Generative AI Misuse: A Taxonomy of Tactics and Insights from Real-World Data

Nahema Marchal[*,1], **Rachel Xu**[*,2], **Rasmi Elasmar**[3], **Iason Gabriel**[1], **Beth Goldberg**[2] **and William Isaac**[1]
[*]Equal contributions, [1]Google DeepMind, [2]Jigsaw, [3]Google.org

Generative, multimodal artificial intelligence (GenAI) offers transformative potential across industries, but its misuse poses significant risks. Prior research has shed light on the potential of advanced AI systems to be exploited for malicious purposes. However, we still lack a concrete understanding of how GenAI models are specifically exploited or abused in practice, including the tactics employed to inflict harm. In this paper, we present a taxonomy of GenAI misuse tactics, informed by existing academic literature and a qualitative analysis of approximately 200 observed incidents of misuse reported between January 2023 and March 2024. Through this analysis, we illuminate key and novel patterns in misuse during this time period, including potential motivations, strategies, and how attackers leverage and abuse system capabilities across modalities (e.g. image, text, audio, video) in the wild.

arXiv:2406.13843v2 [cs.AI] 21 Jun 2024

Corresponding author(s): Nahema Marchal <nahemamarchal@google.com>
© 2024 Google DeepMind. All rights reserved

Contents

Acknowledgements

We would like to thank Mikel Rodriguez, Vikay Bolina, Alexios Mantzarlis, Seliem El-Sayed, Mevan Babakar, Matt Botvinick, Canfer Akbulut, Harry Law, Sébastien Krier, Ziad Reslan, Boxi Wu, Frankie Garcia, and Jennie Brennan, for their feedback and contributions to this paper.

1. Introduction

Generative, multimodal artificial intelligence (GenAI) brings forth new possibilities across industries and creative domains. Over the past year, leading AI labs have unveiled models that demonstrate sophisticated capabilities across tasks: from complex audiovisual understanding and mathematical reasoning (Gemini Team, 2024) to realistic simulation of real-world environments (Brooks et al., 2024). These systems are rapidly being integrated into critical sectors like healthcare (Yim et al., 2024), education (Qadir, 2023) and public services (Bright et al., 2024). Yet, as GenAI capabilities advance, so does awareness of these tools' potential for misuse, including heightened concerns around security, privacy and manipulation (Barrett et al., 2023; Feuerriegel et al., 2023; Golda et al., 2024; Pauly, 2024; Shevlane et al., 2023).

Prior research has shed light on the potential of advanced AI systems to be exploited for malicious purposes using foresight analysis and hypothetical scenarios, which aim to map out future ethical risks in a systematic way (Barrett et al., 2023; Blauth et al., 2022; Brundage et al., 2018; Ferrara, 2024; Goldstein et al., 2023; Rodriguez et al., 2024). Complementing this research, initiatives such as the OECD AI incidents monitor (AIM), the AI, Algorithmic, and Automation Incidents and Controversies repository (AIAAIC) and the AI Incident Database initiatives actively record AI-related incidents across applications and sort their associated harms. While these efforts provide a rich foundation for mapping AI-enabled threats, they tend to be broad in scope and focus on identifying potential risks and downstream harms. In contrast, we still don't know enough about how GenAI tools are specifically exploited and abused by different actors, including the tactics employed. As the technology itself becomes more sophisticated and multimodal, better understanding how these manifest in practice and across modalities, is critical.

In this paper, we first present a taxonomy of GenAI misuse tactics, informed by existing academic literature and a qualitative analysis of 200 media reports of misuse and demonstrations of abuse of GenAI systems published between January 2023 and March 2024). Based on this analysis, we then illuminate key and novel patterns in GenAI misuse during this time period (see Section 4: Findings), including potential motivations, strategies, and how attackers leverage and abuse system capabilities across modalities (e.g. image, text, audio, video) in an uncontrolled environment.

We find that:

1. **Manipulation of human likeness** and **falsification of evidence** underlie the most prevalent tactics in real-world cases of misuse. Most of these were deployed with a discernible intent to influence public opinion, enable scam or fraudulent activities, or to generate profit.
2. The majority of reported cases of misuse do not consist of **technologically sophisticated uses of GenAI systems or attacks**. Instead, we are predominantly seeing an exploitation of easily accessible GenAI capabilities requiring minimal technical expertise.
3. The increased sophistication, availability and accessibility of GenAI tools seemingly introduces **new and lower-level forms of misuse** that are neither overtly malicious nor explicitly violate these tools' terms of services, but still have concerning ethical ramifications. These include the emergence of new forms of communications for political outreach, self-promotion and advocacy that blur the lines between authenticity and deception (see Section 5: Discussion).

Our findings provide policy makers, trust and safety teams, and researchers with an evidence base of these technologies' potential for real-world harm, which can inform their approach to AI governance and mitigations. Second, by providing an overview of salient threats and tactics, this work can also guide the development of safety evaluations and adversarial testing strategies that are more aligned with the rapidly-shifting threat landscape. Lastly, by identifying prominent misuse

tactics across modalities this work can help inform targeted mitigations and interventions, with the possibility to better inoculate the public against specific misuse strategies in the future.

Definitions and Scope

By generative AI ('GenAI'), we refer to a class of large-scale models trained on billions of parameters of text, audio, code, images and video, mostly sourced from publicly available sources.[1] These models, which include large language and diffusion models, can be adapted to a range of downstream tasks, including text and media generation, problem solving (Huang and Chang, 2022), data extraction (Gartlehner et al., 2023), and coding assistance (Tian et al., 2023) among others. They also show several novel capabilities such as the ability to learn how to use external tools (Schick et al., 2023) and to perform new tasks without needing to be retrained on large datasets (Wang et al., 2023; Wei et al., 2022).

Throughout the paper, building on the definition proposed by Blauth et al. (2022) we refer to GenAI 'misuse' as the deliberate use of generative AI tools by individuals and organisations to facilitate, augment or execute actions that may cause downstream harm, as well as attacks on generative AI systems themselves. This definition excludes accidents or cases where harm is caused by malfunctions or limitations of GenAI systems themselves, such as their tendency to hallucinate facts or produce biassed outputs (Ji et al., 2023; Maynez et al., 2020), without a discernible actor involved.

2. Methodology

To develop our taxonomy, we first conducted a review of recent academic and grey literature focusing on malicious uses of generative AI. This initial review provided the initial theoretical foundations for identifying and categorising misuse tactics. We then collected and qualitatively analysed a dataset of media reports of GenAI misuse, as defined above, to validate and augment our taxonomy. Our dataset consists of individual media reports published between January 2023 and March 2024 documenting one or more proof points of real-world misuse involving GenAI. Two authors first independently reviewed each media report in the dataset to identify relevant misuse tactics employed. Our initial taxonomy categories were then continuously updated and expanded based on emerging patterns in the data. Any disagreements between authors were thoroughly discussed to reach consensus, ensuring consistency and accuracy in the classification process.

Whenever possible or clearly identifiable from the reporting, we also extracted information about the actors involved in the misuse, their underlying goals[2], and the individuals organisations or models targeted. This enabled us to identify broader misuse strategies emerging from the combination of specific goals, tactics, uses of GenAI applications, and targets. We present these findings in Section 4 (See Appendix A and Appendix B for a complete list of identified goals and strategies).

To ensure a good coverage of GenAI misuses in our dataset, we employed two data collection approaches. First, we leveraged a proprietary social listening tool that aggregates content from millions of sources — including social media platforms such as X and Reddit, blogs and established news outlets — to detect potential abuses of GenAI tools. We supplemented this with an additional manual search for relevant articles from reputable news sources and blogs published between Jan 1st,

[1] This includes large language and diffusion models such as GPT-4 (OpenAI et al., 2023), Gemini (Gemini Team, 2024), Claude 3 (Anthropic, 2024), LLaMA 2 (Touvron et al., 2023), Sora (Brooks et al., 2024), DALL-E 3 (Betker et al., 2023), and Stable Diffusion that can generate outputs across text, image, video, audio and code modalities.

[2] While inferring motivations or intent with complete certainty is not always possible for every instance of misuse, for our analysis we rely on contextual information provided from the reporting to offer an educated guess.

2023 and March 5th, 2024, based on a list of keywords relevant to GenAI misuse.[3] Data was then parsed to identify and remove duplicates. After de-deduplication and removal of out-of-scope cases, our dataset contains a total of 191 cases.

While using media reports as a primary source of data allows us to capture emerging trends and novel techniques as they are deployed, and emphasise cases that cause significant harm or disruption, there are nevertheless some limitations to this approach (see Section 6: Limitations). Notably, we acknowledge the potential for underreporting of covert misuse operations or instances where the GenAI tactics used are novel or difficult to detect. Additionally, cases that cause less noticeable harm may receive less attention in the press, creating potential blindspots in our data.

3. Taxonomy of Generative AI Misuse Tactics

To systematically categorise GenAI misuse tactics, we propose a taxonomy that distinguishes between two types of tactics: (1) tactics that involve the exploitation of GenAI's capabilities, and (2) tactics that involve compromising or attacking GenAI systems themselves.

3.1. Exploitation of GenAI capabilities

Across modalities, GenAI models are characterised by their ability to synthesise highly-realistic outputs (Cooke et al., 2024; Nightingale and Farid, 2022), including convincingly mimicking writing and artistic styles (Syed et al., 2020). They can produce vast amounts of this content efficiently, allowing for rapid distribution of a high volume of synthetic content in a short time frame. GenAI tools are also characterised by their widespread availability and their ease of use, with intuitive dialogue interfaces that require minimal technical knowledge. While these features enhance the quality and ease of user experiences with GenAI tools, each of these also provide opportunities for exploitation. We identify 10 distinct tactics that exploit GenAI capabilities, summarised in Table 1 below.

The first five tactics listed leverage GenAI models to create realistic depictions of people from natural language descriptions. To distinguish between tactics, we used three framing questions that we believe yield different trust and safety implications: (1) Does the generated output depict a real person or an entirely synthetic one? (2) Is the depiction static or in real-time interaction? and (3) How is the generated content being used?

When the GenAI output depicts a real person, and is ostensibly used to take action on their behalf in real time (e.g. an AI-generated audio clip of someone, in which they claim to be a famous politician), we term this tactic **Impersonation**. Active depictions such as these carry a high potential for harm, as they can easily mislead an unwitting audience due to their realism and resemblance to everyday experiences and scenarios (Sundar, 2008; Vaccari and Chadwick, 2020). When an output depicts a real person in a static manner and at a fixed point in time (e.g. a synthetic image of a celebrity in a country they've never visited), we term this tactic as **Appropriated Likeness**. This often leverages diffusion models' in-painting and out-painting capabilities — the ability to make changes to an existing uploaded image (such as removing details) and to complete missing parts of a photo or image beyond its original (Lugmayr et al., 2022).[4] When a GenAI tool is used to depict entirely synthetic personas and make them take action in the world, we call this tactic **Sockpuppeting**. While not directly targeting specific individuals, the creation of synthetic personas, such as fake experts and

[3] Our search strategy included a combination of GenAI-related terms and models (AI-generated, generative AI, GenAI, deepfake*, Gemini, GPT*, Claude, Llama, DALL-E, Midjourney, Stable Diffusion, Bard, Galactica, Sora) and misuse-related terms (harm*, use*, misuse*, abuse*, harass*, victim*, attack*, safety, malicious, manipulat*, dece*, false, fake, content*, data)

[4] See for e.g. Insert or replace objects with Generative fill

Table 1 | Misuse tactics that exploit GenAI capabilities

	Tactic	Definition	Example
Realistic depictions of human likeness	Impersonation	Assume the identity of a real person and take actions on their behalf	AI robocalls impersonate President Biden in an apparent attempt to suppress votes in New Hampshire
	Appropriated Likeness	Use or alter a person's likeness or other identifying features	Photos of detained protesting Indian wrestlers altered to show them smiling
	Sockpuppeting	Create synthetic online personas or accounts	Army of fake social media accounts defend UAE presidency of climate summit
	Non-consensual intimate imagery (NCII)	Create sexual explicit material using an adult person's likeness	Celebrities injected in sexually explicit "Dream GF" imagery
	Child sexual abuse material (CSAM)	Create child sexual explicit material	Deepfake CSAI on sale on Shopee
Realistic depictions of non-humans	Falsification	Fabricate or falsely represent evidence, incl. reports, IDs, documents	AI-generated images are being shared in relation to the Israel-Hamas conflict
	Intellectual property (IP) infringement	Use a person's IP without their permission	He wrote a book on a rare subject. Then a ChatGPT replica appeared on Amazon.
	Counterfeit	Reproduce or imitate an original work, brand or style and pass as real	Fraudulent copycats of Bard and ChatGPT appear online
Use of generated content	Scaling & Amplification	Automate, amplify, or scale workflows	Researchers use GPT-3 to mass email state legislators, signaling rising verisimilitude of AI-generated emails
	Targeting & Personalisation	Refine outputs to target individuals with tailored attacks	WormGPT can be used to craft effective phishing emails

activists, can lend a powerful force multiplier to actors seeking to shape and manipulate information in digital environments (Harris, 2023).[5]

Finally, we separate out the creation of non-consensual sexually explicit material of adults (**NCII**) and the production of child sexual abuse material (**CSAM**) even though they may deploy any of the three tactics above. These tactics warrant separate categorization due to their uniquely damaging potential, and our understanding of how practitioners treat this content in practice. Unlike any of the tactics listed above, using GenAI to create CSAM or NCII is generally considered policy violative, regardless of how that content is used.

The next three tactics leverage audio, image or video-generation capabilities to create realistic depictions of non-humans, including documents, songs, styles, events, places. This includes **IP Infringement**, where GenAI is used to produce parts or the entirety of someone's intellectual property — such as literary and artistic works — without permission. When GenAI is used to create digital items that mimic an original work or style and falsely represent them as authentic (for example, appropriating a reputable news website's logo and layout), we refer to this tactic as **Counterfeit**. When a GenAI output depicts fabricated events, places or objects presented as real, we refer to this tactic as **Falsification**.

Finally, GenAI models can be used to create high volumes of textual or audio-visual content and facilitate distribution and engagement with that content. For example, operating large networks of fake social media profiles that employ LLMs to generate human-like content, images and audio,

[5] See for e.g. Marco Silva

as well as engage with each other or other users online (DiResta and Goldstein, 2024). We refer to this cluster of tactics as **Scaling and Amplification**. Actors can also leverage similar capabilities to refine model outputs and target them to specific audiences (**Targeting and Personalisation**), such as translating content into different languages to tailor to different geographies (Yang et al., 2023). Many of these tactics can be used in combination. For example, malicious actors orchestrating large-scale influence operations often use Sockpuppeting and Scaling and Amplification methods together to create botnets — fake profiles that appear to be real individuals (Gorwa and Guilbeault, 2020).

3.2. Compromise of GenAI systems

Beyond this, we also identify several tactics aiming to compromise GenAI systems themselves. Unlike the tactics presented above, these tactics do not exploit capabilities but instead vulnerabilities in GenAI systems themselves. While there is already extensive literature documenting the exploitation and abuse of AI systems writ large, attempts to game or manipulate GenAI models are relatively new and rapidly-evolving (Rodriguez et al., 2024). We identify 8 tactics in this vein from our analysis of media reports, most of which map on to research demonstrations of model vulnerabilities and possible compromise pathways, rather than genuine attacks from malicious actors (see Section 4: Findings). These are summarised in Table 2 below.

We separate these tactics based on the part of the system that the compromise is targeted at. Broadly, we distinguish between attacks on: (1) Model integrity (attacks that manipulate the model itself, its structure, settings or input prompts) and (2) Data integrity (attacks that alter the model's training data or compromise its security and privacy).

Adversarial Inputs involve modifying individual input data to cause a model to malfunction. These modifications, which are often imperceptible to humans, exploit how the model makes decisions to produce errors (Wallace et al., 2019) and can be applied to text, but also to images, audio, or video (e.g. changing pixels in an image of a panda in a way that causes a model to label it as a gibbon).[6]

Prompt Injections are a form of Adversarial Input that involve manipulating the text instructions given to a GenAI system (Liu et al., 2023). Prompt Injections exploit loopholes in a model's architectures that have no separation between system instructions and user data to produce a harmful output (Perez and Ribeiro, 2022). While researchers may use similar techniques to test the robustness of GenAI models, malicious actors can also leverage them. For example, they might flood a model with manipulative prompts to cause denial-of-service attacks or to bypass an AI detection software.

Jailbreaking aims to bypass or remove restrictions and safety filters placed on a GenAI model completely (Chao et al., 2023; Shen et al., 2023). This gives the actor free rein to generate any output, regardless of its content being harmful, biassed, or offensive. All three of these are tactics that manipulate the model into producing harmful outputs against its design. The difference is that prompt injections and adversarial inputs usually seek to steer the model towards producing harmful or incorrect outputs from one query, whereas jailbreaking seeks to dismantle a model's safety mechanisms in their entirety.

Model Diversion takes model manipulation one step further, by repurposing (often open-source) generative AI models in a way that diverts them from their intended functionality or from the use cases envisioned by their developers (Lin et al., 2024). An example of this is training the BERT open source model on the DarkWeb to create DarkBert.[7]

[6] See for e.g. Attacking machine learning with adversarial examples

[7] Scientists Train New AI Exclusively on the Dark Web

Table 2 | Misuse tactics to compromise GenAI systems

	Tactic	Definition	Example
Model integrity	Prompt injection	Manipulate model prompts to enable unintended or unauthorised outputs	ChatGPT workaround returns lists of problematic sites if asked for avoidance purposes
	Adversarial input	Add small perturbations to model input to generate incorrect or harmful outputs	Researchers find perturbing images and sounds successfully poisons open source LLMs
	Jailbreaking	Bypass restrictions on model's safeguards	Researchers train LLM to jailbreak other LLMs
	Model diversion	Repurpose pre-trained model to deviate from its intended purpose	We Tested Out The Uncensored Chatbot FreedomGPT
	Model extraction	Obtain model hyperparameters, architecture, or parameters	ChatGPT Spills Secrets in Novel PoC Attack
	Steganography	Hide message within model output to avoid detection	Secret Messages Can Hide in AI-Generated Media
	Poisoning	Manipulate a model's training data to alter behaviour	Researchers plant misinformation as memories in BlenderBot 2.0
Data integrity	Privacy compromise	Compromise the privacy of training data	Samsung bans use of ChatGPT on corporate devices following leak
	Data exfiltration	Compromise the security of training data	Researchers find ways to extract terabytes of training data from ChatGPT

Steganography is the practice of hiding coded messages in GenAI model outputs, which may allow malicious actors to communicate covertly.[8] **Data Poisoning** involves deliberately corrupting a model's training dataset to introduce vulnerabilities, derail its learning process, or cause it to make incorrect predictions (Carlini et al., 2023). For example, the tool Nightshade is a data poisoning tool, which allows artists to add invisible changes to the pixels in their art before uploading online, to break any models that use it for training.[9] Such attacks exploit the fact that most GenAI models are trained on publicly available datasets like images and videos scraped from the web, which malicious actors can easily compromise.

Privacy Compromise attacks reveal sensitive or private information that was used to train a model. For example, personally identifiable information or medical records. **Data Exfiltration** goes beyond revealing private information, and involves illicitly obtaining the training data used to build a model that may be sensitive or proprietary. **Model Extraction** is the same attack, only directed at the model instead of the training data — it involves obtaining the architecture, parameters, or hyper-parameters of a proprietary model (Carlini et al., 2024).

4. Findings

In this section, we summarise findings from our analysisof media reports of GenAI misuse between January 2023and March 2024 to provide an empirically-grounded under-standing of how the threat landscape of GenAI is evolving. Beyond identifying salient misuse tactics, whenever possible and discernible from the reporting we also extracted information for each case in our dataset about

[8] Secret Messages Can Hide in AI-Generated Media
[9] This new data poisoning tool lets artists fight back against generative AI

the actors involved in the misuse, their goals, and which specific GenAI tools were exploited across modalities. Our analysis reveals patterns in how these elements combine into broader misuse strategies. We tabulate this information in Appendix A and Appendix B and use these data points to enrich our discussion.

4.1. Prevalence and modalities of misuse tactics

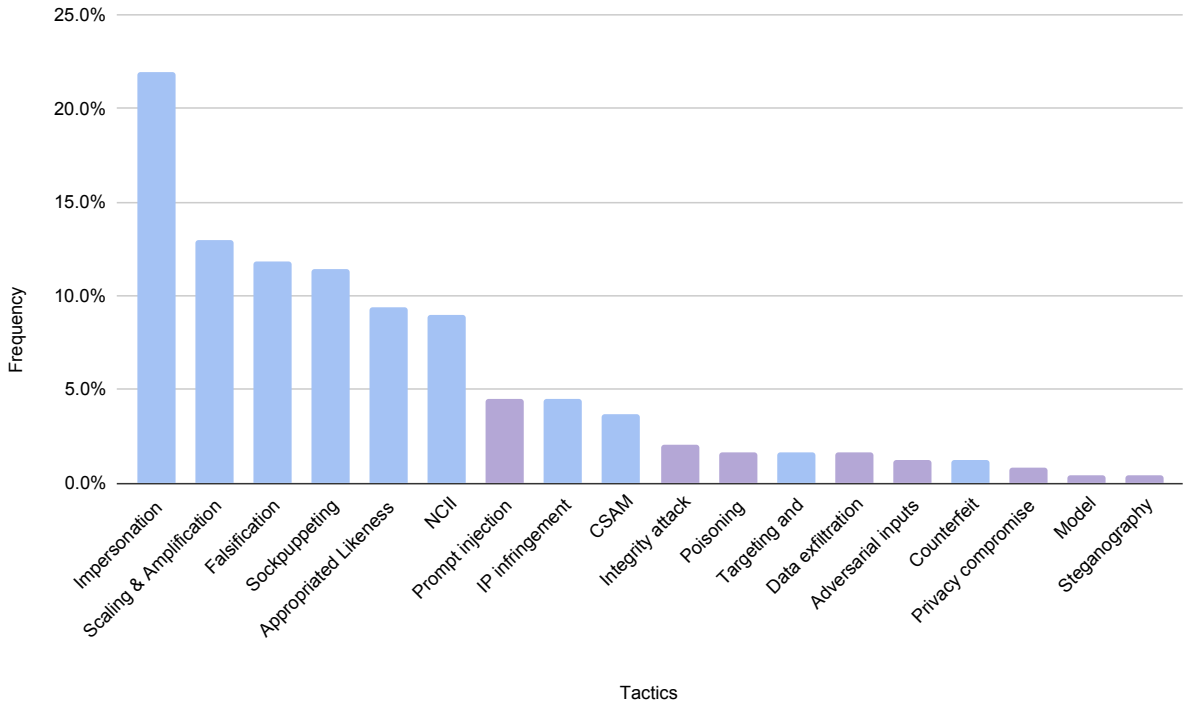

Figure 1 | Frequency of tactics across categories.
Note: Each bar represents the frequency with which a tactic was identified within our dataset. Each case of misuse could involve more than one tactic.

First, we find that most reported cases of GenAI misuse involve actors exploiting the capabilities of these systems, rather than launching direct attacks at the models themselves (see Figure 1). Nearly 9 out of 10 documented cases in our dataset fall into this category. Of these, the most prevalent cluster of tactics involve the manipulation of human likeness, especially Impersonation (followed by Sockpuppeting, Appropriated Likeness and NCII). Scaling & Amplification and Falsification are also prominent tactics, accounting for 13% and 12% of reported cases respectively.

As Table 3 below shows, Impersonation typically involves text-to-speech and video generation tools to replicate people's voices and likeness, especially that of public figures.[10] Falsification of content, on the other hand, mostly draws on text and image generation to create synthetic books, news articles and website copy, and images to accompany these articles, such as synthetic images of events that never took place (for e.g., the fake images of explosions at the Pentagon[11]). NCII primarily relies on manipulating image and video modalities to create suggestive deepfakes of private individuals or celebrities. Sockpuppeting and tactics centred on scaling and amplifying content distribution involve creating synthetic social media profiles with AI-generated profile pictures and descriptions.

[10] See for e.g. Faked AI audio hits Harlem politics
[11] Fake image of Pentagon explosion briefly sends jitters through stock market

Table 3 | Modalities associated with each tactic.

Tactic	Modality				Total
	Image	Text	Audio	Video	
Impersonation	4	3	28	21	56
Sockpuppeting	17	18	7	6	48
Scaling & Amplification	15	24	4	1	44
Falsification	16	12	4	2	34
NCII	11	1	1	11	24
Appropriated Likeness	12	4	2	2	20
IP Infringement	2	7	3		12
CSAM	9	1			10
Targeting/ Personalisation		5	2		7
Counterfeit		3			3
Total	86	78	51	43	258

Note: Counts denote the number of times a tactic was linked with this specific modality in our data.

4.2. Goals and strategies of misuse

GenAI misuse does not happen in a vacuum. Actors often have discernible goals or specific motivations to misuse and abuse GenAI, not all of which are necessarily adversarial. These range from financial gain to harassment, and political disruption (see Appendix A for full breakdown). Understanding these motivations is crucial for assessing the severity of downstream impacts and crafting appropriate countermeasures. By observing how actors combine misuse tactics in pursuit of their goals, we can also identify specific patterns of misuse: we label these combinations as strategies (see Appendix B) Below, we identify the most common goals behind GenAI misuse, along with the most prevalent and novel strategies used to achieve them.

Between 2023-2024, our data shows that **attacks on GenAI systems themselves were mostly conducted as part of research demonstrations or testing** aimed at uncovering vulnerabilities and weaknesses within these systems (See Table 4 above). Within this subset, approximately a third of these attempts employed Prompt Injection as a tactic. In contrast, we find limited evidence of attacks on deployed GenAI systems in the wild. Specifically, we document only two real-world instances of compromise, the goals of which were to prevent the unauthorised scraping of copyrighted materials,[12] and provide users with the ability to generate uncensored content.[13] While we find limited evidence of targeted attacks reported in the news, it's important to note that jailbreaking and other model attacks might be occurring, often without widespread publicity. Therefore, the actual number of real-world compromises may be higher than the two instances documented here.

[12] This new data poisoning tool lets artists fight back against generative AI
[13] We Tested Out The Uncensored Chatbot FreedomGPT

Table 4 | Count of tactics employed per goal

	Goal													
Tactic	Research	Opinion Manipulation	Monetizat' & Profit	Scam & Fraud	Harrass-ment	Reach	Unclear	Parody	Terrorism & Extremism	Child Abuse	Cyber-attacks	Hate	Subversion	Civil Unrest
Exploitation of GenAI capabilities														
Impersonation	2	20	2	22	3	1		4						1
Scaling & Amplification	2	5	14	7		2			1		2	1		
Sockpuppeting	2	12	8	4		2			3					
Falsification	2	14	9	1			2					1		
NCII		1	11	2	8									
Appropriated Likeness	1	12	1	3	1		2							
IP Infringement			4	2		3	2							
CSAM			2		4					3				
Targeting/ Personalisation		2		2							1			
Counterfeit				2		1								
Compromise of GenAI systems														
Prompt injection	11													
Jailbreaking	4													
Poisoning	3												1	
Training data exfiltration	3													
Adversarial inputs	2													
Privacy compromise	2													
Model diversion	1												1	
Steganography	1													
Model extraction	1													
Total (n)	37	66	51	45	16	9	6	4	4	3	3	2	2	1
Total (%)	14.9%	26.5%	20.5%	18.1%	6.4%	3.6%	2.4%	1.6%	1.6%	1.2%	1.2%	0.8%	0.8%	0.4%

4.2.1. Opinion Manipulation

The most common goal for exploiting GenAI capabilities during this period was **to shape or influence public opinion** (27% of all reported cases). In those instances, we saw actors deploy a range of tactics to distort the public's perception of political realities. These included impersonating public figures, using synthetic digital personas to simulate grassroots support for or against a cause ('astroturfing') and creating falsified media.

The majority of cases in our dataset involved the generation of emotionally charged synthetic images around politically divisive topics, such as war, societal unrest or economic decline. For example, images and ads shared during electoral campaigns in the US, Canada and New Zealand by party staffers[14] and state-sponsored actors[15] alike frequently depicted scenes of urban decay, homelessness and insecurity. Purportedly 'leaked' AI-generated videos and audio clips of politicians falsely endorsing controversial political positions — such as Vladimir Putin declaring martial law after Ukrainian forces entered Russian territory[16] — and privately attacking their political opponents were also common.[17] These uses of GenAI are clustered under the 'Disinformation' label in Figure 2 below.

[14] See for e.g. House GOP campaign arm slams Democrats in new AI-generated ad turning national parks into migrant tent cities; A.I.'s Use in Elections Sets Off a Scramble for Guardrails

[15] China Sows Disinformation About Hawaii Fires Using New Techniques; Chinese Influence Campaign Pushes Disunity Before U.S. Election, Study Says

[16] 'Fake Putin' announces Russia under attack as Ukraine goes on offensive

[17] See for e.g. Trolls in Slovakian Election Tap AI Deepfakes to Spread Disinfo

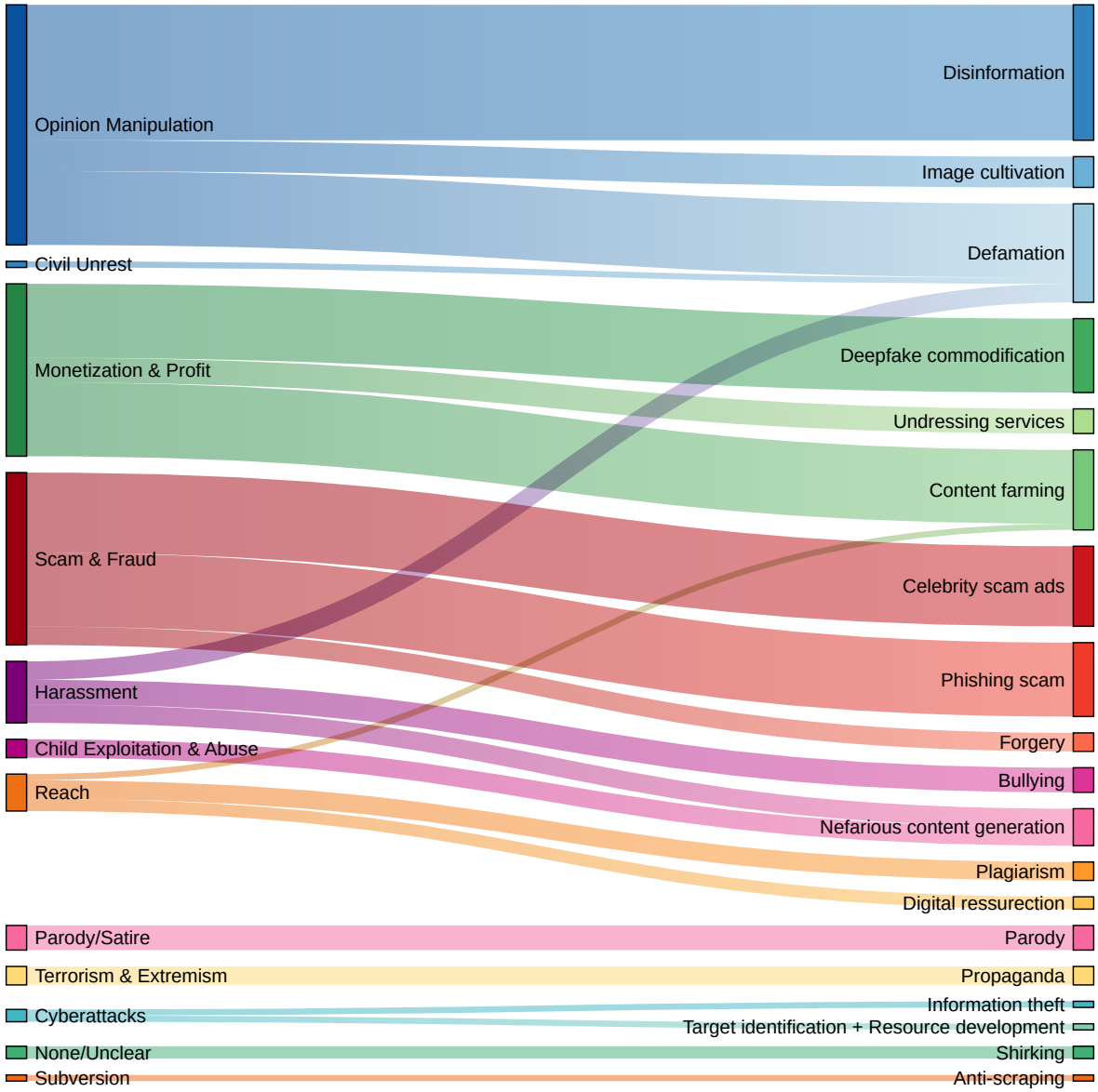

Figure 2 | Top strategies associated with each misuse goal.

Defamation was another central strategy for opinion manipulation, with GenAI tools frequently used to impersonate political figures or dissidents and portray them in compromising situations that undermine their reputation or public standing. Specific instances from our data involved depicting electoral candidates spouting abuse towards protected groups, party staffers, or their own constituents.[18] In other cases, actors shared AI-generated images of politicians appearing visibly aged to make them look unfit for leadership,[19] and showing them in intimate settings with other public figures.[20]

An emerging, though less prevalent, trend was the undisclosed use of AI-generated media by political candidates and their supporters to construct a positive public image. In one case, for example,

[18] See for e.g. Deepfake audio of Sir Keir Starmer released on first day of Labour conference; Deepfaking it: America's 2024 election collides with AI boom

[19] Viral video clip of Imran Khan looking older is AI-generated

[20] Ron DeSantis ad uses AI-generated photos of Trump, Fauci

consultants hired by the team of a Philadelphia sheriff used GenAI to fabricate positive news stories for her campaign website,[21] while the campaign staff of one of Argentina's presidential candidates leveraged GenAI to bolster his image as a strong and charismatic leader, through immersive videos and altered images. These portrayals included depicting him as a soldier in battle and emulating visual styles of Soviet-era propaganda.[22] Finally, a handful of cases involved political actors using GenAI for hypertargeted political outreach, such as simulating their own voice with high fidelity to reach out to their constituents in their native languages, or deploy GenAI-powered campaign callers to engage in tailored conversations with voters on key issues.[23] In all of these cases, the lack of appropriate disclosure around the use of GenAI tools in the context of campaigning risks misleading users and causing harm through deception.

4.2.2. Monetization & Profit

The second most common goal behind GenAI misuse was **to monetize products and services** (21% of reported cases, see Table 4). Profit-driven actors tended to leverage a wide array of tactics, including content scaling, amplification and falsification.

Content farming — the generation of vast quantities of content at scale — was a prevalent strategy observed in relation to this goal. This strategy primarily involved private users and, at times, small corporations churning out low quality AI-generated articles, books and product ads for placement on websites such as Amazon and Etsy to cut costs and capitalise on advertising revenue.[24] Aside from this profit-making motive, it is important to note that content farms are also commonly used by state-sponsored actors to flood the information space with false or misleading information.[25]

The creation of non-consensual intimate imagery (NCII) also represented a significant portion of monetization-driven misuse. In nearly all of these misuse cases, image- and video-generation tools were used to create and sell sexually explicit videos of celebrities who did not consent to the production of that content, or to "nudify" them as a paid service ('undressing services').[26]

4.2.3. Scam & Fraud

Third, actors leveraged GenAI to engage in **scams and fraudulent activities** such as stealing information, money or other assets (18%). Fraud-motivated misuses tended to leverage the power of real identities to deceive victims. While identity-based fraud is a long-standing issue, the photorealism and sophistication of GenAI outputs has empowered malicious actors to create more personalised and highly persuasive scams.

Celebrity scam ads, for example, which involve impersonating influential figures to promote fraudulent crypto and investment schemes, were prominent in our dataset. Another common strategy involved using AI-generated audio or video to impersonate trusted individuals, such as a loved one or senior colleague, in a bid to extort money from victims ('Phishing scams' in Figure 2). We also observed the use of GenAI to create bespoke business email compromise campaigns or to convincingly imitate an organisation's trademark or logo to boost the believability of phishing emails.[27]

[21] Philly Sheriff's Campaign Takes Down Bogus 'News' Stories Posted to Site That Were Generated by AI
[22] Is Argentina the First A.I. Election?
[23] Meet Ashley, the world's first AI-powered political campaign caller
[24] See for e.g. A New Frontier for Travel Scammers: A.I.-Generated Guidebooks; Rise of the Newsbots: AI-Generated News Websites Proliferating Online
[25] Chinese Influence Operations Evolve in Campaigns Targeting Taiwanese Elections, Hong Kong Protests
[26] See for e.g. 'Nudify' Apps That Use AI to 'Undress' Women in Photos Are Soaring in Popularity
[27] Generative AI making it harder to spot fraudulent emails

These misuse tactics not only infringe upon the targeted individual or organisation's rights and reputation, but also inflict a potentially high financial and psychological cost to victims. One of the cases in our dataset, for example, saw a financial worker being tricked into transferring $25m to scammers, who had used GenAI to impersonate several of the employee's co-workers on a video call.[28]

4.2.4. Harassment

Fourth, approximately 6% of observed cases of GenAI misuse involved some form of harassment or intimidation, with a distinctly gendered dimension. Here, most cases in our dataset centred on the non-consensual generation of NCII (non-consensual intimate imagery) targeting both adults and adolescents, all of which were female. A prevalent and disturbing trend in that respect was the creation and sharing of AI-generated nudes of high school students as part of bullying campaigns. Journalists and celebrities were also common targets of this type of abuse.

Beyond the defamation cases discussed in Section 4.2.1 ('Opinion Manipulation'), outside the political domain, we also noted several instances involving the unauthorised cloning of public figures' voices and likeness for abuse (including high school principals and celebrities) or as an attempt to defame them.[29] A rarely observed but seemingly novel form of harassment involved malicious actors generating audio clips of voice actors doxxing themselves (i.e. reading aloud their own addresses) and sharing them online. This case, and others mentioned above, highlight a growing and concerning risk of targeted abuse for content creators and individuals whose data is easily or publicly accessible.

4.2.5. Reach

Finally, a comparatively small proportion of cases involved using GenAI to maximise the reach of a message, piece of content, or brand (3.6%). Though still marginal, one noteworthy and emerging strategy in relation to this goal was the rise of digital resurrections for reach and advocacy. We saw multiple cases of GenAI being used to recreate the likeness of individuals who had passed away, sometimes without the consent of those involved. In Feb 2024, two activist groups used GenAI to create voice recordings of school shooting victims in an attempt to compel Congress to act on gun violence.[30] In Aug 2023, content creators on TikTok used GenAI to "give voices" to deceased or missing children to narrate disturbing details of their experiences in an attempt to "raise awareness."[31]

While not necessarily violative of a model's content policies, practices such as these raise profound ethical concerns by instrumentalizing the likeness of people who cannot express consent or lack agency over how their image is utilised.

5. Discussion

Our analysis of real-world GenAI misuse highlights key patterns with significant implications for trust and safety practitioners, policy makers and researchers.

Our data shows that GenAI tools are primarily exploited to **manipulate human likeness** (through Impersonation, Sockpuppeting, Appropriated Likeness and NCII) and **falsify evidence**. The prevalence of these tactics may be due to the fact that sources of human data (e.g. images, audio, video) abound

[28] Finance worker pays out $25 million after video call with deepfake 'chief financial officer'

[29] See for e.g. AI-Generated Voice Firm Clamps Down After 4chan Makes Celebrity Voices for Abuse; High Schoolers Made a Racist Deepfake of a Principal Threatening Black Students

[30] Voices of the dead: shooting victims plead for gun reform with AI-voice messages

[31] AI is being used to give dead, missing kids a voice they didn't ask for

online, making it relatively easy for bad-faith actors to feed this information into generative AI systems. However, it is also possible that these types of misuses simply attract more media attention than others, due to their broad societal impact. These cases of misuse primarily aimed to shape public opinion, especially through defamation and manipulation of political perceptions, and to facilitate scams, fraud and quick monetization schemes.

Despite widespread concerns around highly sophisticated, state-sponsored uses of GenAI[32], we find that **most cases of GenAI misuse are not sophisticated attacks on AI systems** but readily exploit easily accessible GenAI capabilities that require minimal technical expertise. Many of the salient tactics we observe, such as impersonation scams, forgery, and synthetic personas, pre-date the invention of GenAI and have long been used to influence the information ecosystem and manipulate others. However, by giving these age-old tactics new potency and democratising access, **GenAI has altered the costs and incentives associated with information manipulation**, leading to a wide range of use cases and a wide pool of individuals being involved in these activities. As our data shows, this includes political figures and private citizens alike, and those without significant technical background.

The widespread availability, accessibility and hyperrealism of GenAI outputs across modalities has also enabled **new, lower-level forms of misuse that blur the lines between authentic presentation and deception**. While these uses of GenAI — such as generating and repurposing content at scale and leveraging GenAI for personalised political communication — are often **neither overtly malicious nor explicitly violate these tools' content policies or terms of services**, their potential for harm is significant. GenAI-powered political image cultivation and advocacy without appropriate disclosure, for example, undermines public trust by making it difficult to distinguish between genuine and manufactured portrayals. Likewise, the mass production of low quality, spam-like and nefarious synthetic content[33] risks increasing people's scepticism towards digital information altogether and overloading users with verification tasks. If unaddressed, this contamination of publicly accessible data with AI-generated content could potentially impede information retrieval and distort collective understanding of socio-political reality or scientific consensus. For example, we are already seeing cases of liar's dividend,[34] where high profile individuals are able to explain away unfavourable evidence as AI-generated, shifting the burden of proof in costly and inefficient ways.

These findings carry several consequences for how we approach mitigations. Common misuse tactics such as NCII are exacerbated by technical vulnerabilities within GenAI systems that model developers are actively working to resolve and create safeguards against, such as removing toxic content from training data or restricting prompts that violate these tools' terms of services. However, many of the cases identified (e.g. those involving deceptive portrayals) prey on vulnerabilities in the broader social context in which they are deployed — for example, phishing scams campaigns that rely on an individuals' reasonable expectation of the authenticity of their digital landscape and their interactions with it. While technical interventions may provide some benefit, in these cases, non-technical, user-facing interventions are necessary. Prebunking, for example — a common psychological intervention to protect against information manipulation (Roozenbeek et al., 2022) — could be usefully extended to protect users against GenAI-enabled deceptive and manipulative tactics.

Additionally, many prevalent forms of misuse hinge on exploiting GenAI capabilities that model developers are actively working to enhance (for example, photo-realistic outputs). As GenAI tools

[32] See for e.g. Safety and Security Risks of Generative Artificial Intelligence to 2025; Propaganda, foreign interference, and generative AI

[33] See for e.g. Re: Artificial Intelligence and the Exploitation of Children

[34] See for e.g. An Indian politician says scandalous audio clips are AI deepfakes. We had them tested

become more capable and accessible, we may therefore continue to see an increase in AI-generated content as part of media-based misinformation and manipulation campaigns (Dufour et al., 2024). While several solutions like synthetic media detection tools and watermarking techniques have been proposed and offer promise, they are far from panaceas (Sadasivan et al., 2023). Notably, the inherent adaptability of malicious actors means that as detection methods improve, so will methods of circumvention (Leibowicz, McGregor, and Ovadya, 2021). In these cases, targeted interventions such as restrictions on specific model capabilities and usage restrictions may be warranted when the risk for misuse is high and other interventions are insufficient (Anderljung and Hazell, 2023; Shevlane, 2022).

6. Limitations and further research

While this research provides valuable insights into the current landscape of GenAI misuse, it is important to acknowledge several limitations that may affect the generalisability and comprehensiveness of our findings.

First, relying on media reports as a primary data source can introduce biases. Media outlets often prioritise incidents with sensational elements or those that directly impact human perception, potentially skewing our dataset towards particular types of misuse. Conversely, covert attacks or those that do not necessarily include humans in-the-loop — such as the use of GenAI to obfuscate malicious code to evade filters — may be underrepresented in our data due to limited media coverage or to the fact that companies may keep this information private. This underscores the need for better and more comprehensive sources of anonymized data — akin to the Safety Information Analysis and Sharing (ASIAS) System for the aviation industry for example — to gain a more holistic understanding of the threat landscape and inform effective mitigations.

Second, our analysis is time-bound and therefore only offers a snapshot of GenAI misuse at a specific point in time. Yet, as GenAI models continue to acquire new capabilities, become more agentic and integrated into everyday applications and services, their potential for misuse may expand beyond the current scope of our findings. For example, the progressive integration of GenAI into social media platforms to deliver personalised content may lead to new forms of information manipulation. Keeping up with this dynamic landscape calls for further longitudinal analyses and continued monitoring of emerging tactics and harms.

Finally, we note that the majority of observed cases of misuse in our dataset involve models that take text prompts as input rather than leveraging truly multimodal capabilities. This is possibly due to the fact that we are analysing data at an early stage in this technology's development. However, the field is rapidly progressing towards more sophisticated multimodal models capable of processing and generating diverse forms of content. We anticipate that new modalities and capabilities, such as the ability to prompt models with video and image input will likely lead to new patterns of misuse, which should be thoroughly investigated.

7. Conclusion

This research has sought to illuminate the evolving landscape of GenAI misuse, and its impacts. While fears of sophisticated adversarial attacks have dominated public discourse, our findings reveal a prevalence of low-tech, easily accessible misuses by a broad range of actors, often driven by financial or reputational gain. These misuses, while not always overtly malicious, have far-reaching consequences for trust, authenticity, and the integrity of information ecosystems. We have also seen how GenAI amplifies existing threats by lowering barriers to entry and increasing the potency and accessibility

of previously costly tactics. These findings underscore the need for a multi-faceted approach to mitigating GenAI misuse, involving collaboration between policymakers, researchers, industry leaders, and civil society. Addressing this challenge requires not only technical advancements but also a deeper understanding of the social and psychological factors that contribute to the misuse of these powerful tools.

References

Markus Anderljung and Julian Hazell. Protecting society from AI misuse: When are restrictions on capabilities warranted? March 2023. doi: 10.48550/arXiv.2303.09377. URL http://arxiv.org/abs/2303.09377.

Anthropic. Introducing the next generation of claude. Technical report, Anthropic, March 2024. URL https://www.anthropic.com/news/claude-3-family.

Clark Barrett, Brad Boyd, Elie Bursztein, Nicholas Carlini, Brad Chen, Jihye Choi, Amrita Roy Chowdhury, Mihai Christodorescu, Anupam Datta, Soheil Feizi, Kathleen Fisher, Tatsunori Hashimoto, Dan Hendrycks, Somesh Jha, Daniel Kang, Florian Kerschbaum, Eric Mitchell, John Mitchell, Zulfikar Ramzan, Khawaja Shams, Dawn Song, Ankur Taly, and Diyi Yang. Identifying and mitigating the security risks of generative AI. *Foundations and Trends® in Privacy and Security*, 6(1):1–52, 2023. ISSN 2474-1558. doi: 10.1561/3300000041. URL http://doi.org/10.1561/3300000041.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Wang Jianfeng, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, Wesam Manassra, Prafulla Dhariwal, Casey Chu, Yunxin Jiao, and Aditya Ramesh. Improving image generation with better captions. October 2023. URL https://cdn.openai.com/papers/dall-e-3.pdf.

Taís Fernanda Blauth, Oskar Josef Gstrein, and Andrej Zwitter. Artificial intelligence crime: An overview of malicious use and abuse of AI. *IEEE Access*, 10:77110–77122, 2022. ISSN 2169-3536. doi: 10.1109/ACCESS.2022.3191790. URL http://doi.org/10.1109/ACCESS.2022.3191790.

Jonathan Bright, Florence E Enock, Saba Esnaashari, John Francis, Youmna Hashem, and Deborah Morgan. Generative AI is already widespread in the public sector. January 2024. doi: 10.48550/arXiv.2401.01291. URL http://arxiv.org/abs/2401.01291.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. Technical report, OpenAI, February 2024. URL https://openai.com/research/video-generation-models-as-world-simulators.

Miles Brundage, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, Paul Scharre, Thomas Zeitzoff, Bobby Filar, Hyrum Anderson, Heather Roff, Gregory C Allen, Jacob Steinhardt, Carrick Flynn, Seán Ó hÉigeartaigh, Simon Beard, Haydn Belfield, Sebastian Farquhar, Clare Lyle, Rebecca Crootof, Owain Evans, Michael Page, Joanna Bryson, Roman Yampolskiy, and Dario Amodei. The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. February 2018. doi: 10.48550/arXiv.1802.07228. URL http://arxiv.org/abs/1802.07228.

Nicholas Carlini, Matthew Jagielski, Christopher A Choquette-Choo, Daniel Paleka, Will Pearce, Hyrum Anderson, Andreas Terzis, Kurt Thomas, and Florian Tramèr. Poisoning Web-Scale training datasets is practical. February 2023. doi: 10.48550/arXiv.2302.10149. URL http://arxiv.org/abs/2302.10149.

Nicholas Carlini, Daniel Paleka, Krishnamurthy Dj Dvijotham, Thomas Steinke, Jonathan Hayase, A Feder Cooper, Katherine Lee, Matthew Jagielski, Milad Nasr, Arthur Conmy, Eric Wallace, David Rolnick, and Florian Tramèr. Stealing part of a production language model. March 2024. doi: 10.48550/arXiv.2403.06634. URL http://arxiv.org/abs/2403.06634.

Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J Pappas, and Eric Wong. Jailbreaking black box large language models in twenty queries. October 2023. doi: 10.48550/arXiv.2310.08419. URL http://arxiv.org/abs/2310.08419.

Di Cooke, Abigail Edwards, Sophia Barkoff, and Kathryn Kelly. As good as a coin toss: Human detection of AI-generated images, videos, audio, and audiovisual stimuli. March 2024. doi: 10.48550/arXiv.2403.16760. URL http://arxiv.org/abs/2403.16760.

Renee DiResta and Josh A Goldstein. How spammers and scammers leverage AI-Generated images on facebook for audience growth. March 2024. doi: 10.48550/arXiv.2403.12838. URL http://arxiv.org/abs/2403.12838.

Nicholas Dufour, Arkanath Pathak, Pouya Samangouei, Nikki Hariri, Shashi Deshetti, Andrew Dudfield, Christopher Guess, Pablo Hernández Escayola, Bobby Tran, Mevan Babakar, et al. Ammeba: A large-scale survey and dataset of media-based misinformation in-the-wild. 2024. doi: 10.48550/arXiv.2405.11697. URL http://arxiv.org/abs/2405.11697.

Emilio Ferrara. GenAI against humanity: nefarious applications of generative artificial intelligence and large language models. *Journal of Computational Social Science*, February 2024. ISSN 2432-2725. doi: 10.1007/s42001-024-00250-1. URL https://doi.org/10.1007/s42001-024-00250-1.

Stefan Feuerriegel, Renée DiResta, Josh A Goldstein, Srijan Kumar, Philipp Lorenz-Spreen, Michael Tomz, and Nicolas Pröllochs. Research can help to tackle AI-generated disinformation. *Nat Hum Behav*, 7(11):1818–1821, November 2023. ISSN 2397-3374. doi: 10.1038/s41562-023-01726-2. URL http://doi.org/10.1038/s41562-023-01726-2.

G Gartlehner, L Kahwati, R Hilscher, I Thomas, S Kugley, K Crotty, M Viswanathan, B Nussbaumer-Streit, G Booth, N Erskine, A Konet, and R Chew. Data extraction for evidence synthesis using a large language model: A proof-of-concept study. October 2023. doi: 10.1101/2023.10.02.23296415. URL https://www.medrxiv.org/content/10.1101/2023.10.02.23296415v1.

2024 Gemini Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. March 2024. doi: 10.48550/arXiv.2403.05530. URL http://arxiv.org/abs/2403.05530.

Abenezer Golda, Kidus Mekonen, Amit Pandey, Anushka Singh, Vikas Hassija, Vinay Chamola, and Biplab Sikdar. Privacy and security concerns in generative AI: A comprehensive survey. *IEEE Access*, 12:48126–48144, 2024. ISSN 2169-3536. doi: 10.1109/ACCESS.2024.3381611. URL http://doi.org/10.1109/ACCESS.2024.3381611.

Josh A Goldstein, Girish Sastry, Micah Musser, Renee DiResta, Matthew Gentzel, and Katerina Sedova. Generative language models and automated influence operations: Emerging threats and potential mitigations. January 2023. doi: 10.48550/arXiv.2301.04246. URL http://arxiv.org/abs/2301.04246.

Robert Gorwa and Douglas Guilbeault. Unpacking the social media bot: A typology to guide research and policy. *Policy Internet*, 12(2):225–248, June 2020. ISSN 2194-6019, 1944-2866. doi: 10.1002/poi3.184. URL http://doi.org/10.1002/poi3.184.

Keith Raymond Harris. Liars and trolls and bots online: The problem of fake persons. *Philos. Technol.*, 36(2):35, May 2023. ISSN 2210-5433. doi: 10.1007/s13347-023-00640-9. URL https://doi.org/10.1007/s13347-023-00640-9.

Jie Huang and Kevin Chen-Chuan Chang. Towards reasoning in large language models: A survey. December 2022. doi: 10.48550/arXiv.2212.10403. URL http://arxiv.org/abs/2212.10403.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12):1–38, March 2023. ISSN 0360-0300. doi: 10.1145/3571730. URL https://doi.org/10.1145/3571730.

Claire Leibowicz, Sean McGregor, and Aviv Ovadya. The deepfake detection dilemma: A multistakeholder exploration of adversarial dynamics in synthetic media. February 2021. doi: 10.48550/arXiv.2102.06109. URL http://arxiv.org/abs/2102.06109.

Zilong Lin, Jian Cui, Xiaojing Liao, and Xiaofeng Wang. Malla: Demystifying real-world large language model integrated malicious services. January 2024. doi: 10.48550/arXiv.2401.03315. URL http://arxiv.org/abs/2401.03315.

Yi Liu, Gelei Deng, Zhengzi Xu, Yuekang Li, Yaowen Zheng, Ying Zhang, Lida Zhao, Tianwei Zhang, Kailong Wang, and Yang Liu. Jailbreaking ChatGPT via prompt engineering: An empirical study. May 2023. doi: 10.48550/arXiv.2305.13860. URL http://arxiv.org/abs/2305.13860.

Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. RePaint: Inpainting using denoising diffusion probabilistic models. January 2022. doi: 10.48550/arXiv.2201.09865. URL http://arxiv.org/abs/2201.09865.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.173. URL http://doi.org/10.18653/v1/2020.acl-main.173.

Sophie J Nightingale and Hany Farid. AI-synthesized faces are indistinguishable from real faces and more trustworthy. *Proc. Natl. Acad. Sci. U. S. A.*, 119(8), February 2022. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2120481119. URL http://doi.org/10.1073/pnas.2120481119.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook

Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, C J Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. GPT-4 technical report. March 2023. doi: 10.48550/arXiv.2303.08774. URL http://arxiv.org/abs/2303.08774.

Jintro Pauly. Technology: Disconnecting the gordian node. In Tobias Bunde, Sophie Eisentraut, and Leonard Schütte, editor, *Munich Security Report 2024*, pages 95–101. February 2024. doi: 10.47342/BMQK9457. URL https://securityconference.org/en/publications/munich-security-report-2024/technology/.

Fábio Perez and Ian Ribeiro. Ignore previous prompt: Attack techniques for language models. November 2022. doi: 10.48550/arXiv.2211.09527. URL http://arxiv.org/abs/2211.09527.

Junaid Qadir. Engineering education in the era of ChatGPT: Promise and pitfalls of generative AI for education. In *2023 IEEE Global Engineering Education Conference (EDUCON)*, pages 1–9. IEEE, May 2023. doi: 10.1109/EDUCON54358.2023.10125121. URL http://dx.doi.org/10.1109/EDUCON54358.2023.10125121.

Mikel Rodriguez, Andrew Trask, Vijay Bolina, Geoff Keeling, and Iason Gabriel. Malicious uses. In Iason Gabriel, Arianna Manzini, Geoff Keeling, Lisa Anne Hendricks, Verena Rieser, Hasan Iqbal, Nenad Tomašev, Ira Ktena, Zachary Kenton, Mikel Rodriguez, Seliem El-Sayed, Sasha Brown, Canfer Akbulut, Andrew Trask, Edward Hughes, A Stevie Bergman, Renee Shelby, Nahema Marchal, Conor Griffin, Juan Mateos-Garcia, Laura Weidinger, Winnie Street, Benjamin Lange, Alex Ingerman, Alison Lentz, Reed Enger, Andrew Barakat, Victoria Krakovna, John Oliver Siy, Zeb Kurth-Nelson, Amanda McCroskery, Vijay Bolina, Harry Law, Murray Shanahan, Lize Alberts, Borja Balle, Sarah de Haas, Yetunde Ibitoye, Allan Dafoe, Beth Goldberg, Sébastien Krier, Alexander Reese, Sims Witherspoon, Will Hawkins, Maribeth Rauh, Don Wallace, Matija Franklin, Josh A Goldstein, Joel

Lehman, Michael Klenk, Shannon Vallor, Courtney Biles, Meredith Ringel Morris, Helen King, Blaise Agüera y Arcas, William Isaac, and James Manyika, editors, *The Ethics of Advanced AI Assistants*. April 2024. doi: 10.48550/arXiv.2404.16244. URL http://arxiv.org/abs/2404.16244.

Jon Roozenbeek, Sander van der Linden, Beth Goldberg, Steve Rathje, and Stephan Lewandowsky. Psychological inoculation improves resilience against misinformation on social media. *Sci Adv*, 8(34):eabo6254, August 2022. ISSN 2375-2548. doi: 10.1126/sciadv.abo6254. URL http://doi.org/10.1126/sciadv.abo6254.

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. Can AI-Generated text be reliably detected? March 2023. doi: 10.48550/arXiv.2303.11156. URL http://arxiv.org/abs/2303.11156.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. February 2023. doi: 10.48550/arXiv.2302.04761. URL http://arxiv.org/abs/2302.04761.

Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. "do anything now": Characterizing and evaluating In-The-Wild jailbreak prompts on large language models. August 2023. doi: 10.48550/arXiv.2308.03825. URL http://arxiv.org/abs/2308.03825.

Toby Shevlane. Structured access: an emerging paradigm for safe AI deployment. January 2022. doi: 10.48550/arXiv.2201.05159. URL http://arxiv.org/abs/2201.05159.

Toby Shevlane, Sebastian Farquhar, Ben Garfinkel, Mary Phuong, Jess Whittlestone, Jade Leung, Daniel Kokotajlo, Nahema Marchal, Markus Anderljung, Noam Kolt, Lewis Ho, Divya Siddarth, Shahar Avin, Will Hawkins, Been Kim, Iason Gabriel, Vijay Bolina, Jack Clark, Yoshua Bengio, Paul Christiano, and Allan Dafoe. Model evaluation for extreme risks. May 2023. doi: 10.48550/arXiv.2305.15324. URL http://arxiv.org/abs/2305.15324.

S S Sundar. The MAIN model: A heuristic approach to understanding technology effects on credibility. In Miriam J Metzger and Andrew J Flanagin, editors, *Digital Media, Youth, and Credibility*, pages 73–100. MIT Press, 2008. ISBN 9780262062732. URL https://library.oapen.org/handle/20.500.12657/26088.

Bakhtiyar Syed, Gaurav Verma, Balaji Vasan Srinivasan, Anandhavelu Natarajan, and Vasudeva Varma. Adapting language models for Non-Parallel Author-Stylized rewriting. *AAAI*, 34(05): 9008–9015, April 2020. ISSN 2374-3468, 2374-3468. doi: 10.1609/aaai.v34i05.6433. URL https://ojs.aaai.org/index.php/AAAI/article/view/6433.

Haoye Tian, Weiqi Lu, Tsz On Li, Xunzhu Tang, Shing-Chi Cheung, Jacques Klein, and Tegawendé F Bissyandé. Is ChatGPT the ultimate programming assistant – how far is it? April 2023. doi: 10.48550/arXiv.2304.11938. URL http://arxiv.org/abs/2304.11938.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael

Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and Fine-Tuned chat models. July 2023. doi: 10.48550/arXiv.2307.09288. URL http://arxiv.org/abs/2307.09288.

Cristian Vaccari and Andrew Chadwick. Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media + Society*, 6(1): 2056305120903408, January 2020. ISSN 2056-3051. doi: 10.1177/2056305120903408. URL https://doi.org/10.1177/2056305120903408.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. Universal adversarial triggers for attacking and analyzing NLP. August 2019. doi: 10.48550/arXiv.1908.07125. URL http://arxiv.org/abs/1908.07125.

Zhendong Wang, Yifan Jiang, Yadong Lu, Yelong Shen, Pengcheng He, Weizhu Chen, Zhangyang Wang, and Mingyuan Zhou. In-Context learning unlocked for diffusion models. May 2023. doi: 10.48550/arXiv.2305.01115. URL http://arxiv.org/abs/2305.01115.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. June 2022. doi: 10.48550/arXiv.2206.07682. URL http://arxiv.org/abs/2206.07682.

Wen Yang, Chong Li, Jiajun Zhang, and Chengqing Zong. BigTranslate: Augmenting large language models with multilingual translation capability over 100 languages. May 2023. doi: 10.48550/arXiv.2305.18098. URL http://arxiv.org/abs/2305.18098.

Dobin Yim, Jiban Khuntia, Vijaya Parameswaran, and Arlen Meyers. Preliminary evidence of the use of generative AI in health care clinical services: Systematic narrative review. *JMIR Med Inform*, 12: e52073, March 2024. ISSN 2291-9694. doi: 10.2196/52073. URL http://doi.org/10.2196/52073.

Appendices

A. Goals

Analysing our dataset, we have identified 16 distinct goals driving GenAI misuse (see Table A.1 below).

Table A.1 | Goals of GenAI misuse

	Goal	Definition
Adversarial	Scam & Fraud	Access or steal money, information, or other assets (property) from individuals or organisations
	Market Manipulation	Control or artificially affect the price and demand of securities
	Terrorism & Extremism	Facilitate, promote or glorify terrorism or extremist causes.
	Cyberattacks	Conduct cyberattacks and disruption of technical systems and networks
	Civil Unrest	Encourage defiance of the law or political violence.
	Surveillance	Facilitate surveillance of specific individuals or organisations.
	Child Exploitation & Abuse	Exploit, abuse or harm children
	Hate	Disparage and dehumanise individual/s based on their membership to a protected group, or encourage discrimination/hatred.
	Harassment	Intimidate, pressure, annoy, or upset an individual (e.g. doxxing)
Not inherently adversarial	Opinion Manipulation	Influence or shape individual or public opinion.
	Monetization & Profit	Drive monetization or maximise a business's profitability
	Reach	Maximise the number of individuals exposed to a message, piece of content, or brand.
	Subversion	Oppose or undermine the advantage held by powerful people or institutions.
	Parody/Satire	Imitate someone in an exaggerated way for comedic effect or criticism, using humour or satire.
Benevolent	Research	Research or conduct testing on malicious attacks and model vulnerabilities
	Unclear	Unclear or outside other listed goal category

Importantly, not all goals associated with misuse are necessarily adversarial or involve malicious intent. Some goals are, by definition, malicious in that they involve engaging in explicitly illegal activities or demonstrate a clear intent to cause harm. This includes leveraging GenAI models to

engage in fraudulent activities such as stealing information, money or other assets (**Scam & Fraud**), encouraging political violence and defiance of the law (**Civil Unrest**), spying on individuals' online activities and data (**Surveillance**), conducting cyberattacks, or promoting terrorism and other extreme ideologies (**Terrorism & Extremism**). Other goals that fit in this category include exploiting GenAI tools to attack, dehumanise or disparage protected groups (**Hate**), harass individuals (**Harassment**) and exploit or harm children (**Child Exploitation and Abuse**).

In contrast, some goals may be pursued without an explicit intention to cause harm, but could still result in adverse consequences for individuals and society. This includes cases where actors use GenAI to try to influence public opinion about socio-political issues (**Opinion Manipulation**), to sell products and services or to maximise a business profitability (**Monetization & Profit**), or to manipulate the price and demand of stocks and securities (**Market Manipulation**). Other relevant goals include leveraging GenAI to aid the distribution of specific messages or content (**Reach**), to oppose or undermine power structures (**Subversion**), or for satirical purposes (**Parody**).

Finally, we observe several instances where generative AI is only misused for ostensibly benevolent purposes, as part of research efforts and testing aimed at exposing model vulnerabilities (**Research**). Of course, it is possible for any given case of misuse to have more than one goal at a time. For simplicity and clarity, in our dataset, we track only what we believed to be the primary goal of each case of misuse, based on the contextual information provided by media reports.

B. Strategies

Our observations also reveal distinct combinations of goals, tactics, uses of GenAI and targets of misuse coalescing into broader 'misuse strategies.' These strategies are useful to delineate as they reveal the calculated steps taken to leverage GenAI towards different ends, which may demand tailored interventions or mitigation strategies. In the following table, we enumerate these strategies, organised by goal, and outline the tactics they employ, along with salient examples from our dataset.

Table B.1 | Strategies of GenAI misuse per goal

Goal	Strategy	Use of GenAI	Tactic
Scam & Fraud	Celebrity scam ads	Impersonate celebrities or public figures to promote fraudulent investment scams	Impersonation
	Forgery	Forge documents to bypass identity verification	Falsification
		Generate media (songs or books) that appear to have been created by known artists and sell them as authentic	Appropriateness Likeness + Counterfeit
	Phishing scam	Generate content to run targeted phishing scams at scale (e.g. business email compromise (BEC) campaigns.)	Scale & Amplification + Targeting/Personalization
		Mimic an organisation's trademark to increase legitimacy of phishing.	Targeting/Personalization + Counterfeit
		Create fake personas to carry out romance scams at scale	Sockpuppeting + Scale & Amplification
		Impersonate a trusted individual (a loved one in distress, or a senior colleague) to steal funds	Impersonation
	Sextortion	Generate NCII of individuals from public social media photos to run sextortion schemes.	NCII + Scale & Amplification
		Impersonate authorities to blackmail individuals	Impersonation
	Information theft	Impersonate public figures to access privileged information.	Impersonation
	Malware	Create copycat websites to trick users into downloading malware.	Counterfeit
		Create fake tutorial videos to influence individuals to download stealer malware.	Sockpuppeting

(continued on the next page)

Table B.1 | Strategies of GenAI misuse per goal *(continued)*

Opinion Manipulation	Astroturfing	Create the impression of widespread grassroots support for or opposition against a cause	Sockpuppeting + Scale & Amplification
		Create the impression of popular approval of a product (covert/spammy advertising) at scale	Scale & Amplification + Sockpuppeting
	Defamation	Alter appearance of politicians to make them look older	Appropriated Likeness
		Impersonate politicians or political dissidents making abusive statements	Impersonation
		Generate media of politicians in compromising situation	Appropriated Likeness
	Digital resurrection	Impersonate deceased victims to plead for a cause (e.g. gun reform)	Impersonation
	Disinformation	Impersonate politicians falsely endorsing specific political positions or claiming electoral victory	Impersonation
		Generate false images of emotionally charged and politically divisive issues.	Falsification
		Alter the likeness of dissidents or protesters	Appropriated Likeness
		Generate images or stories of fake crisis events	Falsification
	Political outreach	Create personalised campaign robocallers on behalf of candidates.	Sockpuppeting + Targeting & Personalization
		Generate robocalls to conduct outreach to voters in their language.	Appropriated Likeness
	Image cultivation	Generate media that create a positive impression of a public figure.	Appropriated Likeness + Falsification
	Voter suppression	Impersonate politician giving misleading election information	Impersonation
	News hijacking	Interrupt news broadcasts to air AI-generated media.	Sockpuppeting + Falsification

(continued on the next page)

Table B.1 | Strategies of GenAI misuse per goal *(continued)*

Monetization & Profit	Botnet	Operate botnets to perform revenue-generating actions	Sockpuppeting + Scale & Amplification
	Content farming	Generate high volumes of fake clickbait articles to optimise ad revenue	Falsification + Scale & Amplification
		Mimic or recycle existing original content at scale by adding text and voiceovers.	IP Infringement + Scale & Amplification
	Deepfake commodification	Generate sexually explicit deepfakes of celebrities for sale	NCII
		Create and sell chatbot impersonating politicians to answer election-related questions	Impersonation
	Plagiarism	Plagiarise original content across media for monetisation	IP Infringement
	Shirking	Generate fake documentation (e.g. legal filings) to automate one's labour.	Falsification
		Create fake personas to produce product reviews.	Sockpuppeting
	Undressing services	Generate NCII of individuals as a paid service.	NCII
Cyberattacks	Target identification	Automate research and identification of high-value organisational targets and their vulnerabilities.	Scale & Amplification
	Resource development	Coding assistance for operation and automation of cyberattack-related tasks (e.g. targeted malware development)	Scale & Amplification + Targeting & Personalization
Harassment	Bullying	Generate NCII of private individuals or public figures (e.g. journalists) to bully or silence them.	NCII + CSAM
	Defamation	Generate audio and video clips of celebrities and private individuals making abusive or racist statements	Impersonation
	Doxxing	Generate audio clips of content creators reading aloud their own address.	Appropriated Likeness
Reach	Plagiarism	Plagiarise competitors' website content to maximise reach.	IP Infringement
	Digital resurrection	Create fake videos of deceased individuals narrating the events of their death.	Sockpuppeting
	Content farming	Generate bogus news articles at scale to boost website in search results.	Falsification + Scale & Amplification
Subversion	Anti-scraping	Poisoning training data to prevent copyrighted data scraping	Poisoning