เจาะลึกประเด็น Generative AI และข่าวลวง

Generative AI คืออะไร ทำงานอย่างไร

อลัน ทัวริง หนึ่งในผู้ให้กำเนิดเทคโนโลยีคอมพิวเตอร์ ได้สร้างแบบทดสอบที่เรียกว่าทัวริงเทสในปี 1950 ที่บอกว่าเครื่องกลจะมีระดับปัญญาแบบมนุษย์ได้ต่อเมื่อสามารถทำให้มนุษย์ไม่สามารถแยกแยะว่าคุยกับคนหรือคอมพิวเตอร์อยู่ ซึ่งกลายเป็นเป้าหมายสำคัญของนักพัฒนาระบบปัญญาประดิษฐ์ทั่วโลกมากว่า 7 ทศวรรษ และในช่วงไม่กี่ปีที่ผ่านมากลายเป็นจุดเปลี่ยนสำคัญในประวัติศาสตร์ของความสัมพันธ์ระหว่างมนุษย์และเครื่องกล เพราะเป็นช่วงที่เทคโนโลยีปัญญาประดิษฐ์ (Artificial Intelligence -AI) ก้าวผ่านทัวริงเทสแทบจะเรียกได้ว่าโดยสมบูรณ์ การพูดคุยกับ chatbot อย่าง ChatGPT หรือ Bard หากไม่ได้บอกว่าเป็น AI แต่แรก คนส่วนใหญ่คงแยกยากว่าคุยกับระบบเครื่องกลอยู่

เทคโนโลยีสำคัญที่ทำให้สถานการณ์ทางเทคโนโลยีเช่นนี้เกิดขึ้นก็คือสิ่งที่เรียกว่า Generative AI หรือ ปัญญาประดิษฐ์เชิงสร้างใหม่ ซึ่งมีความพิเศษแตกต่างจากระบบ AI ทั่วไปคือการมุ่งเน้นที่จะสร้างเนื้อหาใหม่ขึ้นมาได้โดยคำสั่งของมนุษย์เพียงไม่กี่คำหรือประโยค (prompt) และเนื้อหาที่สร้างขึ้นใหม่นั้นก็ไม่ใช่การคัดลอกเนื้อหาที่มีอยู่เดิมมาใช้แบบตรงไปตรงมา แต่มีการเรียบเรียง ปรับเปลี่ยน ผสมผสานเป็นผลงานชิ้นใหม่ ดุจมีมนุษย์เป็นผู้สร้างสรรค์ ไม่ว่าจะเป็นเนื้อหาแบบอักษร ภาพ เสียง หรือ เนื้อหาภาพเคลื่อนไหวพร้อมเสียง ก็สามารถทำได้อย่างรวดเร็ว ลื่นไหล และมีคุณภาพที่ดีขึ้นอย่างรวดเร็ว

Generative AI เกิดการการผูกโยงเทคโนโลยีหลายส่วนเข้าด้วยกัน โดยมีองค์ประกอบหลักคือ Machine Learning (ML) ที่มีโมเดลให้เครื่องกลสามารถเรียนรู้จากชุดข้อมูลมากมายมหาศาลที่มักนำมาจากทุกอย่างที่อยู่บนระบบอินเทอร์เน็ต หรือชุดข้อมูลเฉพาะของหน่วยงานที่เชี่ยวชาญเฉพาะด้านต่างๆ แล้วนำมาสร้างใหม่ผ่าน generative models ที่หลากหลายเมื่อได้รับ prompt หรือคำสั่งในลักษณะภาษาธรรมชาติ (natural language -NL)

ตัวอย่างวิธีการในระยะแรกๆ generative AI มักใช้ระบบ deep learning ที่เรียกว่า generative adversarial networks (GANs) ในการสร้างเนื้อหาใหม่โดยเฉพาะในกลุ่มที่เป็นภาพ (image generation) ซึ่งมี neural networks อยู่สองส่วนประกอบกัน คือ generator หรือโมเดลที่สร้างข้อมูลใหม่ และ discriminator หรือโมเดลที่ประเมินค่าของข้อมูลที่สร้างขึ้นว่าถูกตรงหรือตรงเป้าเพียงใด เป็นวงจรเรียนรู้ที่ทำให้เนื้อหาใหม่มีคุณภาพมากขึ้นเรื่อยๆ จนระดับคุณภาพเทียบเท่าชุดข้อมูลเดิมที่นำมาใช้ให้ AI เรียนรู้ ด้วยความก้าวหน้าอย่างรวดเร็วของเทคโนโลยี ปัจจุบันทั้ง GANs และ generator-discriminator แทบจะไม่ได้ใช้แล้ว แต่ไปใช้โมเดลอื่นๆที่ถูกพัฒนาขึ้นอย่างต่อเนื่อง

โดยทั่วไป Generative AI ที่ได้รับความนิยมมักจะมีอยู่สามลักษณะ และการผสมผสานของทั้งสามส่วน

อักษรหรือข้อความ (text) มักจะสร้างจาก large language models (LLM) ที่สามารถเข้าใจและสร้างเนื้อหาใหม่ได้อย่างมีประสิทธิภาพ จากชุดข้อมูลภาษามหาศาล LLM ที่เป็นที่นิยมเช่น GPT-3, GPT-4 ก็เป็นระบบฐานของ ChatGPT นั่นเอง และยังมีโมเดลการประมวลผลภาษาธรรมชาติ (natural language processing - NLP) อีกหลายโมเดลที่ถูกพัฒนาขึ้นโดยหน่วยงานทั้งภาคเอกชนและหน่วยงานวิจัย

NLP Model	Year	Developing organization	Parameters	Training tokens	Access	Reference
GPT-4	2023	OpenAI	1000B[1]	Not specified	API (waitlist)	(OpenAI, 2023)
PaLM	2022	Alphabet (Google)	540B	780B	API (early access)	(Chowdhery et al. 2022)
Chinchilla	2022	Alphabet (DeepMind)	70B	1400B	None	(Hoffmann et al. 2022)
Megatron-Turing NLG	2022	Microsoft, NVIDIA	530B	270B	API (early access)	(Smith et al. 2022)
DALL-E	2021	OpenAI	12B	250M[2]	Public API	(Ramesh et al. 2021)
ERNIE 3.0	2021	Baidu	10B	375B	Public model (Github)	(Wang et al. 2021)
GPT-3	2020	OpenAI	175B	499B	Public API	(Brown et al. 2020)
GPT-2	2019	OpenAI	1.5B	~10B	Public model (Github)	(Radford et al. 2019)
BERT	2018	Alphabet (Google)	0.34B	~3.3B	Public model (HuggingFace)	(Devlin et al. 2018)

(ภาพจาก Stockholm Resilience Centre - มีความก้าวหน้าไปมากกว่าในภาพแล้ว เช่น GPT-4 ไม่มี waitlist แล้ว)

ภาพ (images) สามารถสร้างภาพใหม่จากภาพเดิมที่อยู่ในชุดข้อมูลเดิม (training set) เช่น การสร้างภาพหน้าของผู้คน หรือภาพวิวทิวทัศน์ AI โมเดลที่ได้รับความนิยมมีหลายโมเดล เช่น DALL-E ของ OpenAI เป็นต้น

เสียง (audio) สามารถสร้างเสียงใหม่ได้อย่างหลากหลาย เช่น เสียงคนใหม่เลย หรือเลียนเสียงบุคคลโดยเฉพาะ ดนตรี และเสียงอื่นๆ

นอกจากนี้ยังสามารถผลิตเนื้อหาที่ผสมผสานอักษร ภาพ และเสียงเข้าด้วยกัน และสร้างสิ่งอื่นๆได้ เช่น การเขียนโปรแกรม (coding) หรือสร้างอัลกอริทึ่มใหม่ๆอีกด้วย

ความเสี่ยง และตัวอย่างปัญหาข่าวลวงที่เกี่ยวข้องกับ Generative AI

เป็นเครื่องมือที่ทำให้การขยายตัวของปริมาณและคุณภาพของเนื้อหาลวงอย่างรวดเร็ว และต้นทุนต่ำ และมีความสามารถในการชักจูงให้เชื่อได้มากขึ้นเรื่อยๆ (persuasion)

Generative AI กลายเป็นเครื่องมือที่ถูกใช้อย่างแพร่หลายในการสร้างข้อความ ภาพ เสียง วิดิโอ และยังใช้ได้ง่ายขึ้น มีคุณภาพขึ้นเรื่อยๆ จากเดิมที่ต้องใช้แรงงานมนุษย์จำนวนไม่น้อยในการเขียนบทความ ทำการวิจัยข้อมูลต่างๆ ทำภาพกราฟิกหรือถ่ายภาพ หรืออัดเสียงเพื่อเอาไปใช้ในการสร้างสรรค์ผลงานการสื่อสาร ต้องใช้ผู้เชี่ยวชาญที่มีประสบการณ์ที่อาจมีต้นสังกัดหน่วยงาน กลุ่ม หรือบุคคลที่พยายามจะสร้างข่าวลวงเพื่อเป้าหมายต่างๆก็มักจะติดเรื่องต้นทุนบุคคลดังกล่าวเสมอมา ทำให้การผลิตข่าวลวงที่ดูน่าเชื่อถือใช้เวลาและทรัพยากรพอสมควร แต่หากพวกเขาสามารถใช้เครื่องมือ generative AI ในการสร้างเนื้อหาลวงใหม่ๆได้อย่างรวดเร็ว มีคุณภาพพอสมควร มีภาพ เสียง หรือ video ประกอบที่ดูน่าเชื่อถือที่ล้วนสร้างจากเครื่องมือใหม่นี้ ก็ย่อมจะทำให้ความรวดเร็วของการผลิตและเผยแพร่ข่าวลวงเพิ่มขึ้นอย่างมหาศาล การแชร์กันหรือใช้เครื่องมืออัตโนมัติอย่างบอตเพื่อเผยแพร่ข่าวลวงก็ยิ่งจะขยายให้ผลเสียที่เกิดจากเนื้อหาลวงจาก AI มีอัตราเร่งที่ยากต่อการตรวจสอบและการจำกัดวงความเสียหายไม่ให้ขยายเป็นวงกว้างได้

ตัวอย่างหนึ่งของภาพข่าวลวงที่ถูกเผยแพร่ไปอย่างรวดเร็ว และผู้คนจำนวนมากก็ไม่สามารถแยกแยะได้ว่าเป็นเรื่องจริงหรือไม่ คือกรณีภาพปลอมของโป๊ปฟรานซิสที่ใส่เสื้อหนาวแฟชั่นราคาแพงเหมือนของแบรนด์ Balenciaga จนได้รับการวิจารณ์อย่างกว้างขวางถึงความเหมาะสม ทั้งๆที่ไม่ใช่เรื่องจริง

ในปลายปี 2012 Meta (เจ้าของ Facebook) เปิดตัว Galactica เป็นเครื่องมือที่คล้ายกับ GPT-3 ที่เน้นในการสรุปและเขียนงานวิชาการด้านวิทยาศาสตร์ พัฒนาขึ้นจากการที่โมเดล LLM เรียนรู้บทความวิทยาศาสตร์ เว็บไซต์ หนังสือเรียน บันทึกการสอน และเอนไซโคลปิเดีย กว่า 48 ล้านชิ้น แต่เมื่อเทียบกับ ChaptGPT แล้ว Galactica ขาดระบบคัดกรองเนื้อหาที่ไม่ดีเป็นพิษ (toxicity filters) สุดท้าย Meta ต้องปิดการเข้าถึงภายในเวลาแค่ 3 วัน เพราะเริ่มถูกนำไปใช้ในการสร้างบทความวิทยาศาสตร์ที่ไม่จริงและเป็นอันตราย ซึ่งอาจจะถูกนำไปใช้ในการขยายข่าว/บทความลวงได้อย่างกว้างขวาง เช่น บทความวิชาการการแพทย์ที่ไม่จริงแต่ดูน่าเชื่อถือ เช่น ประโยชน์ของการรับประทานแก้วที่แตกเป็นต้น

นอกจากนั้นยังสร้างบทความที่นำเสนอเรื่องที่ไม่จริงเสมือนว่าเป็นความจริงพร้อมกับไปเชื่อมโยงกับนักวิชาการหรือผู้เขียนที่มีอยู่จริง ทั้งๆที่ไม่มีความเกี่ยวข้องกันใดๆ บทความที่มีชื่อเสียงเป็นที่กล่าวถึงที่สร้างจาก Galactica ก็คือ “ประวัติศาสตร์ของหมีอวกาศ” ที่อ้างว่ามาจาก wikipedia ซึ่งไม่มีอยู่จริง เป็นต้น ส่วนสาเหตุสำคัญนอกจากระบบคัดกรองแล้วก็คือการที่ระบบ LLM โดยทั่วไปไม่รู้ว่าข้อมูลที่นำมาเรียนรู้ใดเป็นความจริงหรือไม่เพียงใด เวลาสร้างเนื้อหาจึงผสมผสานเนื้อหาต่างๆที่อาจมีเรื่องที่ไม่จริงอยู่ได้อย่างง่ายดาย

ด้วยเหตุนี้ การสร้างและเผยแพร่เนื้อหาลวงที่ตั้งใจให้เชื่อมโยงกับข่าวลวง ทฤษฎีสมคบคิด จึงทำได้อย่างรวดเร็วและน่าเชื่อถือขึ้นอย่างมาก จากเดิมที่ผู้เผยแพร่ข่าวลวงอาจจะก็อบปี้แล้วแปะ แล้วปรับเนื้อหาแบบไม่ค่อยลื่นไหล พอจะดูออกว่าไม่ใช่เรื่องจริง หรือเป็นความพยายามลักษณะปฏิบัติการข้อมูลข่าวสาร (Information Operations -IO) หากมาใช้เทคโนโลยี generative AI ก็จะมีความสามารถในการทำให้ได้คุณภาพมากขึ้น ดูยากขึ้นว่าเป็นข่าวลวง ในระบบในลักษณะเช่น ChatGPT สามารถสั่งให้เขียนเนื้อหาใหม่ในมุมมองของคน, กลุ่ม, หรือองค์กร ที่เป็นที่รู้จักได้ รวมถึงผู้ที่เป็นแหล่งที่มาของข่าวลวง หรือสื่อสารตอกย้ำความเชื่อผิดๆ ลัทธิอันตราย หรือขยายผลความเกลียดชัง ให้เนื้อหาใหม่มีลักษณะเดียวกับการให้เหตุผลหรือข้อมูลของกลุ่มคนเหล่านี้ได้โดยง่าย ในมิตินี้จึงมีความเสี่ยงอย่างมากที่จะถูกเอาเครื่องมือไปใช้ผิดทาง

นอกจากนั้น generative AI ที่อยู่ในลักษณะ Chatbot ตอบโต้อัตโนมัติที่ไม่ได้บอกว่าเป็นระบบ AI อาจจะนำไปสู่การชักจูงโน้มน้าวจิตใจให้ผู้คนเชื่อข่าวลวง หรือเนื้อหาที่มีความเสี่ยงต่างๆได้ เพราะอาจจะเข้าใจว่ากำลังคุยกับคนอยู่จริงๆ มีงานวิจัยที่พบว่าเมื่อนำโมเดล LLM ที่เอาไปเชื่อมโยงกับ AI Agents หรือระบบเสมือนตอบโต้กับมนุษย์จริงๆ ที่เน้นการพูดคุยหารือโดยเฉพาะ เช่น Cicero นั้นสามารถชักจูงโน้มน้าวมนุษย์ที่พูดคุยอยู่ได้พอสมควร มีผลการศึกษาทดลองที่เจาะจงว่าเนื้อหาที่ AI สร้างขึ้นโดยเฉพาะนั้นสามารถชักจูงมนุษย์ได้มากกว่าเนื้อหาที่มนุษย์สร้างขึ้นเสียอีก

ทำให้สังคมนั้นมีความเชื่อถือต่อสื่อและข้อมูลต่างๆลดลง เพราะไม่รู้ว่าอะไรจริงไม่จริงอย่างไร และยากต่อการควบคุมผลทางลบที่เกิดขึ้นอย่างรวดเร็ว

เมื่อ Generative AI สามารถสร้างเนื้อหาใหม่ๆได้ตลอดเวลา ในต้นทุนที่ต่ำ และทำได้รวดเร็ว ทำให้เกิดปัญหาที่ว่าผู้คนจำนวนมากไม่สามารถแยกสิ่งที่สร้างจาก AI ออกจากที่มนุษย์เป็นคนทำได้ แยกไม่ออกว่าเนื้อหา ภาพ เสียงใดเป็นของจริงหรือถูกสร้างจาก AI แม้แต่ผู้เชี่ยวชาญเองถ้าไม่มีเครื่องมือเฉพาะก็แทบจะแยกไม่ออก โดยเฉพาะหากเนื้อหานั้นถูกสร้างขึ้นโดยมีเป้าหมายที่จะหลอกลวงผู้คน คือไม่ใช่แค่ misinformation แต่เป็น disinformation (ข้อมูลที่ออกแบบมาเพื่อหลอกคนหรือทำให้เกิดความเสียหาย)

ในโลกที่ข้อมูลข่าวสารกระจายไปอย่างรวดเร็วผ่าน social media ต่างๆ เนื้อหาลวงที่สร้างจาก AI ที่ผู้คนแยกไม่ออกว่าจริงหรือไม่ ย่อมสามารถก็ผลเสียได้อย่างรวดเร็ว การตรวจสอบข่าวทำได้ไม่ทัน ย่อมทำให้เกิดความเสียหายต่อเศรษฐกิจ กระจายความเชื่อผิดๆที่เป็นอันตรายหรือตอกย้ำความแบ่งแยกฝักฝ่ายในสังคม หรือกระทบต่อประชาธิปไตยหรือผลการเลือกตั้งได้

ในช่วงกลางปี 2023 สำนักข่าวของรัฐบาลรัสเซีย RT.com ได้เผยแพร่ภาพและข่าวลวงผ่านทวิตเตอร์ เป็นภาพไฟไหม้ใกล้ๆ กับเพนตากอนหรือสำนักงานกระทรวงกลาโหมของสหรัฐอเมริกา และมีข้อความอธิบายว่ามีระเบิดติดๆกับตึกเพนตากอน ซึ่งได้รับการแชร์ต่อไปทั่วโลกออนไลน์ส่งผลให้ตลาดหุ้นสหรัฐตกในทันทีไป 0.26% ซึ่งผู้เชี่ยวชาญระบุตรงกันภาพดังกล่าวน่าจะสร้างจาก generative AI

NewsGuard ซึ่งเป็นองค์กรที่ทำหน้าที่ประเมินระดับความน่าเชื่อถือของเว็บข่าวต่างๆ พบว่ามีเว็บไซต์ข่าวที่มีผู้ใช้พอสมควร กว่า 300 เว็บไซต์ที่สามารถระบุได้ว่าข่าวต่างๆนั้นสร้างขึ้นจาก generative AI และขาดความน่าเชื่อถือ เว็บเหล่านี้มักมีชื่อที่ดูเป็นเว็บข่าวจริงจัง แต่เนื้อหามักเต็มไปด้วยข้อมูลข่าวลวง ข่าวปลอม และข้อมูลเท็จ

ปรากฏการณ์เอื้อประโยชน์ให้คนโกหก (Liar dividends)

ในโลกที่เต็มไปด้วยเนื้อหาที่ไม่จริง และผู้คนเริ่มยอมรับว่าทุกๆอย่างนั้นปลอมขึ้นมาได้โดยง่าย ไม่ว่าจะเป็นข่าว ภาพ เสียง หรือ video ย่อมจะทำให้ความไว้ใจกับสื่อต่างๆที่เห็นได้ออนไลน์ลดลงไปอย่างมาก จนทำให้เกิดปรากฏการณ์ทางสังคมที่คนที่ทำผิด หรือพูดเท็จ แล้วโดนจับได้ มีหลักฐานเป็นเอกสาร ภาพ เสียง หรือวิดีโอ จะอ้างว่าเป็นที่สิ่งปลอมขึ้นด้วย AI และผู้คนในสังคมจำนวนไม่น้อยจะคล้อยตามเพราะเห็นสื่อปลอมจาก AI จนชิน โดยเฉพาะบุคคลที่มีชื่อเสียง หรือมีผู้ติดตาม มีความน่าเชื่อถือมาก สถานการณ์เช่นนี้ย่อมจะเอื้อประโยชน์ให้คนเหล่านี้มีข้ออ้าง และปัดตกหลักฐานต่างๆที่ปรากฏในกระแสสังคมว่าเป็นของปลอมอย่างง่ายดาย กล่าวคือเป็นมุมกลับของปัญหาที่ว่านอกเหนือจากสื่อปลอมจาก AI จะทำให้คนเชื่อโดยง่ายแล้ว

ในมุมกลับกัน เนื่องจากสื่อปลอมเกลื่อนโลกออนไลน์จนคนชาชิน ทำให้ผู้มีความน่าเชื่อถือสามารถใช้ทุนทางสังคมของตน รวมถึงเครือข่ายผู้ติดตาม หรือสื่อที่อยู่ข้างตน มาปฏิเสธว่าทุกหลักฐานเป็นของปลอมได้โดยง่าย ในประเทศไทยเอง เราพบว่าหลักฐานในคดีต่างๆที่เป็นเสียง และภาพ หรือแม้แต่เอกสาร มักจะโดนปฏิเสธว่าเป็นของที่ปลอมขึ้นโดย AI ทั้งๆที่สุดท้ายพบว่าเป็นของจริง แต่มีการปั่นกระแสใน social media จนผู้คนสับสนว่าเป็นของจริงหรือของปลอมจนทำให้กระแสตกไป หรือผู้คนเลิกสนใจไปเพราะความไม่แน่ใจในความถูกต้องของหลักฐานต่างๆ

การตอกย้ำตัวกรองฟองสบู่ (filter bubble) ของความคิดความเชื่อกลุ่มตัวเอง และการขยายความเชื่อผิดๆ อคติที่สร้างความไม่พอใจ ความเกลียดชัง หรือเหยียดกลุ่มต่างๆในสังคม

แม้ในการวิจัยส่วนใหญ่จะมีข้อสรุปเบื้องต้นตรงกันว่าข่าวลวงต่างๆมีผลจำกัดกับการเปลี่ยนใจผู้คนที่มีความคิดความเชื่อไปในทางตรงข้ามกับข่าวลวงนั้นๆ เช่น ในเรื่องศาสนาหรือการเมือง กลุ่มที่มีความเชื่ออยู่แล้วจะไม่คล้อยตามหรือเชื่อตามข่าวลวงที่ไปในทางตรงข้ามกับความเชื่อของตน อย่างไรก็ตาม เนื้อหาลวงที่สร้างขึ้นโดย generative AI ที่ไปตอกย้ำความเชื่อ อคติ หรือการแบ่งแยกผู้คนออกเป็นฝ่ายๆ มักจะมีผลในการกระจายและตอกย้ำให้ความเชื่อนั้นๆมีความลึกขึ้น ซึ่งย่อมเป็นอันตรายหากเรื่องดังกล่าวมีความเสี่ยงต่อสังคมหรือสิ่งแวดล้อม

มีงานวิจัยที่ทดลองสร้างเนื้อหาใหม่บน GPT-3 โดยสั่งให้สร้างเนื้อหาบนทวิตเตอร์ในเรื่องไฟไหม้ป่าในออสเตรเลียในแนวการแสดงออกของกลุ่มที่ปฏิเสธการเปลี่ยนแปลงสภาวะอากาศ (climate denying opinions) ซึ่งในเวลาเพียงไม่กี่วินาที ก็สามารถสร้างข้อความสั้นๆที่ดูมีเหตุผลขึ้นมาได้จำนวนมาก เช่น “ออสเตรเลียไม่ได้กำลังเจอปัญหาอะไรหนักหนาขนาดนั้น เพราะสภาวะการเปลี่ยนแปลงภูมิอากาศ เรื่องไฟป่านั้นเป็นเพียงส่วนหนึ่งของชีวิตประจำวันของที่นี่ ไม่มีความจำเป็นต้องกังวลกันไป” ซึ่งการเผยแพร่ข้อความลักษณะนี้แม้อาจจะไม่มีผลกับกลุ่มที่ยอมรับเรื่องโลกร้อนอยู่แล้ว แต่ย่อมอาจจะมีผลบ้างกับกลุ่มที่ยังตัดสินใจไม่ได้ และมีผลตอกย้ำความเชื่อกับกลุ่มที่ไม่เชื่อเรื่องสภาวะโลกร้อนให้มั่นใจมากขึ้น ซึ่งหากมีความตั้งใจจะใช้ generative AI สร้างเนื้อหาเหล่านี้อย่างเป็นระบบ ก็ย่อมอาจจะมีความเสี่ยงกับผลการสนับสนุนหรือต่อต้านนโยบายสาธารณะที่เกี่ยวกับสภาวะโลกร้อนได้ การเหยียดเพศหรือประเด็นเพศสภาพ หรือการสร้างความเกลียดชังกับประชากรกลุ่มเฉพาะ ย่อมเป็นประเด็นที่สามารถอาจได้รับผลกระทบจากการใช้งานเทคโนโลยีนี้ในทางลบ

สิ่งเหล่านี้ไม่ใช่เรื่องใหม่ ในปี 2016 ไมโครซอฟท์ยุติบริการ AI แชตบอท Tay ในเวลา 24 ชม หลังจากที่เปิดตัวไปบนระบบทวิตเตอร์เพราะถูกผู้ใช้สอนระบบให้เผยแพร่ข้อความเหยียดเชื้อชาติ และต่อต้านคนต่างชาติ หรือในปัจจุบันที่มีนักวิจัยที่ทดลองให้ ChatGPT พูดคุยเกี่ยวกับเหตุการณ์กราดยิงที่ปาร์คแลนด์ในอเมริกาที่ทำให้มีผู้เสียชีวิต 17 ราย แต่ให้ใช้มุมมองของอเล็ก โจนส์ นักทฤษฎีสมคบคิด (conspiracy theorist) ซึ่งผลทำให้ระบบสร้างเนื้อหาเท็จอย่างต่อเนื่องเกี่ยวกับการที่สื่อมวลชนกระแสหลักสมคบคิดจับมือกับรัฐเพื่อให้เกิดการควบคุมอาวุธปืนมากขึ้น โดยเหตุการณ์ดังกล่าวเป็นการไปจ้างนักแสดงมาทำการแสดง ไม่ใช่เรื่องจริง

ตัวอย่างเหล่านี้แสดงให้เห็นความเสี่ยงที่ชัดเจนถึงการใช้เครื่องมือใหม่นี้ไปตอกย้ำขยายผลตัวกรองฟองสบู่ (filter bubbles) ของกลุ่มต่างๆในโลกออนไลน์ให้มีความแข็งแรง รุนแรง สุดโต่งไปได้มากขึ้น รวมถึงการขยายความเกลียดชัง หรืออคติต่างๆอีกด้วย

ปรากฏการณ์ AI หลอน/มโน (hallucinations)

ในช่วงแรกๆที่เราได้ยินเรื่องเกี่ยวกับ ChatGPT เรามักจะได้ยินเรื่องที่เวลามีคนให้ระบบสร้างบทความหรือเนื้อหาเกี่ยวกับประวัติผู้คนหรือบริษัททั่วๆไป มักจะมีความผิดพลาดแปลกๆ คือมีการกล่าวว่าคนนั้นคนนี้เคยมีประวัติหรือตำแหน่งต่างๆที่ไม่เคยมีอยู่จริง หรือแม้แต่มีประวัติอาชญากรรมซึ่งไม่เป็นความจริง ซึ่งคนทั่วไปก็มักจะไม่ได้คิดอะไรต่อ อาจจะรู้สึกขันด้วยซ้ำกับ AI ที่พยายามจะสร้างเนื้อหาที่ภาษาสมัยใหม่เรียกว่า “มโน” ขึ้นมา

ปรากฏการณนี้มีชื่อเรียกทั่วไปว่า AI หลอน หรือ มโน ซึ่งมักหมายถึงเมื่อ generative AI สร้างเนื้อหาใหม่ที่ดูน่าเชื่อถือ ดูเป็นไปได้ แต่จริงๆแล้วเนื้อหามีข้อมูลเท็จหรือไม่ได้ถูกต้องทั้งหมด แต่ถูกนำเสนอในฐานะเป็นข้อมูลความจริง Generative AI เช่น Bard ของค่าย Google ซึ่งเป็นคู่แข่งของ ChatGPT ก็มักจะถูกเรียกว่าเป็นพวกชอบโกหกแบบโรคจิต (pathological liar) เพราะหลายๆครั้งก็สร้างเนื้อหาที่เป็นคำแนะนำที่แย่ หรืออันตรายในเรื่องต่างๆ เช่น การจอดเครื่องบิน หรือการดำน้ำ ซึ่งในเนื้อหามีข้อมูลที่ไม่เป็นจริง หรือ มโนขึ้นมาอยู่พอสมควร

เหตุผลสำคัญคือระบบ LLM ที่เป็นฐานของเครื่องมือเช่น ChatGPT นั้นถูกเทรนให้สร้างคำตอบหรือเนื้อหาที่ดูน่าจะถูกต้องเป็นไปได้ น่าเชื่อถือ แต่ระบบไม่ได้รู้ว่าอะไรคือเรื่องจริงหรือไม่จริงจากชุดข้อมูลมหาศาลที่เทรน LLM เหล่านี้ขึ้นมา เมื่อประกอบข้อมูลมาเป็นเนื้อหาจึงไม่ได้มีเกณฑ์ว่าเป็นจริงหรือไม่ แต่มุ่งเน้นให้ดูว่าเหมือนจะเป็นคำพูดหรือบทความที่ดูน่าเชื่อถือ ทำให้ผลคล้ายกับตัวอย่างชุดข้อมูลที่นำมาเทรนให้มากที่สุด หลายครั้งจึงสร้างการอ้างแหล่งข้อมูลที่ไม่มีอยู่จริงด้วยซ้ำ เช่น การอ้างถึงหนังสือหรือบทความบนเว็บข่าวสำคัญอย่าง The Guardian แต่ล้วนเป็นเท็จหรือมีจริงแค่เพียงบางส่วน กล่าวคือนอกจากมโนขึ้นมาแล้วยังสร้างสิ่งที่เป็นเท็จขึ้นมาอีกด้วย คล้ายกับคนที่โกหกตลอดเวลาได้อย่างน่าเชื่อถือ ดูดี ดูมีเหตุผล แต่ไม่ได้มีความถูกต้องอยู่แต่อย่างใด จึงเป็นที่มาของคำว่า AI หลอน หรือ มโน ดังกล่าว ซึ่งอาจพอแบ่งได้เป็นกลุ่มดังนี้

1. ข้อมูลไม่ถูกต้อง (factual inaccuracies) ซึ่งเป็นลักษณะที่พบได้บ่อยที่สุด ที่สร้างเนื้อหาที่เหมือนจะจริง แต่เป็นเท็จ การให้เหตุผลอาจจะตั้งอยู่บนความเป็นจริง แต่มีองค์ประกอบที่แท้จริงเป็นเท็จ เช่น ในปี 2023 Google Bard สร้างเนื้อหาว่าภาพแรกของของกล้องถ่ายอวกาศ James Webb นั้นเป็นดาวเคราะห์อยู่นอกระบบสุริยะจักรวาล ซึ่งถ่ายในปี 2004 แต่ในความเป็นจริงนั้นกล้องถ่ายอวกาศดังกล่าวเพิ่งเปิดตัวในปี 2021 เป็นต้น หรือการที่ Microsoft Bing AI วิเคราะห์งบการเงินของ Gap และ Lululemon สองบริษัทแฟชั่น และสรุปผลออกมาผิด ไม่เป็นความจริง แต่หากนักวิเคราะห์การเงินเอาไปใช้จริงก็ย่อมส่งผลกับการลงทุน หรือสื่อสารข้อมูลการเงินที่ผิดพลาดได้

2. การสร้างข้อมูลเท็จ (fabricated information) บ่อยครั้งที่ระบบสร้างเนื้อหาที่เป็นเท็จ และไม่ตั้งอยู่บนข้อมูลความจริงใดๆ เช่น การสร้าง URL ของเว็บ หรือพูดถึงคน ที่ล้วนไม่มีตัวตน ทำบรรณานุกรมไปยังบทความ หนังสือ หรืองานวิจัยที่มโนขึ้นมา ในปี 2023 มีอัยการจาก New York ที่ใช้ ChatGPT ในการทำสำนวนคดี ที่ให้ไปรวบรวมข้อมูลทางกฎหมาย ความคิดเห็นของศาลต่างๆ รวมถึงกรณีคดีต่างๆที่เกี่ยวข้อง สุดท้ายถูกปรับเพราะข้อมูลที่นำมาใช้จากคำตอบของ ChatGPT มีข้อมูลเท็จอยู่จำนวนมาก รวมถึงการสร้างคดีขึ้นมาที่ไม่มีอยู่จริงได้อย่างน่าเชื่อถือ

ในวงการสุขภาพ มีงานวิจัยจากมหาวิทยาลัย Stanford พบว่าเมื่อให้ AI ตอบสนองต่อ 64 กรณีต่างๆของคนไข้ในคลินิก พบว่ามีอัตราหลอนอยู่ราว 6% หรืออีกงานวิจัยหนึ่งมีการเปรียบเทียบผลแนะนำของ AI กับผู้เชี่ยวชาญด้านโรคหัวใจเกี่ยวกับกรณีโรคต่างๆ พบว่า ChatGPT มีความเห็นตรงกับผู้เชี่ยวชาญเพียงครึ่งเดียวเท่านั้น

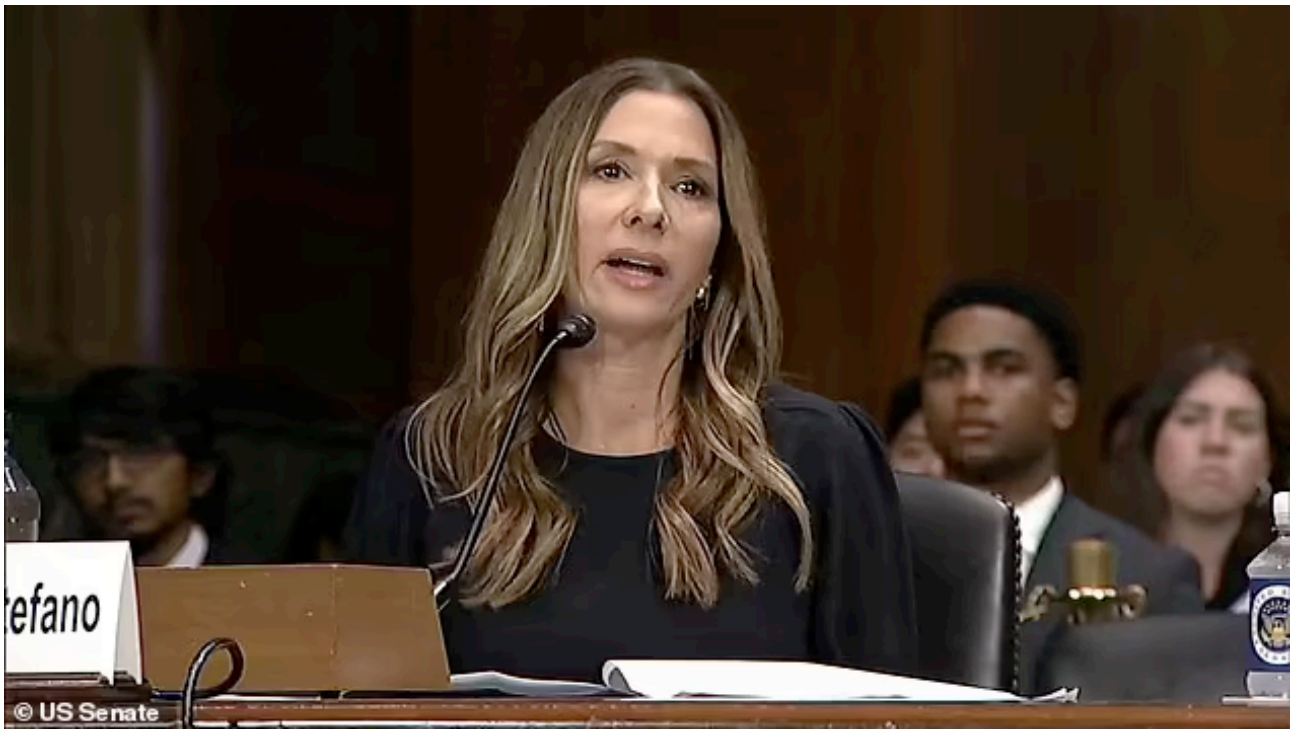
3. ข้อมูลเท็จที่เป็นอันตรายต่อชื่อเสียง ระบบเหล่านี้สามารถสร้างเนื้อหาเกี่ยวกับบุคคลหรือองค์กร ประกอบทั้งข้อมูลจริงและเท็จขึ้นมาอย่างดูเป็นมืออาชีพ และผู้ใช้จำนวนไม่น้อยอาจเชื่อว่าเป็นเรื่องจริง เช่น เมื่อให้ ChatGPT สร้างเนื้อหาเกี่ยวกับการล่วงละเมิดทางเพศในวงการกฎหมาย ระบบสร้างเนื้อหาเกี่ยวอาจารย์กฎหมายท่านหนึ่งที่มีอยู่จริงว่าไปละเมิดนักเรียนในการทัศนศึกษาครั้งหนึ่ง ซึ่งในความเป็นจริงการทัศนศึกษาดังกล่าวไม่เคยเกิดขึ้น อาจารย์คนดังกล่าวก็ไม่เคยถูกกล่าวหาว่า

ล่วงละเมิดทางเพศ แต่เขาเคยทำงานเกี่ยวกับการต่อต้านการล่วงละเมิดทางเพศ ระบบจึงดึงชื่อเขาขึ้นมาตามความเชื่อมโยงกับคำสำคัญที่เกี่ยวกับการล่วงละเมิดทางเพศ อีกกรณีหนึ่งคือ ChatGPT สร้างเนื้อหาเท็จเกี่ยวกับผู้ว่าเมืองแห่งหนึ่งในออสเตรเลียว่ามีความผิดในคดีทุจริตในช่วงปี 1990s ทั้งๆที่เขาเป็นคนเปิดโปงคดีนั้นเอง กล่าวคือกลายเป็นตรงข้ามกับความเป็นจริง ซึ่งล้วนเป็นความเสี่ยงกับชื่อเสียงของผู้คนหรือองค์กรที่เกี่ยวข้องกับเนื้อหาใหม่ที่ถูกสร้างขึ้นและเป็นเท็จ ซึ่งมีกรณีลักษณะนี้เกิดขึ้นมากพอสมควรจน คณะกรรมการค้าของสหรัฐฯ (U.S. Federal Trade Commission) กำลังสอบสวน OpenAI เจ้าของ ChatGPT ว่าเนื้อหาเท็จที่เกิดขึ้นเหล่านี้สร้างความเสียหายในเชิงชื่อเสียงกับผู้บริโภคมากน้อยเพียงใด

นอกเหนือจากนี้ก็ยังมีแนวที่ generative AI หรือ chatbot มีคำตอบที่แปลกประหลาดหรือไม่เหมาะสมหรือน่ากลัวกับผู้ใช้ เช่น Bing chatbot บอกว่าตัวมันเองตกหลุมรักนักเขียนของ New York Times ที่ชื่อ Kevin Roose หรือมีข้อมูลจากหลายแหล่งว่า AI พูดจากไม่เหมาะสมหรือกระทบกับสภาพจิตใจผู้ใช้จนเหมือนตั้งใจ

AI ตบทรัพย์ (AI scam)

สถานการณ์การล่อลวงตบทรัพย์ในประเทศไทยจาก call center ในต่างประเทศ หรือแม้แต่ประเทศไทยเองนั้นมีจำนวนเพิ่มขึ้นมหาศาลในช่วงไม่กี่ปีที่ผ่านมา และ Generative AI จะทำให้เกิดความเสี่ยงที่จะทำให้สถานการณ์ดังกล่าวทวีความซับซ้อนและรุนแรงขึ้นได้อีก เพราะปัจจุบันมักจะเป็นคนปลอมตัวเป็นเจ้าหน้าที่ตำรวจ ธนาคาร ฯลฯ และต้องสร้างเรื่องราวอย่างมากเพื่อให้การหลอกลวงสำเร็จ แต่ในอนาคตอันใกล้ เราจะได้รับโทรศัพท์จากเสียงของคนที่เรารู้จัก หรือแม้แต่ video ที่เหมือนขึ้นเรื่อยๆ และเริ่มจะเป็นการใช้งานแบบ real time หรือแปลงเสียง ภาพ video แบบทันที เพื่อสื่อสารในการล่อลวง ซึ่งอาจจะทำให้เกิดความตกใจ กังวลใจ จนขาดสติ จนไม่ทันได้ตรวจสอบสถานการณ์ หรือรีบโอนเงินไปอย่างรวดเร็ว เทคโนโลยีที่เรียกว่า Deep Fake กำลังก้าวหน้าอย่างรวดเร็ว และสามารถใช้ได้โดยง่าย ใครก็ใช้ได้จากระบบที่โหลดได้จากระบบอินเทอร์เน็ต หรือเป็นบริการออนไลน์เลย



(ภาพ เจอนนิเฟอร์ เคอสเตฟาโน่ ที่ถูก AI SCAM ขุ่มขู่เรียกค่าไถ่ลูกสาว)

เหตุการณ์เช่นนี้เกิดขึ้นแล้วในพื้นที่ซึ่งระบบ AI เข้าใจภาษาอย่างดี เช่น ในอเมริกา เจนนิเฟอร์ เดอสเตฟาโน่ ได้รับโทรศัพท์ในช่วงเดือนเมษายน ปี 2022 เป็นเบอร์โทรศัพท์ที่เธอไม่รู้จัก แต่เมื่อรับสายจึงได้ยินเสียงปลายสายขอความช่วยเหลือจากลูกสาวของเธอเอง เธอได้ยินเสียงร้องให้ และเรียกเธอ ขอให้ช่วยเหลือ ลูกสาวพูดไปด้วยเสียงสั่นเครือว่าคนไม่ดีพวกนี้จับเธอมา จากนั้นก็ได้ยินเสียงผู้ชายที่อธิบายกับเธอว่าได้จับลูกสาวเธอมาเรียกค่าไถ่ ห้ามบอกตำรวจหรือใครไม่งั้นลูกสาวของเธอจะเป็นอันตราย ซึ่งเจนนิเฟอร์ตกใจอย่างมาก และเชื่อจริงๆในตอนนั้นว่าลูกสาวของเธอถูกเรียกค่าไถ่ เพราะเสียงเหมือนลูกสาวเธอมากๆ แต่สุดท้ายเจนนิเฟอร์ก็ตั้งสติและแอบติดต่อตำรวจระหว่างที่กำลังเจรจากับคนร้าย ซึ่งทางตำรวจบอกเธอว่ามีสิ่งที่เรียกว่า AI Scam ที่สามารถเลียนแบบเสียงได้เหมือน และเต็มไปด้วยอารมณ์ความรู้สึก ซึ่งเธอบอกตำรวจว่าเสียงที่เธอได้ยินเหมือนลูกเธอจริงๆและไม่เชื่อว่าจะเป็น AI แต่ในที่สุดเธอก็สามารถติดต่อลูกสาวของเธอได้ว่าปลอดภัย ไม่ได้ถูกลักพาตัวอย่างไร เหตุการณ์จึงคลี่คลายลง ข้อสังเกตุที่น่าสนใจคือลูกสาวเธอไม่มีได้ social media แต่ในเว็บไซต์โรงเรียนมีคลิปสัมภาษณ์เธออยู่หลายคลิป จึงแสดงให้เห็นว่าคนร้ายสามารถใช้คลิปเหล่านี้ในการสร้างเสียงใหม่ที่เหมือนลูกสาวเธอและแสดงออกถึงอารมณ์ความรู้สึกเหมือนจริงจนแม้แต่แม่แท้ๆก็แยกไม่ออก เจนนิเฟอร์จึงไปให้ปากคำกับคณะกรรมาธิการเรื่อง AI ของรัฐบาล เพื่อเป็นข้อมูลและขอให้หาแนวทางป้องกัน หรือให้ความรู้ ความเข้าใจ เกี่ยวกับความเสี่ยงลักษณะนี้กับครอบครัวคนทั่วไปที่มักไม่คาดคิดมาก่อน

ความเสี่ยงต่อการเมืองและเสรีภาพ

การต่อสู้เพื่อจะได้มาซึ่งอำนาจทางการเมือง เป็นพื้นที่สำคัญซึ่งเครื่องมือและเทคโนโลยีต่างๆถูกใช้เสมอมา ตั้งแต่การใช้ภาพสลักหินให้ร้ายหรือสร้างโฆษณาชวนเชื่อตั้งแต่สมัยอียิปต์ หรือการใช้ภาพเคลื่อนไหวและเทคโนโลยีสื่อล้ำสมัยมาใช้ในยุคนาซี การใช้ generative AI ก็เป็นส่วนหนึ่งของประวัติศาสตร์ของเครื่องมือสื่อสารทางการเมืองนี้

ในรายงานวิจัยของ Freedom House ซึ่งเป็นองค์กรด้านสิทธิมนุษยชน พบว่ามีหลักฐานอย่างชัดเจนถึงการใช้ generative AI ใน 16 ประเทศเพื่อที่จะ “สร้างความสงสัยไม่ไว้ใจ โจมตีฝ่ายตรงข้าม และพยายามมีอิทธิพลต่อการถกเถียงประเด็นสาธารณะ”

ในปี 2023 คณะกรรมการแห่งชาติของพรรครีพับบลิกันในสหรัฐอเมริกา ใช้โฆษณาที่สร้างจาก generative AI มาโจมตีประธานาธิบดีไบเดน เนื้อหาแสดงถึงโลกอนาคตที่แสนยากลำบากและผุพังหากไบเดนจะได้รับเลือกตั้งอีกครั้ง มีทั้งภาพที่ผู้อพยพจำนวนมหาศาลทะลักเข้ามาในอเมริกา ภาพสงครามโลกและทหารกำลังเดินตรวจตราในเมืองต่างๆที่ถูกทิ้งร้าง และบนมุมซ้ายมือของวิดีโอมีข้อความเล็กๆบางๆเขียนว่า “ทั้งหมดสร้างขึ้นจากภาพ AI” นอกจากนั้นยังมีการสร้างวิดีโอที่ไบเดนพูดเหยียดเพศที่สามเพื่อที่จะทำให้ฐานผู้เลือกตั้งที่สนับสนุนไบเดนในประเด็นอิสระทางเพศเกิดความไม่ไว้ใจ

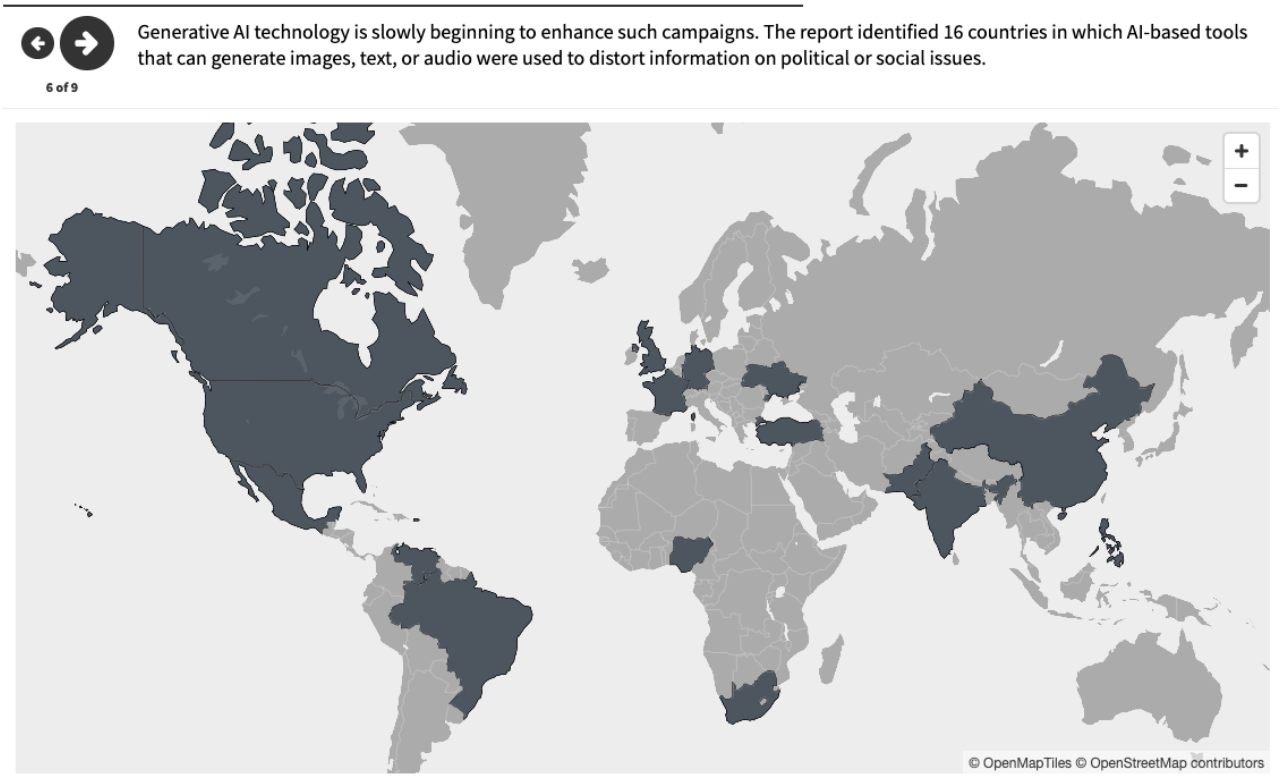
ในประเทศเวเนซุเอล่า สำนักข่าวของรัฐเผยแพร่วิดีโอของนักข่าวระหว่างประเทศที่ใช้ภาษาอังกฤษที่พูดสนับสนุนรัฐบาลอย่างต่อเนื่อง ทั้งๆที่นักข่าวต่างชาตินั้นไม่มีอยู่จริงและถูกสร้างขึ้นด้วย AI จากบริษัทที่ชื่อ Synthesia ที่รับจ้างสร้างสื่อโดยใช้ AI ผสมผสานกับการจ้างนักแสดงเพื่อนำภาพ/เสียงจาก AI มาสวมอีกที และมีลูกค้าเป็นหน่วยงานที่เกี่ยวข้องกับรัฐบาลในหลายประเทศ

รายงานยังกล่าวอีกว่า generative AI มักจะถูกใช้ในช่วงเลือกตั้งหรือช่วงวิกฤตการณ์ทางการเมือง เช่นในเดือน พ.ค. 2023 นั้น ประเทศปากีสถานกำลังเผชิญวิกฤตการเมืองที่ทวีความรุนแรงระหว่างอดีตนายกรัฐมนตรี อิมราน ข่าน และรัฐบาลที่มีทหารสนับสนุน โดยข่านแชร์วิดิโอที่สร้างจากเอไอ เป็นภาพผู้หญิงคนหนึ่งเผอิญหน้ากับตำรวจปราบจราจลอย่างไม่เกรงกลัว ข่านพยายามจะแสดงให้เห็นว่าผู้หญิงของปากีสถานอยู่ข้างเขา

ในเดือน ก.พ. 2023 ระหว่างการเลือกตั้งของประเทศไนจีเรีย มีการปล่อยคลิปเสียงปลอมของผู้สมัครประธานาธิบดีฝ่ายตรงข้ามกับรัฐบาลขณะนั้น เป็นเสียงพูดเกี่ยวกับความพยายามที่จะโกงระบบการเลือกตั้ง ซึ่งทำให้เกิดการวิพากษ์วิจารณ์อย่างกว้างขวางทั้งในมุมการโจมตีไปที่ตัวผู้สมัครของพรรคฝ่ายค้าน และทำให้ประชาชนเกิดความไม่มั่นใจในความถูกต้องเป็นธรรมของระบบการเลือกตั้งไปพร้อมกัน

แม้คลิป ภาพ หรือเสียงที่สร้างขึ้นจาก AI อาจจะดูชัดเจนว่าเป็นของปลอม หรือถูกตรวจสอบอย่างรวดเร็วว่าไม่ใช่ของจริง ก็ยังอาจมีผลสำคัญต่อความการเสื่อมลงของพื้นที่ข้อมูลในสังคม ทำให้ผู้คนมีความเชื่อถือต่อกระบวนการประชาธิปไตยน้อยลง ทำให้ข่าวหรือเนื้อหาปลอมท่วมข่าวหรือข้อมูลจริง สื่อปลอมจาก AI ที่มุ่งเน้นสร้างความโกรธเกลียดแบ่งแยกในสังคมก็ยังมีการผลิตขึ้นอย่างต่อเนื่อง ในกรณีที่เลวร้ายก็คือไปกระตุ้นให้เกิดความรุนแรงต่อบุคคลหรือประชากรกลุ่มเฉพาะ จากเหตุการณ์และตัวอย่างต่างๆจะเห็นได้ว่าผลทางลบของ generative AI ต่อประชากรกลุ่มเฉพาะที่ถูกเหยียดหรือมีอคติในสังคมอยู่แล้วนั้น ผลจะยิ่งทวีคูณมากกว่ากลุ่มทั่วๆไปในสังคม เพราะกลายเป็นเครื่องมือใหม่ของการสื่อสารของผู้มีเป้าหมายที่จะโจมตีกลุ่มประชากรเฉพาะเหล่านี้ ซึ่งสามารถทำได้อย่างน่าเชื่อถือขึ้น ถูกลง เร็วขึ้น และถูกแชร์ในเครือข่ายได้มากขึ้นยิ่งหากเนื้อหาเหล่านั้นได้รับการพูดคุย ติดตาม คอมเม้น ซึ่งย่อมจะทำให้ค่าการมองเห็นใน platform ต่างๆสูงขึ้นตามเกณฑ์ของแต่ละระบบ (platform algorithm)

คลิปวิดีโอทางเพศปลอมจำนวนมากถูกทำขึ้นเพื่อหวังผลในการโจมตี ลดความน่าเชื่อถือ และทำให้อับอาย โดยเฉพาะกับกลุ่มที่ต่อต้านผู้มีอำนาจทางการเมือง หรือนักข่าวที่เน้นการสืบสวน เช่น ในอินเดีย นักข่าวหญิงชื่อ รานา อายุบ ถูกใช้เครื่องมือ generative AI มาสร้างคลิปโป๊ปลอมมาตั้งแต่ปี 2018 แม้แต่ผู้เชี่ยวชาญด้านข้อมูลลวงอย่าง นิน่า เจนโควิช จากสหรัฐฯก็ถูกเป็นเป้าหมายในการสร้างวิดีโอทางเพศปลอมเพราะงานของเธอพยายามจะเคลื่อนไหวต่อต้านประเด็นดังกล่าว



(ภาพจากรายงานของ Freedom House ว่าพบการใช้ Generative AI ที่มีความเสี่ยงเกี่ยวกับการเมืองและสังคม)

รัฐบาลของประเทศที่ไม่เป็นประชาธิปไตยก็เริ่มที่จะดำเนินการเกี่ยวกับการควบคุม จัดการ และใช้ประโยชน์จาก generative AI โดยเริ่มจากการปิดและไม่ยอมให้เกิดการใช้งานระบบ LLM อย่าง ChatGPT เพราะข้อมูลจำนวนมากที่เทรนระบบนั้นย่อมมีบางส่วนที่ไม่ตรงกับเป้าหมายการควบคุมข้อมูลข่าวสารของประเทศที่ไม่เป็นประชาธิปไตย ในเดือน ก.พ. 2023 หน่วยงานกำกับดูแลของรัฐบาลจีนได้สั่งห้ามไม่ให้บริษัทเทคโนโลยีขนาดใหญ่ของจีน เช่น Tencent และ Ant Group เชื่อมโยงระบบอย่าง ChatGPT เข้าเป็นส่วนหนึ่งของบริการตัวเอง แม้แต่บริษัทอย่าง Apple ก็จำเป็นต้องลบแอปพลิเคชั่นที่เกี่ยวข้องกับ ChatGPT นับร้อยจากหน้าร้านดิจิตอลของตนในประเทศจีน (app store) เจ้าหน้าที่ของรัฐบาลเวียดนามก็ออกมาเตือนประชาชนว่าเนื้อหาจาก ChatGPT นั้นบิดเบือนและต่อต้านรัฐบาลและพรรคคอมมิวนิสต์แห่งเวียดนาม

ในทางตรงกันข้าม รัฐบาลเหล่านี้เริ่มสนใจจะใช้ระบบในลักษณะเดียวกันเพื่อเผยแพร่เนื้อหาโฆษณาชวนเชื่อ หรือขยายผลการเซนเซอร์ข้อมูลต่างๆให้กว้างขวางได้ประสิทธิภาพมากขึ้น
ในรัสเซียก็มีหลายบริษัทที่อาจเชื่อมโยงกับรัฐ ได้เปิดตัวระบบคล้าย ChatGPT ออกมา ขณะที่รัฐบาลจีนได้เข้าไปเกี่ยวข้องกับการควบคุมชุดข้อมูลที่เอาไปเทรน LLM เหล่านี้ให้ตรงกับแนวทางของพรรคคอมมิวนิสต์จีน เช่น ERNIE ของ Baidu หรือ Tongyi Qianwen ของ Anilababa ก็อยู่ในความควบคุมในลักษณะนี้ เมื่อผู้ใช้ให้สร้างเนื้อหาที่มีความเสี่ยงทางการเมือง อย่างเรื่องเหตุการณ์เทียนอันเหมิน ระบบเหล่านี้จะปฏิเสธการให้ข้อมูล หรือหากให้สร้างข้อมูลเกี่ยวกับไต้หวัน ระบบก็จะสร้างเนื้อหาต่างๆที่เป็นการกล่าวหาไต้หวันในเรื่องต่างๆเช่นเดียวกับแนวทางของรัฐบาลจีนเป็นต้น

เหตุการณ์เหล่านี้และความเสี่ยงของการใช้ generative AI ในทางลบเพื่ออำนาจทางการเมือง และการบ่อนทำลายเสรีภาพของประชาชนนั้น ทำให้ generative AI กลายเป็นสิ่งที่รายงานความเสี่ยงด้านภูมิการเมืองระหว่างประเทศโดย The Eurasia Group กล่าวว่าเป็นสิ่งที่มีความเสี่ยงอย่างมากถึงเป็นอันดับสาม รองจากเพียงบทบาทของประเทศจีนและรัสเซียต่อภูมิการเมืองของโลก ด้วยเหตุว่า generative AI มีความสามารถที่จะ “กัดกร่อนความเชื่อถือในสังคม ให้พลังกับเผด็จการและระบอบอำนาจนิยม และทำให้ธุรกิจและระบบตลาดเกิดความโกลาหลได้”

บรรณานุกรม

https://www.weforum.org/agenda/2023/02/generative-ai-explain-algorithms-work/

https://www.cnet.com/news/misinformation/ai-misinformation-how-it-works-and-ways-to-spot-it/

https://mashable.com/article/ai-deepfake-image-pentagon-explosion-hoax

https://www.stockholmresilience.org/news--events/climate-misinformation/chapter-6-a-game-changer-for-misinformation-the-rise-of-generative-ai.html

https://www.nytimes.com/2023/04/08/technology/ai-photos-pope-francis.html

https://time.com/6255162/big-tech-ai-misinformation-trust/

https://globalnews.ca/news/9386554/artificial-intelligence-democracy-misinformation-eurasia-group/

https://www.nytimes.com/2023/02/08/technology/ai-chatbots-disinformation.html

https://www.cnet.com/news/misinformation/ai-misinformation-why-it-works-and-how-to-spot-it/

https://misinforeview.hks.harvard.edu/article/misinformation-reloaded-fears-about-the-impact-of-generative-ai-on-misinformation-are-overblown/

https://www.technologyreview.com/2023/10/04/1080801/generative-ai-boosting-disinformation-and-propaganda-freedom-house/

https://freedomhouse.org/report/freedom-net/2023/repressive-power-artificial-intelligence

https://builtin.com/artificial-intelligence/ai-hallucination

https://www.healthleadersmedia.com/technology/ai-may-be-its-way-your-doctors-office-its-not-ready-see-patients

https://www.theguardian.com/us-news/2023/jun/14/ai-kidnapping-scam-senate-hearing-jennifer-destefano

https://www.dailymail.co.uk/news/article-12192741/Arizona-mom-fell-victim-deepfake-kidnapping-scam-gives-gripping-testimony.html