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

Nahema Marchal, Rachel Xu, Rasmi Elasmar,
Iason Gabriel, Beth Goldberg, William Isaac

Google DeepMind

Presented by:

Laurel Aquino, Mohammad Abdul Basit

UNIVERSITY OF TORONTO

DEFY GRAVITY

Background

Challenge:

- Increased sophistication and capabilities of generative models result in increased potential for misuse
- Growing need for assessment of emerging risks and harms

Current Research Landscape:

- Existing research focuses on identifying potential risks, not actual misuse
- Limited insight into:
  - How GenAI tools are exploited and abused in practice
  - Motivations of the malicious actors
  - Tactics used to cause harm
- This paper presents a taxonomy of GenAI misuse tactics to fill the gap

Introduction

Analyze academic work and 200 media reports on cases of GenAI misuse

Identify novel patterns of misuse that emerge as GenAI becomes more sophisticated

Develop evidence base of real harms and better understanding of threat landscape

Methodology

Literature Review

- Reviewed academic and grey literature on malicious uses of GenAI
- Developed initial theoretical framework of misuse tactics

Data Collection and Analysis

- Collected dataset of media reports detailing real-world misuses
- Updated misuse tactic categories based on patterns

Broader Misuse Strategies

- Collected data on actors, goals, and targets

Categorization of Misuse Tactics

- Exploit GenAI capabilities
- Compromise GenAI system

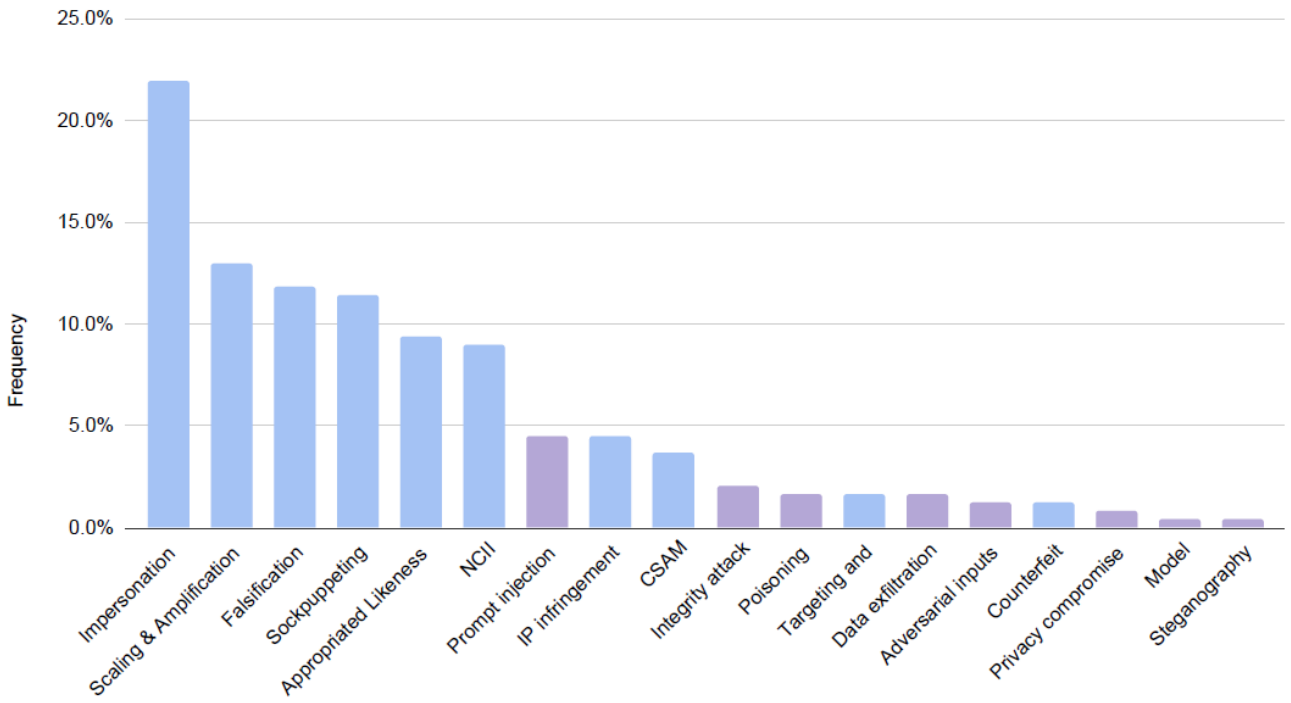
Taxonomy: Exploitation of GenAI Capabilities

	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

Taxonomy: Compromise of 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

Findings



	Modality				
Tactic	Image 🖼️	Text 📄	Audio 🔊	Video 📽️	Total
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

Goals and Motivations

Opinion Manipulation

- Shape or influence public opinion
- Defamation- impersonate popular figures in compromising situations
- Construct positive public image of a political candidate

Monetization and Profit

- Content farming- low quality AI generated articles, books and product ads for placements on Ecom websites
- Creation of non consensual intimate imagery (NCII)

Scam and Fraud

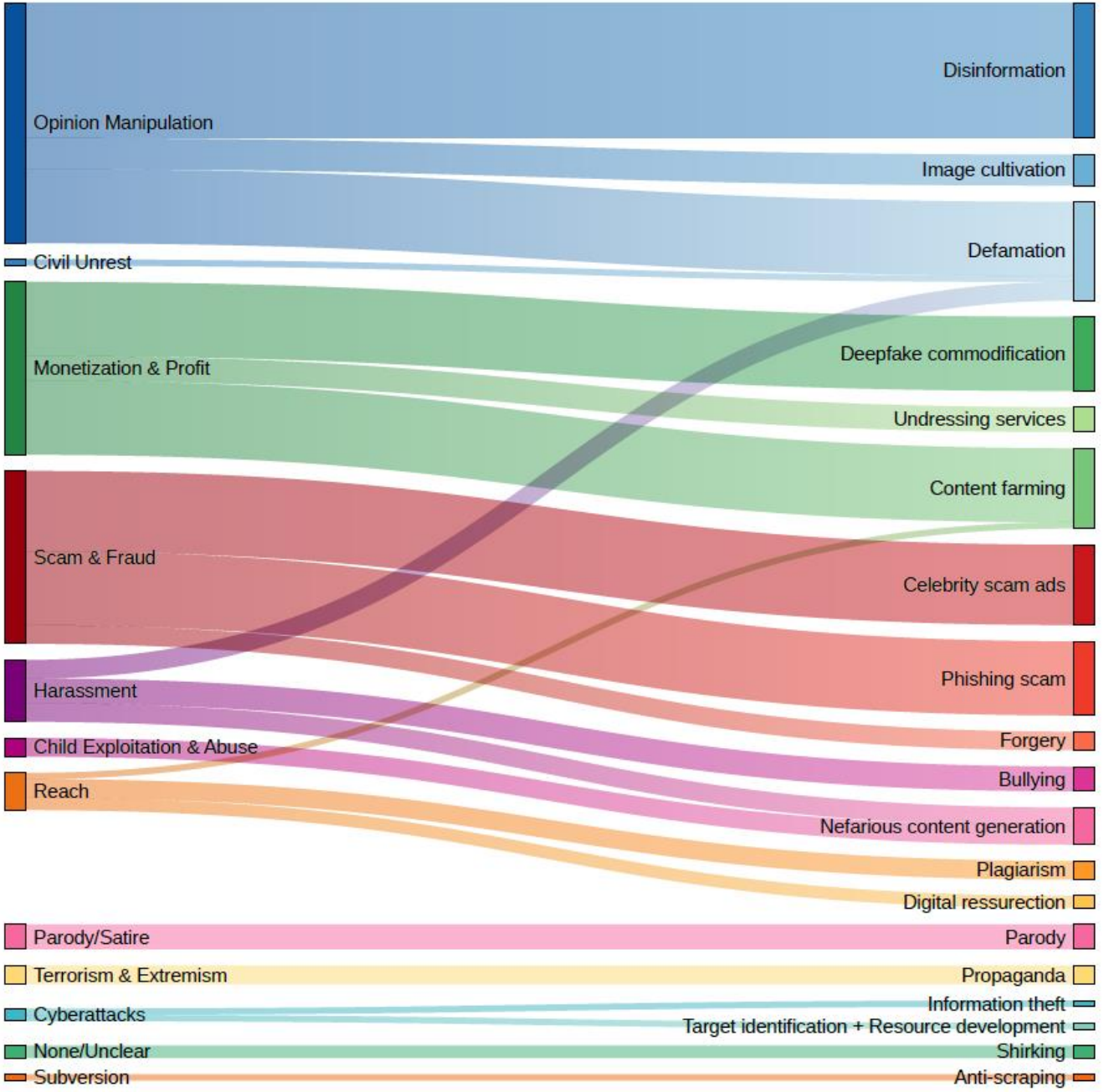
- Photorealism and sophistication of GenAI – more personalized and persuasive scams
- Celebrity scam ads, impersonate trusted individuals, bespoke business emails, imitating organization trademark/logo

Harassment

- Creation of NCII targeting adults and adolescents – all females
- Cloning public figure voices and likeness for abuse

Reach

- Use GenAI to increase brand/content reach.
- Digital resurrections – likeness of people without their consent



UNIVERSITY OF TORONTO

DISCUSSION

•GenAI tools used to exploit human likeness and falsify evidence- abundance of source of data and broader societal impact

•Most attacks not sophisticated, instead exploit easily available GenAI capabilities which require minimal technical expertise.

•Mass production of low quality and nefarious synthetic content- increases people's skepticism, overload with verification tasks, impede information retrieval and distort collective understanding of socio-political reality.

•Steps taken towards mitigation- model developers create safeguards by removing toxic content from training data or restricting prompts that violate these tools' terms of services.

•Technical and non technical interventions necessary to prevent risk of high misuse

Limitations

Reliance on Media reports

- Introduction of bias- they prioritise elements with sensational elements or those that directly impact human perception, hence the skeweness.
- Covert attacks with no human in the loop are underrepresented, as companies keep this information private – use of GenAI to obfuscate malicious code to evade filters
- Need for better and more comprehensive sources of anonymized data

Time bound analysis

- With every passing day GenAI models acquire new capabilities and become more agentic, their potential for misuse increases.
- Progressive integration of GenAI into social media to deliver content leads to new forms of information manipulations
- To keep up with dynamic landscape, need further longitudinal analysis and continued monitoring of emerging tactics.

Analysis limited to text prompt models

- Most misuse cases were observed in models that take in text prompts are input rather than leveraging multimodal models
- New modalities and capabilities will enable new forms of misuse.

Conclusion

- A review of the evolving landscape of GenAI misuse and its impacts.
- Fears of sophisticated adversarial attacks dominate public space, but our research proves otherwise.
- Mostly low tech easily accessible misuses by a range of actors driven by financial or reputational gain
- GenAI amplifies existing threats by lowering barriers to entry and increasing potency and accessibility of previously costly tactics.
- Multi-faceted approach to mitigating GenAI misuse.
- Addressing this challenge requires not only technical advancements but also a deeper understanding of the social and psychological factors.