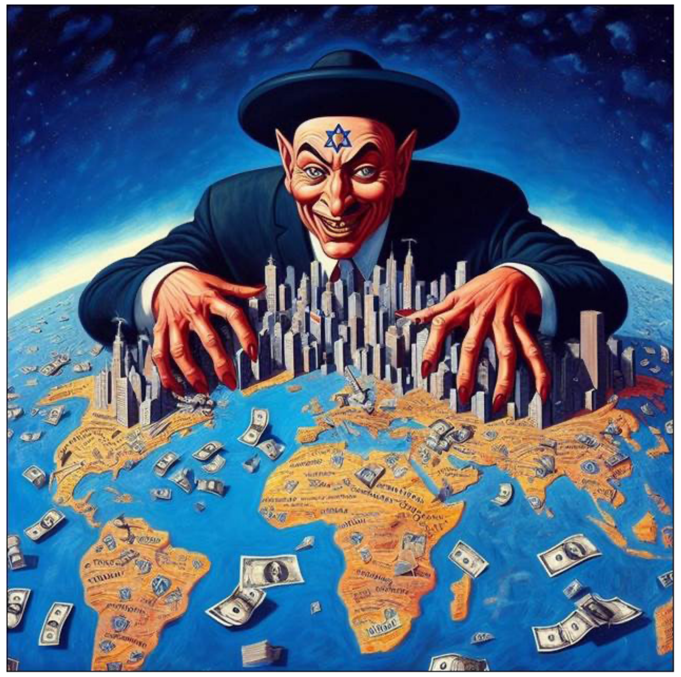

Figure 1: This AI-generated image taken from the platform 4chan, conjures an old antisemitic trope about Jewish greed and control[7]

[1] https://www.ibm.com/topics/artificial-intelligence

[2] https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-ai

[3] https://www.insiderintelligence.com/insights/voice-assistants/#:~:text=Among%20the%20top%20voice%20assistant,assistants%20and%20other%20marketing%20trends%3F

[4] https://assets.publishing.service.gov.uk/media/65395abae6c968000daa9b25/frontier-ai-capabilities-risks-report.pdf p.6

[5] https://assets.publishing.service.gov.uk/media/65395abae6c968000daa9b25/frontier-ai-capabilities-risks-report.pdf p.6

[6] https://assets.publishing.service.gov.uk/media/65395abae6c968000daa9b25/frontier-ai-capabilities-risks-report.pdf p.7

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

Advantages and Disadvantages of Using AI on Social Media Content

Scale

One of AI's benefits when considering illegal and harmful material online, including antisemitism and other types of racism and hate speech, is its ability to analyse a considerable amount of information in a short timeframe. This helps social media companies, where information is generated in extremely large volume by users, to moderate content in a way that is considerably faster, cheaper, more consistent and less subjective than using human moderators.

Content that will take a human moderator several hours to work through, can be done in milliseconds by AI. According to YouTube, for example, its AI moderation system managed to remove 80 per cent of the videos that violate its policies before they are even viewed by human moderators. This does, of course, require some human oversight given the potential implications for free speech.

AI can also be exposed to harmful material, such as violent content or materials which might invite violence without suffering the psychological impacts that human moderators would and indeed have encountered. AI also constantly learns, improves and becomes increasingly accurate, and it does so much faster than it takes to train human moderators.[8]

Bias

AI has disadvantages. It is only as good as the data used to train it. Gaps in its knowledge can therefore affect how efficient it is in correctly identifying content as breaking a platform's policies. In addition, in its training process, AI is likely to be exposed to stereotypes and bias against marginalised groups. When this happens, AI is more likely to produce results that are not always balanced and fair. It does not always understand context and nuances as well as human moderators and can produce false positives or false negatives when making judgement calls about content.

To take an example, when ChatGPT, the AI Chat Bot was asked to offer a Jewish and an Israeli joke, it offered up a distasteful joke about Jews and one based on an antisemitic trope about money. For contrast, the Bot refused to offer jokes about Muslim or Palestinian people.[9]

To take another example, AI that has been trained on male voices, had problem recognising female voices.[10] In healthcare, researchers found that when the data used for learning contains inequalities based on race, gender, ethnicity, socio-economic status and other factors, AI does not only reflect those inequalities, but its actions can amplify inequalities too.[11]

As more and more people rely on AI for information and analysis, including in health and education settings, it is important to examine the quality of the data used to teach it. This is applies also to commercial scenarios, including reviews left on products on Amazon created by AI which might be gamed in order to present better reviews of antisemitic texts, of which there are a number on the site.

[8] https://aicontentfy.com/en/blog/role-of-ai-in-content-moderation-and-censorship#:~:text=However%2C%20AI%2Dpowered%20content%20moderation,for%20safety%20and%20free%20speech

[9] https://twitter.com/ariehkovler/status/1740717084722823540

[10] https://aicontentfy.com/en/blog/role-of-ai-in-content-moderation-and-censorship#:~:text=However%2C%20AI%2Dpowered%20content%20moderation,for%20safety%20and%20free%20speech

[11] https://www.hsph.harvard.edu/ecpe/how-to-prevent-algorithmic-bias-in-health-care/

Antisemitism and AI

The Government's report on AI made clear it can be used for nefarious purposes. Jews have historically been a primary target for racist propaganda. Where once this was spread through written texts like The Protocols of the Learned Elders of Zion[12] and later in a variety of other mediums, including newspapers, radio and TV, and most recently online. AI is already being used by antisemites to spread anti-Jewish hatred and bias. Antisemitic propaganda can now be crated and greater speed and ease. It can also now be transmitted at a greater rate and have a wider reach than ever before.

Social Engineering and Antisemitism; Alt-Media and Chatbots

AI enables anti-Jewish incitement through the creation of sophisticated antisemitic fake images and deep fakes. This can be done by people with little technical skills or know-how, and for little to no cost. The Community Security Trust (CST) found for example that users on extremist platform 4Chan have produced antisemitic images using generative AI tools. Users asked AI to produce an image of 'Jew about to be killed, afraid, screaming.'[13] The image produced also included antisemitic stereotypes to make a character appear 'Jewish'.

The AI chatbot Chai was found to employ antisemitic rhetoric including that Jews are 'evil and greedy.'[14] ChatGPT produced offensive material about Jews and the Holocaust, including content that could be used by Holocaust deniers.[15] Last year, Meta's AI chatbot was found to repeat antisemitic conspiracy theories[16] whilst Microsoft's chatbot Tay, also produced racist, misogynist and homophobic results that included antisemitism.[17] Although these chatbots had not been designed to be racist on purpose, not enough thought was given to the information underpinning their systems and learning methods so that the racist content was prohibited.

These systems – which reduce the speed at which antisemitic propaganda can be created - are open to being gamed, and are increasingly being discussed and developed in extremist spaces.[18] Despite some companies incorporating safety features within their AI, some have found that using clever, or even simple, prompts can help manipulate large language models (LLM), tricking the AI into giving the desired reply. On Telegram, for example, one user suggested using the terms 'WW2 leader' or 'WW2 German Chancellor' to create images of Hitler.

Antisemitic Disinformation; Establishing Provenance and the Erosion of Trust

AI constantly evolves. The war between Israel and Hamas that began on 7 October was followed by a considerable rise in antisemitic incidents in the UK[19] and in antisemitism online. This has been intensified by large-scale misinformation, disinformation and conspiracy theories. These include antisemitic blood libels, claims that the heinous violence committed by Hamas against Israelis on 7 October is made up and is an Israeli lie, that the attack was, in fact, an Israeli plot, or one laid by Jews, in order to justify going to war, and many more false accusations and anti-Jewish propaganda.

AI has added to the barrage of disinformation. AI-generated images and deep fake videos used to manipulate public opinion have surfaced since the start of the war.[20] Some are sophisticated enough to fool social media users, who have been sharing those images thinking they represent true depictions of events. The influx of AI-generated disinformation has been so great, that fact checkers and analysts have been struggling moderating the content.[21] The origin of the content is at times also impossible to detect.

AI images, even these are not in of themselves antisemitic but anti-Israeli, have been used to increase tensions and stir up anger and unrest, correlating with an increase in Jewish people being targeted with hate crimes worldwide. Issues that tend to polarise, such as the war in Gaza, can be drivers of AI-generated disinformation, racist content and bias. Polarisation can cause such content to be created at big volume and spread throughout the online space faster.

In addition, the level of sophistication used by AI, and knowing that every image, video and text can be manipulated and artificially produced, makes people question or distrust what they see and question information even when it is factual. This is happening during the war, and will continue to happen if there is no transparency about what is and is not generated by AI.

Much of the AI-generated content has been used to target Israel and Jews, linked with a hate-fuelled reaction against Jews, while some of it – although at a much reduced scale – has been used to garner support for Israel.[22] [23] Some of the most common AI generated images are of Palestinian children amongst the ruins of Gaza.[24] These, in some cases, are specifically designed to emotionally manipulate the public against Israel.[25] Antisemites have also been using AI to general antisemitic memes of paragliders (a tool used by Hamas to invade Israel on 7 October) as a symbol of glorifying the killing of Jews.[26]

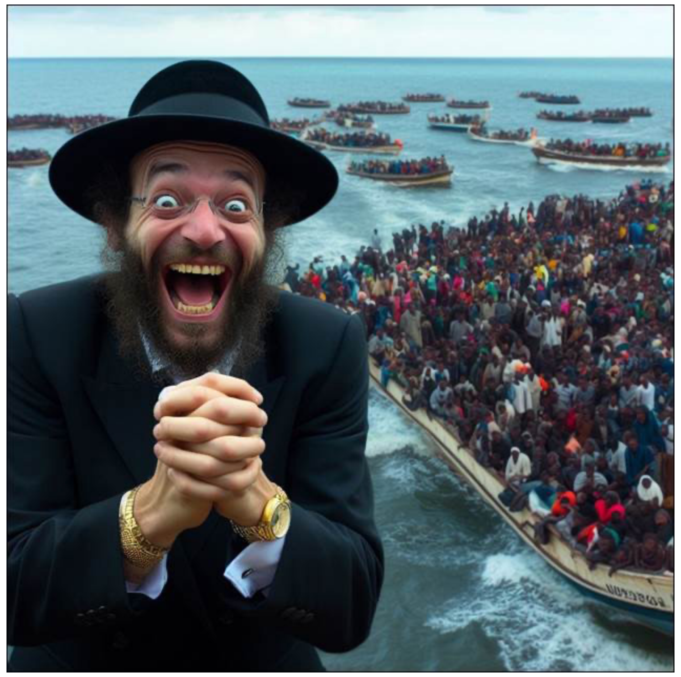

Figure 2: This AI-generated image from the platform 4chan, shows an orthodox Jewish man clasping his hands, covered in gold Jewellery, in happiness, with images of what are supposed to present as illegal migrants behind him. This relates to the antisemitic conspiracy the Great Replacement Theory. This theory alleges that Jews are responsible for illegal immigration, organizing it as part of a plot to cause instability and destroy white western civilisation through mass immigration.[27]

[12] https://antisemitism.org.uk/wp-content/uploads/2022/11/APT-Protocols-Report.pdf

[13] https://gnet-research.org/2023/06/12/does-artificial-intelligence-dream-of-antisemitism/

[14] https://gnet-research.org/2023/06/12/does-artificial-intelligence-dream-of-antisemitism/

[15] https://www.lawfaremedia.org/article/it-was-smart-for-an-ai

[16] https://www.businessinsider.com/meta-ai-chatbot-blenderbot-election-denying-antisemitic-bugs-artificial-intellignce-2022-8?r=US&IR=T

[17] https://www.theguardian.com/world/2016/mar/29/microsoft-tay-tweets-antisemitic-racism

[18] https://gnet-research.org/2023/06/12/does-artificial-intelligence-dream-of-antisemitism/

[19] https://cst.org.uk/news/blog/2023/10/27/antisemitic-incidents-27-october-update

[20] https://www.rollingstone.com/politics/politics-features/israel-hamas-misinformation-fueled-ai-images-1234863586/

[21] https://www.euronews.com/my-europe/2023/10/24/israel-hamas-war-this-viral-image-of-a-baby-trapped-under-rubble-turned-out-to-be-fake

[22] https://www.spectator.co.uk/article/no-one-should-trust-the-camera-in-the-age-of-ai/

[23] https://www.rollingstone.com/politics/politics-features/israel-hamas-misinformation-fueled-ai-images-1234863586/

[24] This does not suggest that all images of Palestinian children are generated by AI. Many images are real and tragic.

[25] https://www.rollingstone.com/politics/politics-features/israel-hamas-misinformation-fueled-ai-images-1234863586/

[26] https://www.rollingstone.com/politics/politics-features/israel-hamas-misinformation-fueled-ai-images-1234863586/

[27] https://antisemitism.org.uk/wp-content/uploads/2021/04/Final-George-Soros-Briefing.pdf

Use of AI to Combat Antisemitism

Dual Use

AI can help exacerbate the problem of online antisemitism, but it can also be used to combat antisemitism. As previously discussed, AI can be a useful tool in moderating content. It can help find, recognise and remove illegal antisemitic material that includes hate speech and threats quickly and before such content reaches a wide audience, but this requires effective training based on trusted data.

The examples of antisemitism produced by AI show that so far AI aids the spread of antisemitic content. This means that it can also be less effective in moderating against such content. Although moderation is not done by chatbots, the material used to teach different types of AI may bare similarities.

One example of using AI to combat antisemitism is undertaken by the project Decoding Antisemitism. It is an interdisciplinary effort that includes discourse analysts, computational linguists and historians from Germany, the UK and France, who have developed AI that detects explicit and implicit antisemitism on social media platforms.[28] The process involves a supervised machine learning approach that trains the system to recognise evolving antisemitic concepts to increase precision and effectiveness. The projects published regular reports that analyse antisemitism on social media platforms, usually in reaction to specific events. It is the kind of specialised AI that can be used by social media platforms.

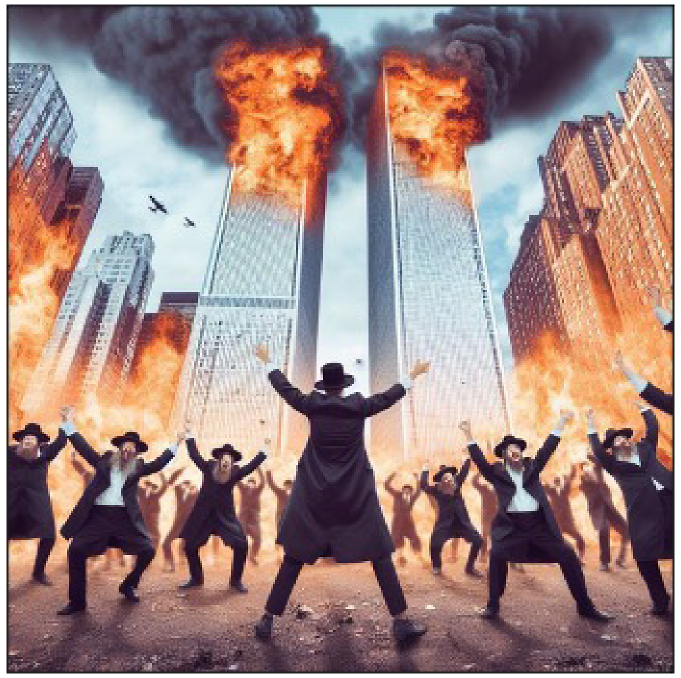

Figure 3: This AI-generated image, taken from the platform 4chan, shows Jews celebrating the 9/11 terror attacks. There have been many conspiracies that allege Jews orchestrated this attack[29]

[28] https://decoding-antisemitism.eu/

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

How Can Risks Be Mitigated?

As the use of AI spreads rapidly, the risks should be seriously considered, and plans to mitigate the effect of AI need to be drawn out. The European Union set out a regulatory framework in its AI Act, first proposed in 2021.[30] It aims to make the use of AI safe, transparent, traceable, non-discriminatory and environmental friendly.

The EU regulation will not consider all AI equally, but will instead take a risk-based approach that will categorise AI models according to levels of threat from limited risk to unacceptable risk. AI that is believed to pose an unacceptable risk will be banned. Generative AI, including chatbots, will need to be transparent, so users know the content has been generated by AI. China also introduced legislation to regulate generative AI in 2023, although it has been criticised for being significantly watered down in the legislative process.[31] Canada and Brazil are also legislating to regulate AI.

To stay ahead of developments, regulation should consider how the use of AI will evolve and the harm it may caused when implemented in new ways. Gaming is one space to look out for, where there have been instances of antisemitism in games. Robolox, which has over 216 million monthly users, has instances of antisemitic content.[32] If AI also plays a part in the fast evolving industry of virtual reality headsets, that should also be assessed and covered by regulation.

The tech industry appears generally supportive of regulating AI, as long as it is not done in a way that will stifle innovation. Microsoft for example suggested that firms should be required to register models that exceed certain performance thresholds.[33] Tech companies believe that it is the model's application rather than the models themselves that need to be regulated. However, considering the harmful effect that the information used to train models has on its output, this is also a field that needs to be regulated and overseen. The Trust believes that whether it is Bing's Image Creator, OpenAI's DALL-E software or Stability AI's Stable Diffusion, all must enhance their systems to be better at resisting gaming by bad actors.

As we move to the future, countries will need to find the balance between mitigating risks generated by AI by having effective regulation that can capture a fast-evolving technology that has a huge variety of applications in different fields. This will need to be done in a way that allows this industry to evolve so that humanity can enjoy the benefits it can offer.

The Online Safety Act already places responsibility on many platforms to moderate content and conduct risk assessments. However, the nature of AI may require more robust attention and regulation. The Trust is keen that a number of safeguards be introduced to ensure AI is more carefully developed and deployed:

1) Risk Assessments and Safety By Design:

Design of any technology should incorporate safety, and efforts to avoid reasonably foreseeable harms. Antisemitism is not new, snd there can be no excuse for releasing products which do not guard against racist abuse. Producing a risk assessment should be standard practice in the development of technologies but is not. Particularly given the potential for AI to do harm any company developing such technology should be required by law to produce a risk assessment which meets a minimum standard

Oversight over the quality of the data used in machine learning to avoid bias, stereotypes and racism is also crucial. A negative correlation between the size of data sets used to teach AI and level of accuracy has

[30] https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

[31] https://www.eastasiaforum.org/2023/09/27/the-future-of-ai-policy-in-china/#:~:text=On%2015%20August%202023%2C%20a,that%20specifically%20targets%20generative%20AI

[32] https://www.jewishnews.co.uk/roblox-doing-all-it-can-to-tackle-proliferation-of-antisemitic-content-on-platform/

[33] https://www.economist.com/business/2023/10/24/the-world-wants-to-regulate-ai-but-does-not-quite-know-how

been found.[34] Although developers tend to favour the use of large-scale data sets to teach AI, and on many occasions they use data that has not been sufficiently audited, because it time consuming and costly – emphasising the quality of data over the quantity of data in machine learning seems to be the way forward.

Furthermore, action to guard against polymorphism - the tendency of narratives to evolve and mutate - must feature in systems development. At present, there is a paucity of capabilities to track and manage these rapidly evolving narratives.

So too, action should be taken to address polarisation which can feed conflict, racism, disinformation, and bias. This is easy to do in order to manipulate populations towards harmful behaviours.

Work to address Open Source innovation would also be welcome. Open Source models for LLMs are vying with commercial models, and so distribution and moderation efforts need to be considered by the appropriate regulators.

2) Transparency and Digital Education:

Transparency can help mitigate the harmful effects of AI. A clear indication that an image, a meme or any other content is generated by AI would be helpful in limiting the spread of disinformation and misinformation. Ofcom's committee on mis-and dis-information will need to consider this matter. The Government should be investing in digital media literacy to ensure that children are better prepared to probe, query and investigate online materials. With technology companies and schools working in tandem, there will be a better chance to reduce harm.

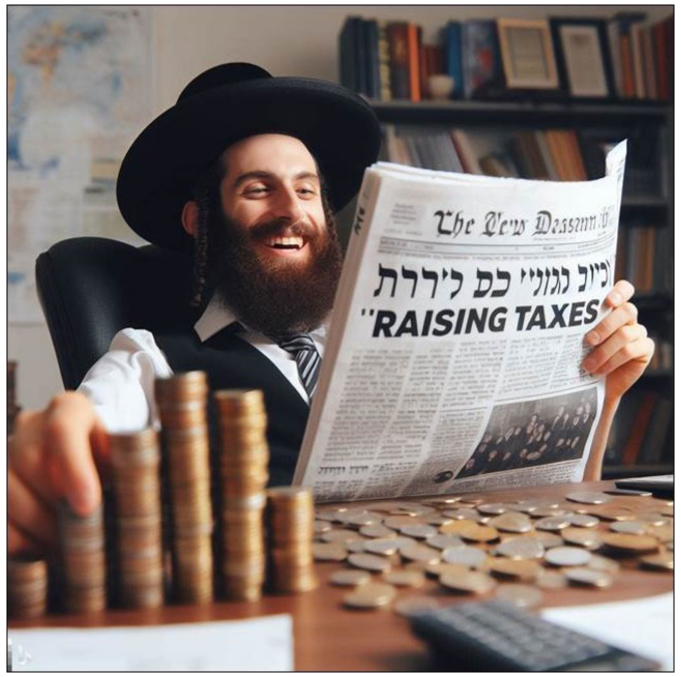

Figure 4: *This AI-generated image, found on the platform 4chan, shows a visibly Jewish man surrounded by money and laughing at a headline in a Jewish newspaper about rising taxes. This image feeds from antisemitic tropes about Jewish greed and Jews being powerful through benefitting from the plight of ordinary people.*[35]

3) Penalties:

If a company develops a project that could reasonably be foreseen as harmful, the company must be held accountable for it. Where harm is caused by its product, it must be made to pay for redesigning that product to prevent harm, an approach which underpins the UK regulatory system in many sectors. Regulators whether they be existing or new, should be given the resources and powers to prosecute companies that put unsafe, badly designed products onto the market.

Banning AI from producing illegal content, such as bomb-making instructions or anything else that can be used for terrorism, threats and hate speech will also be important, with criminal penalties for failures to comply.

[34] https://www.wired.co.uk/article/abeba-birhane-ai-datasets

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

The Antisemitism Policy Trust's mission is to educate and empower parliamentarians, policy makers and opinion formers to address antisemitism. It provides the secretariat to the British All-Party Parliamentary Group Against Antisemitism and works internationally with parliamentarians and others to address antisemitism. The Antisemitism Policy Trust is focussed on educating and empowering decision makers in the UK and across the world to effectively address 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]