เราจะป้องกัน และแก้ปัญหาความเสี่ยงของ generative AI ต่อประเด็นข่าวลวงได้อย่างไร

ความรับผิดชอบของบริษัทเทคโนโลยีผู้พัฒนา generative AI

หลายๆความเสี่ยงที่ได้กล่าวถึง โดยเฉพาะในส่วนของ AI หลอน หรือการใช้ generative AI ในการสร้างเนื้อหา ภาพ หรือภาพเคลื่อนไหวที่ตอกย้ำความเชื่อผิดๆ ข่าวลวง หรือการสร้างความเกลียดชังขัดแย้งในสังคมนั้น ในแง่มุมนึงก็คือประเด็นความใหม่ของเทคโนโลยีที่เพิ่งเข้าสู่การใช้อย่างกว้างขวางไม่นาน ยังขาดการพัฒนาระบบป้องกันและการติดตามแก้ปัญหาให้เท่าทันกับสถานการณ์ ซึ่งเมื่อเวลาผ่านไป เทคโนโลยีก็ก้าวหน้าขึ้นเรื่อยๆ มีระบบป้องกัน (guardrails) และแก้ปัญหาในเรื่องที่เกี่ยวข้องกับข้อมูลลวงมากขึ้นอย่างต่อเนื่อง ตัวอย่างระบบอย่าง ChatGBT หรือ Bard ก็สามารถสร้างเนื้อหาในประเด็นละเอียดอ่อนต่างๆได้ดีขึ้น ระบบได้รับการแก้ไขปรับปรุงอยู่ตลอดเวลา แต่อย่างไรก็ตามในทางหลักการแล้วนั้น เทคนิควิธีการของ LLM ส่วนใหญ่ก็ยังมีปัญหาในตัวของมันเอง เช่น การที่ไม่รู้ว่าอะไรจริงไม่จริง อะไรคือประเด็นและความซับซ้อนของหัวข้อที่มีความละเอียดอ่อนต่างๆ นอกจากนั้นก็ยังมีเรื่องความปลอดภัยของผู้บริโภคในการใช้งานระบบเหล่านี้ หรือแม้แต่ความพยายามของกลุ่มที่ไม่หวังดีที่ตั้งใจจะใช้เป็นเครื่องมือในทางลบ ล้วนเป็นประเด็นปัญหาที่ซับซ้อนทั้งสิ้น

หลายครั้งมักมีผู้กล่าวว่าเราควรใช้ระบบ AI เหมือนกัน ในการตรวจสอบว่าเนื้อหาใดๆถูกสร้างโดย generative AI หรือไม่ แต่ในความเป็นจริงเป็นเรื่องที่ยากมากในปัจจุบัน แม้แต่ OpenAI ก็เคยเปิดตัวเครื่องมือ AI ที่มาใช้ตรวจสอบเนื้อหาที่สร้างขึ้นโดย generative AI แต่ในเวลาไม่นานนักก็ต้องปิดตัวลง เพราะความแม่นยำในการตรวจสอบต่ำมาก อยู่ที่ราวยี่สิบกว่าเปอร์เซ็นเท่านั้น อย่างไรก็ตาม ปัจจุบันก็มีเครื่องมือ AI ที่กำลังพัฒนาเพื่อการเฝ้าระวัง ดักจับ และตรวจสอบเนื้อหาว่าสร้างจาก AI อยู่พอสมควร แต่ก็ยังไม่ได้ถูกนำมาใช้อย่างกว้างขวางได้สำเร็จ

ดังนั้นการปล่อยให้บริษัทเทคโนโลยีต่างๆบอกว่าจะจัดการความเสี่ยงต่างๆของระบบด้วยตัวเองทั้งหมดนั้นอาจจะไม่ได้ประสิทธิภาพ เพราะมักขาดความชัดเจน โปร่งใส และแม้แต่แรงจูงใจในการพัฒนาระบบที่ปลอดภัย เมื่อเทียบต้นทุนที่ต้องใช้กับการเร่งพัฒนาระบบและตัวธุรกิจให้ได้กำไร บริษัทเทคโนโลยีส่วนใหญ่ก็ย่อมจะเลือกการทำกำไรเป็นหลัก ซึ่งหลายครั้งข่าวลวงหรือข้อมูลที่ไม่จริงแต่กระตุ้นอารมณ์ความรู้สึก รวมถึงภาพ และ vdo ปลอมนั้นได้รับความสนใจจากผู้ใช้อย่างมากมายเหนือกว่าข่าวจริงอย่างมาก ซึ่งก็ย่อมจะทำกำไรให้กับบริษัทมากขึ้นไปด้วย เช่น ผ่านการขายหรือวางโฆษณาอัตโนมัติ การจัดการเรื่องเหล่านี้โดยบริษัทเทคโนโลยีเองจึงมีความยากลำบาก ทั้งในมุมความซับซ้อนของเทคโนโลยี และแรงจูงใจทางเศรษฐกิจ ในกระแสการเลิกจ้างงานครั้งใหญ่รอบล่าสุดของเหล่าบริษัทเทคโนโลยีต่างๆในปี 2023 นั้น พบว่าพนักงานที่ทำหน้าที่ดูแลควบคุมเนื้อหา และความปลอดภัยต่อผู้ใช้ หรือการต่อต้านเนื้อหาลวง เป็นกลุ่มที่ถูกเลิกจ้างไปเป็นจำนวนมาก

ความพยายามเชิงโนบายและการกำกับดูแล

สหรัฐอเมริกา

ในเดือน ก.ค. 2023 บริษัทเทคโนโลยีชั้นนำ 7 บริษัทประกอบด้วย Google, Microsoft, Meta (facebook), OpenAI, Amazon, Anthropic และ Inflection ได้พบกับประธานาธิบดีไบเดน และได้มีข้อตกลงร่วมกันเบื้องต้นในการป้องกันความเสี่ยงที่จะเกิดขึ้นจากระบบปัญญาประดิษฐ์ (AI) โดยแต่ละบริษัทจะไปดำเนินการในลักษณะของอาสาสมัคร ไม่ได้มีกลไกบังคับโดยเฉพาะ เนื้อหาหลักของข้อตกลงนี้คือการให้ความสำคัญกับ “ความปลอดภัย ความมั่นคง และความน่าเชื่อถือ (safety, security & trust)” ในการพัฒนาเทคโนโลยี AI

ด้านความปลอดภัย บริษัทจะทดสอบความปลอดภัยและศักยภาพของระบบ AI ของตน รวมถึงให้หน่วยงานภายนอกมาทดสอบ ประเมินความเสี่ยงในด้านต่างๆต่อสังคม ความเสี่ยงด้านชีวภาพ และความมั่นคงไซเบอร์ และเปิดเผยผลการทดสอบเหล่านั้นให้กับสาธารณะ

ด้านความมั่นคง บริษัทจะป้องกันระบบ AI จากการโจมตีทางไซเบอร์ และความเสี่ยงจากคนใน แล้วแบ่งปันองค์ความรู้ วิธีการ และมาตรฐานต่างๆในการป้องกันการใช้ที่ผิดทาง ลดความเสี่ยงต่อสังคม และปกป้องความมั่นคงของชาติ

ความน่าเชื่อถือ ซึ่งเป็นสิ่งที่บริษัทต่างๆที่เข้าร่วมนั้นเห็นพ้องกันมากที่สุดก็คือการทำให้ผู้ใช้สามารถรับรู้ได้อย่างง่ายดายว่าข้อมูลหรือภาพที่เห็นนั้นถูกสร้างหรือปรับปรุงขึ้นโดย AI นอกจากนั้นก็มีประเด็นเรื่องการทำให้ AI ไม่ขยายการเหยียดหรืออคติ ป้องกันเด็กเยาวชนจากอันตราย และใช้ AI เพื่อแก้ปัญหาสำคัญเช่นการเปลี่ยนแปลงสภาวะอากาศ และมะเร็ง

หลังจากการหารือข้อตกลงดังกล่าว ในเดือน ต.ค. 2023 ประธานาธิบดีไบเดนก็ออกคำสั่งประธานาธิบดีเรื่อง “ปัญญาประดิษฐ์ที่ปลอดภัย มั่นคง และน่าเชื่อถือ” ซึ่งเป็นเนื้อหาที่สอดคล้องกับข้อตกลงความร่วมมือในช่วงหลายเดือนก่อนหน้า

ในคำสั่งนี้ บริษัทด้านเทคโนโลยีที่เกี่ยวข้องกับการพัฒนา AI จะต้องให้ข้อมูลผลการทดสอบความปลอดภัยและประเด็นอื่นๆกับรัฐบาลก่อนที่จะเปิดใช้อย่างเป็นทางการกับสาธารณะ โดยรัฐบาลจะกำหนดมาตรฐานการทดสอบ โดยมีสถาบันมาตรฐานและเทคโนโลยีแห่งชาติเป็นแกนหลักในการพัฒนามาตรฐาน

นอกจากนั้นยังมีประเด็นสำคัญอีกหลายอย่าง เช่น

- ข้อแนะนำอย่างเป็นทางการ (official guidance) เกี่ยวกับการใส่ลายน้ำ (watermark) หรือการกำหนดฉลากดิจิตอล (digital labeling) เพื่อให้ผู้ใช้รับรู้ว่าเนื้อหาใดถูกสร้างหรือปรับปรุงด้วย AI ซึ่งจะทำให้สามารถแยกแยะได้ง่ายขึ้น ป้องกันหรือลดการเผยแพร่ข่าวปลอมหรือเนื้อหาลวงในลักษณะต่างๆได้ สามารถตรวจสอบได้ง่ายขึ้น

- บริษัทที่พัฒนาโมเดล AI ที่มีความเสี่ยงต่อความมั่นคงของชาติ-เศรษฐกิจ หรือมีความเสี่ยงต่อสุขภาพและความปลอดภัยของประชาชน ต้องส่งผลทดสอบความเสี่ยงเหล่านั้นให้กับหน่วยงานของรัฐ

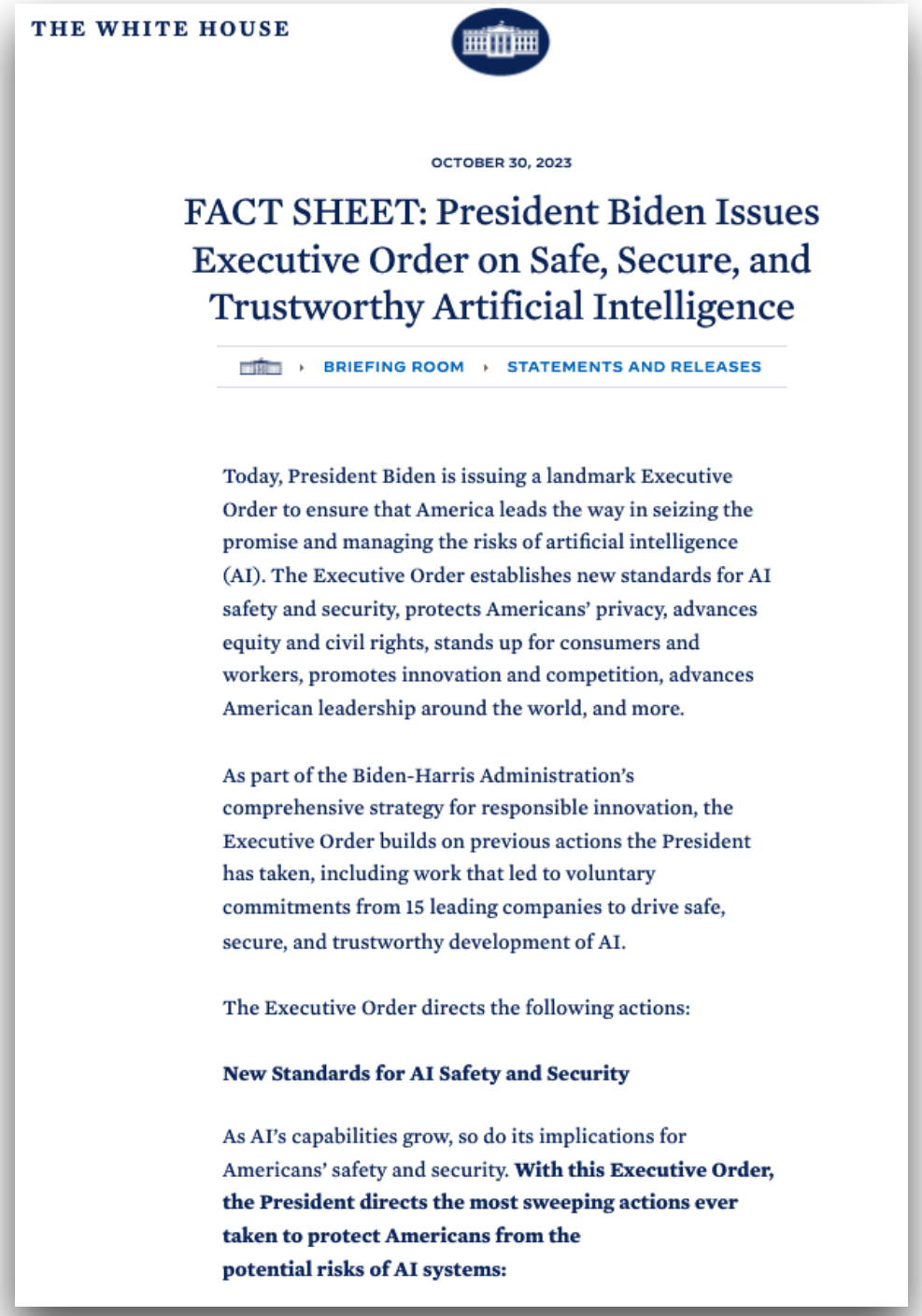
- รัฐจะกำหนดแนวปฏิบัติในการทดสอบระบบแบบ red-team testing หรือการที่ผู้ทดสอบจำลองสถานการณ์เป็นแฮกเกอร์เพื่อทดสอบระบบ

- มาตรฐานเกี่ยวกับการคัดกรองความเสี่ยงของโครงการ AI ที่เกี่ยวกับชีววิทยาและยีนส์ ที่อาจนำไปสู่การสร้างอาวุธชีวภาพได้

- การพัฒนาแนวปฏิบัติที่ดีที่สุดเพื่อป้องกันการใช้ AI และ algorithm แล้วทำให้เกิดการการเลือกปฏิบัติต่อกลุ่มเฉพาะต่างๆ ในสังคม ทั้งในมุมของการจ้างงาน การรับสิทธิประโยชน์ของรัฐ และการเข้าสู่กระบวนการยุติธรรมอย่างเป็นธรรม

- มาตรการต่างๆสำหรับหน่วยงานภาครัฐเพื่อให้เกิดการใช้งานระบบ AI อย่างปลอดภัย มีประสิทธิภาพ รวดเร็ว คำนึงถึงสิทธิของประชาชน

(ศึกษาข้อมูลเพิ่มเติมเกี่ยวกับคำสั่งประธานาธิบดีได้ที่ [https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/])



THE WHITE HOUSE

OCTOBER 30, 2023

FACT SHEET: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence

BRIEFING ROOM › STATEMENTS AND RELEASES

Today, President Biden is issuing a landmark Executive Order to ensure that America leads the way in seizing the promise and managing the risks of artificial intelligence (AI). The Executive Order establishes new standards for AI safety and security, protects Americans' privacy, advances equity and civil rights, stands up for consumers and workers, promotes innovation and competition, advances American leadership around the world, and more.

As part of the Biden-Harris Administration's comprehensive strategy for responsible innovation, the Executive Order builds on previous actions the President has taken, including work that led to voluntary commitments from 15 leading companies to drive safe, secure, and trustworthy development of AI.

The Executive Order directs the following actions:

New Standards for AI Safety and Security

As AI's capabilities grow, so do its implications for Americans' safety and security. **With this Executive Order, the President directs the most sweeping actions ever taken to protect Americans from the potential risks of AI systems:**

กลุ่มประเทศ G7

ในเดือน พ.ค. 2023 มีการประชุมประจำปีผู้นำกลุ่มประเทศ G7 ที่เมืองฮิโรชิมา ประเทศญี่ปุ่น โดยมีการจัดตั้ง “Hiroshima AI Process” ซึ่งเป็นคณะทำงานของ G7 ในการศึกษาโอกาสและความท้าทายของ AI ซึ่งหนึ่งในเป้าหมายที่ชัดเจนก็คือการแก้ปัญหาข้อมูลลวง (misinformation) โดยเฉพาะในบริบทของ generative AI

และในเดือน ต.ค. 2023 ก็ได้มีการเปิดตัวแถลงการณ์ Hiroshima AI Process พร้อมด้วยหลักปฏิบัติ (code of conduct) และหลักการณ์สำคัญ (principles) เกี่ยวกับระบบ AI ที่มีความก้าวหน้า (advanced AI systems)

ในหลักปฏิบัตินั้น กำหนดขั้นตอนการดำเนินการขององค์กรใดๆในการพัฒนาและใช้ระบบ AI โดยเป็นแนวปฏิบัติที่เชื่อมโยงกับหลักการ 11 ข้อ

- กระบวนการกำหนด ประเมิน และป้องกันความเสี่ยงของชั่วชีวิตของระบบ AI ทั้งก่อนและหลังการเปิดตัวในตลาด
- กระบวนการค้นหาและป้องกันจุดอ่อน ปัญหาที่เกิดขึ้นจากการใช้ และการใช้ในทางที่ผิด เมื่อได้เปิดตัวในตลาดแล้ว
- รายงานต่อสาธารณะในเรื่องศักยภาพความสามารถของระบบ AI ข้อจำกัด บริบทของการใช้งานที่เหมาะสมและไม่เหมาะสม เพื่อความโปร่งใสและความรับผิดชอบ
- แลกเปลี่ยนข้อมูลปัญหาต่างระหว่างองค์กรผู้พัฒนาระบบ AI ด้วยกันอย่างรับชอบ เพื่อการสร้างความร่วมมือในการพัฒนามาตรฐาน เครื่องมือ กลไก และแนวปฏิบัติที่ดีที่สุดเพื่อให้เกิดความปลอดภัย มั่นคง และน่าเชื่อถือของระบบ AI
- พัฒนา ดำเนินการ และเปิดเผยข้อมูลเกี่ยวกับระบบธรรมาภิบาล AI และนโยบายจัดการความเสี่ยง
- ดำเนินการจัดการควบคุมความมั่นคงของระบบ ทั้งในเชิงกายภาพ ไซเบอร์ และการป้องกันความเสี่ยงจากคนใน
- พัฒนาและใช้งานกลไกสำหรับให้ผู้ใช้สามารถรับรู้ว่าเนื้อหาใดสร้างขึ้นโดย AI (เช่น มีข้อมูลระบุได้ว่าสร้างจากบริการหรือ AI model ใด) หรือกำลังปฏิสัมพันธ์กับระบบ AI อยู่ โดยจะต้องรับรู้ได้โดยง่ายและน่าเชื่อถือ
- ให้ความสำคัญกับการวิจัยพัฒนาความปลอดภัย ความมั่นคง และความน่าเชื่อถือของระบบ แก้ปัญหาและป้องกันความเสี่ยงสำคัญต่างๆ
- ให้ความสำคัญกับการพัฒนาระบบ AI เพื่อแก้ปัญหาสำคัญที่สุดของโลกต่างๆ โดยเฉพาะในเป้าหมายการพัฒนาที่ยั่งยืน (Sustainabel Development Gaols -SDGs) วิกฤตสภาพอากาศ สุขภาพของโลกและการศึกษา เป็นต้น และควรสนับสนุนการให้ความรู้พื้นฐานเกี่ยวกับเทคโนโลยีดิจิตอลเพื่อให้ประชาชนได้ประโยชน์จากระบบ AI
- พัฒนาและปรับใช้มาตรฐานทางเทคนิค และแนวปฏิบัติที่ดีที่สุด ในระดับระหว่างประเทศ เช่น การใช้ลายน้ำ วิธีการทดสอบระบบ การตรวจสอบเนื้อหาว่าเป็นของจริง ฯลฯ
- ดำเนินการการนำเข้าข้อมูลอย่างเหมาะสม โปร่งใส ได้คุณภาพ ป้องกันอคติที่เป็นอันตราย และมีการป้องกันการละเมิดข้อมูลส่วนบุคคลและทรัพยสินทางปัญญา

สหภาพยุโรป

ในยุโรปมีความพยายามอย่างต่อเนื่องในการกำกับดูแล AI ในหลากหลายลักษณะ โดยเฉพาะในมุมของเนื้อหาลวง (misinformation) ในปี 2022 มีการจัดตั้ง "หลักปฏิบัติของสหภาพยุโรปด้านเนื้อหาลวง (EU Code of Practice on Disinformation)" ซึ่งเป็นกลไกการกำกับดูแลตัวเองในหมู่ผู้พัฒนาเทคโนโลยี AI และองค์กรที่เกี่ยวข้อง เช่น อุตสาหกรรมโฆษณา และภาคประชาสังคม โดยมีความร่วมมืออย่างใกล้ชิดกับสหภาพยุโรปและหน่วยงานที่เกี่ยวข้อง ปัจจุบันมีหน่วยงานสมาชิกเข้าร่วมกว่า 44 องค์กร รวมบริษัทเช่น Meta (facebook), Google, Youtube, TikTok, OpenAI, LinkedIn ฯลฯ

โดยองค์กรสมาชิกร่วมกันพัฒนาหลักปฏิบัติที่ประกอบไปด้วยข้อตกลงดำเนินการ (committments) 44 ข้อ และมาตรการต่างอีก 28 ข้อ ในประเด็นสำคัญ เช่น การป้องกันระบบโฆษณาไม่ให้ถูกใช้เรื่องข้อมูลลวง โฆษณาทางการเมืองต้องโปร่งใส การป้องกันพฤติกรรมในการขยายข่าวลวงข้าม platforms ต่างๆ การดูแลผู้ใช้ให้สามารถรับรู้ว่าอะไรเป็นข้อมูลลวงหรือไม่ใช่ความจริงได้โดยง่ายรวมถึงการเข้าถึงแหล่งข้อมูลที่น่าเชื่อถือ การร่วมมือกับหน่วยงานวิจัยเพื่อศึกษาและจัดการปัญหาข้อมูลลวงรวมถึงการให้การเข้าถึงชุดข้อมูลวิจัยที่สำคัญต่อเป้าหมายดังกล่าว การสนับสนุนเครือข่ายการตรวจสอบข่าว (fact checking) การรายงานผลอย่างเป็นระบบและโปร่งใสเกี่ยวกับการปฏิบัติให้ code of practice นี้ผ่านระบบข้อมูลกลางที่เรียกว่า Transparency Centre ฯลฯ

ในเดือนกันยายนได้มีการรายงานผลรอบ 6 เดือนในการประยุกต์ใช้หลักปฏิบัติฯนี้ในแต่ละ platform หรือระบบที่เกี่ยวข้องและเป็นการรายงานที่ประชาชนหรือผู้สนใจสามารถไปติดตามข้อมูลผลรายงานได้อย่างโปร่งใส

ในรายงานมีเนื้อหาเกี่ยวกับบริบทเฉพาะที่มีความสำคัญกับประเด็นข่าวลวงตามที่สหภาพยุโรปให้ความสำคัญ โดยเฉพาะประเด็นสงครามยูเครน โดยมีการรายงานความคืบหน้าที่น่าสนใจเช่น Youtube ได้ยุติช่องในระบบกว่า 400 ช่องที่เกี่ยวกับปฏิบัติการข้อมูลข่าวสาร (IO) เชื่อมโยงกับหน่วยงาน Internet Research Agency (IRA) ซึ่งได้รับการสนับสนุนจากรัฐบาลรัสเซียในช่วงระหว่างเดือน ม.ค. ถึง เม.ย. 2023 และ Google ยังลบโฆษณาจากกว่า 300 เว็บไซต์ที่เชื่อมโยงไปยังเว็บโฆษณาชวนเชื่อของรัฐ Meta (facebook) ขยายความร่วมมือในการตรวจสอบข่าว (fact-checking) กับ 26 หน่วยงาน ครอบคลุม 22 ภาษาในสหภาพยุโรป ขณะที่ TikTok มีการตรวจสอบข่าวในภาษารัสเซีย ยูเครน เบลารุส และอีก 17 ภาษาในยุโรป และมีความร่วมมือกับ Reuter ในการตรวจสอบข่าว ทำให้มีการตรวจสอบวิดีโอ 832 ชิ้น และมี 211 ชิ้นที่ถูกลบออกไปจากระบบ

นอกจากนั้นยังมีการทำงานร่วมกับหน่วยวิจัย TrustLab ในการทดลองการวิเคราะห์เชิงลึกใน 3 ประเทศ คือ โปแลนด์, สโลวาเกีย และสเปน โดยมีการคัดกรองแล้ววิเคราะห์ข้อความจาก social media 6,155 ชิ้น และบัญชีผู้ใช้ 4,460 ราย จาก Facebook, Instagram, LinkedIn, TikTok, Twitter (ปัจจุบันคือ X) และ Youtube เพื่อให้สามารถแสดงข้อมูลสำคัญในหลักปฏิบัติด้านข่าวลวงของ EU ที่เรียกว่า "ตัวชี้วัดเชิงโครงสร้าง (structural indiciators)" ที่จะทำให้การติดตามการปฏิบัติตามหลักปฏิบัตินั้นมีผลอย่างเป็นรูปธรรม ซึ่งมีตัวชี้วัดสำคัญและผลลัพธ์ดังต่อไปนี้

> **Discoverability** หรือความสามารถในการค้นเจอ คืออัตราส่วนระหว่างข้อมูลลวงและเนื้อหาที่ความอ่อนไหวในประเด็นต่างๆ ซึ่งพบว่า Twitter (X) มีอัตราสูงสุดคือ 0.428 รองลงมาคือ Facebook (0.313) ขณะที่ Youtube มีน้อยที่สุดคือ (0.082)
>
> **Absolute post engagement** คือค่าเฉลี่ยของปฏิสัมพันธ์ (engagement) ของเนื้อหาลวงแต่ละชิ้น และ Relative post engagement คืออัตราส่วนระหว่างค่าเฉลี่ยของ Absolute post engagement ของเนื้อหาลวงเมื่อเทียบกับค่าเฉลี่ยของ abosolute post engagement ของเนื้อหาที่ไม่ใช่ข่าวลวง ซึ่งยิ่งมีค่า relative post engagment สูงเท่าใดก็ย่อมแสดงว่าผู้ใช้ในระบบนั้นๆมีความเสี่ยงสูงที่จะเป็นอันตราย เพราะอาจเจอพื้นที่ที่ความ

ปฏิสัมพันธ์กับเนื้อหาลวงมากกว่าเนื้อหาทั่วไปอย่างมาก ซึ่งในส่วนของ absolute post engagement นั้นค่าสูงสุดคือ TikTok และ Youtube ตามลำดับ แต่ในค่า relative post engagement กลายเป็น Twitter มีค่าสูงที่สุด รองลงมาจึงเป็น Youtube

Ratio of disformation actors หรือสัดส่วนของผู้ตั้งใจเผยแพร่ข้อมูลลวงเมื่อเทียบกับผู้ใช้ทั่วไปที่เป็นกลุ่มตัวอย่างในการวิเคราะห์ โดยพบว่า Twitter และ Facebook มีสัดส่วนผู้ตั้งใจเผยแพร่ข้อมูลลวงเทียบกับกลุ่มตัวอย่างทั่วไปถึง 8-9 % ขณะที่ Youtube นั้นต่ำสุดคือมีอัตราส่วนเพียง 0.8% และพบว่าผู้ที่ตั้งใจจะเผยแพร่เนื้อหาลวงจะติดตามผู้ใช้มากกว่าผู้ใช้ทั่วๆไปมาก และมักจะเพิ่งสมัครเข้าใช้ระบบต่างๆได้ไม่นาน

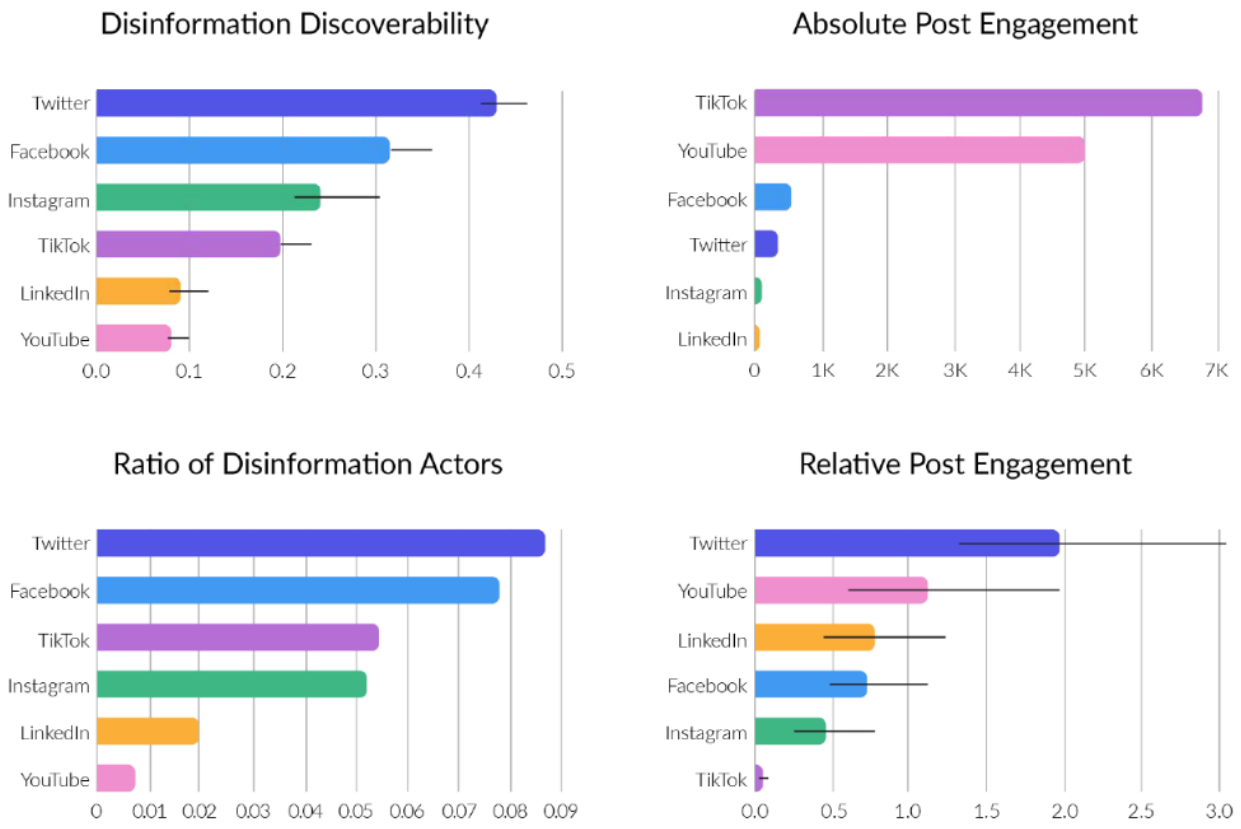


Figure 1: Platform Performance in Discoverability, Relative Post Engagement, Absolute Post Engagement, and Ratio of Disinformation Actors.
See Terminology for metric definitions.

จะเห็นได้ว่า Twitter หรือ X ในปัจจุบันนั้น หากดูข้อมูลเฉพาะ 3 ประเทศในยุโรปที่ทำวิจัย กลายเป็นระบบที่มีการเผยแพร่ข้อมูลลวงมากที่สุดเทียบในเชิงสัดส่วนเมื่อเทียบกับ platform อื่นๆ และ X เลือกที่จะถอนตัวไม่เข้าร่วมกับ EU Code of Conduct ดังกล่าวอีกด้วย

ในช่วงกลางปี 2023 คณะกรรมาธิการยุโรป (European Commission) ได้ขอให้บริษัทเทคโนโลยีสำคัญเช่น Google, Meta, Microsoft, TikTok และสมาชิกของ Code of Conduct ให้ช่วยตรวจจับภาพ วิดิโอ และเนื้อหาที่สร้างขึ้นจาก AI และทำให้ผู้ใช้ทราบว่าเนื้อหานั้นๆถูกสร้างด้วย AI ด้วยการปิดป้ายที่ทำให้เข้าใจง่าย

นอกจากแนวทางการกำกับดูแลกันเองในลักษณะ Code of Conduct ดังกล่าวแล้ว ในเดือน ส.ค. 2023 ได้มีการผ่านกฎหมาย Digital Services Act -DSA ในระดับสหภาพยุโรป ที่บังคับให้ platforms ที่มีผู้ใช้อย่างน้อย 45 ล้านคนต่อเดือนต้องมีการทำแผนประเมินและป้องกัน/จัดการความเสี่ยง ต้องยินยอมให้หน่วยงานวิจัยเข้าไปวิเคราะห์ความเสี่ยงถึงอันตราย และความโปร่งใสของระบบต่างๆโดยเฉพาะ algorithms รวมถึงการที่ต้องรับผิดชอบในการจัดการกับเนื้อหาอันตราย และเนื้อหาลวงต่างๆ ซึ่งมีบทลงโทษที่หนักพอสมควรหากมีการฝ่าฝืน เพราะสามารถปรับเป็นเงินได้ถึง 6% ของรายได้ทั่วโลกของ platform นั้นๆ หรืออาจถูกแบนหรือยุติการใช้งานในสหภาพยุโรปได้

สหภาพยุโรปกำลังอยู่ในกระบวนการพิจารณาร่างกฎหมายอีกหนึ่งฉบับที่เกี่ยวข้องที่เรียกว่า EU AI Act ซึ่งปัจจุบันกำลังเข้าสู่กระบวนการนิติบัญญัติของสหภาพยุโรป โดยมีเนื้อหาที่จะกำกับดูแลความเสี่ยงที่เกี่ยวข้องกับ AI ในภาพรวม และส่งเสริมพัฒนาการในยุโรป ในส่วนที่เกี่ยวกับ generative AI และข้อมูลลวงนั้น จะมีเนื้อหาที่บังคับให้โมเดลพื้นฐาน (Foundation Module -FM) จะต้องถูกประเมินความเสี่ยงต่างๆ ทั้งในเรื่องความถูกต้อง การสร้างเนื้อหาลวง หรือความเสี่ยงเรื่องผลในเชิงการเลือกปฏิบัติหรืออคติต่างๆ และมีแนวทางการจัดการป้องกัน รวมถึงการทดสอบถึงความเสี่ยงต่างๆของ FM เหล่านี้ (FM หมายถึง LLM นั่นเอง) และการสร้างความโปร่งใสของความเสี่ยงและผลทดสอบต่างๆให้กับสาธารณะ รวมถึงการไม่ละเมิดสิทธิข้อมูลส่วนบุคคลและสินทรัพย์ทางปัญญา โดยเฉพาะการนำเข้าชุดข้อมูลมาเทรน FM ต่างๆ

และในเดือน ธ.ค. 2023 สหภาพยุโรปได้บรรลุข้อตกลงร่วมกันเกี่ยวกับกฎเกณฑ์แนวทางสำคัญในการควบคุมดูแล AI ที่จะเป็นองค์ประกอบสำคัญในกฎหมาย AI ของสหภาพยุโรป ซึ่งย่อมจะมีผลต่อการดำเนินธุรกิจหรือโครงการด้าน AI ต่างๆทั่วโลก โดยมีหลักการสำคัญคือระดับการควบคุมดูแลจะต้องสอดคล้องกับความเสี่ยงในระดับที่แตกต่างกัน แบ่งเป็น

ความเสี่ยงในระดับที่ยอมรับไม่ได้ (unacceptable risk) เป็นความเสี่ยงอย่างมากต่อประชาชนและจะถูกห้ามใช้

- การพยายามควบคุม/เปลี่ยนพฤติกรรมของผู้คนและกลุ่มเสี่ยง (cognitive behavioural manipulation), การให้คะแนนเชิงสังคม (social scoring), การใช้ข้อมูลชีวมิติ (biometric) ในการระบุตัวตนและจัดกลุ่มผู้คน รวมถึงการใช้ภาพใบหน้าเพื่อระบุตัวตน ซึ่งอาจมีข้อยกเว้นในส่วนของการบังคับใช้กฎหมายที่สำคัญและจำเป็นจริงๆเท่านั้น

ความเสี่ยงสูง คือระบบ AI ที่มีผลทางลบต่อความปลอดภัย และสิทธิขั้นพื้นฐาน จะต้องมีกระบวนการประเมินและจัดการความเสี่ยงก่อนที่จะเข้าสู่ตลาด และต้องดูแลตลอดอายุของสินค้าบริการนั้นๆ โมเดล AI ที่อาจสร้างความเสี่ยงเชิงระบบในมิติสำคัญต่างๆจะต้องถูกประเมินและทดสอบอย่างต่อเนื่องและผลต่างๆจะต้องส่งให้กับคณะกรรมาธิการยุโรป

Generative AI เช่น ChatGPT จะต้องปฏิบัติตามเงื่อนไขด้านความโปร่งใส

- การระบุให้ชัดเจนว่าเนื้อหานี้ถูกสร้างด้วย AI
- ออกแบบป้องกันไม่ให้สร้างเนื้อหาผิดกฎหมาย
- เปิดเผยข้อมูลโดยสรุปถึงชุดข้อมูลที่มีลิขสิทธิ์ที่ถูกใช้ในการพัฒนา AI (training data)

ความเสี่ยงจำกัด (limited risk) จะต้องปฏิบัติตามเงื่อนไขด้านความโปร่งใสเพื่อให้ผู้ใช้สามารถตัดสินใจต่างๆได้อย่างเหมาะสม หากผู้ใช้มีการปฏิสัมพันธ์กับ AI ก็ต้องแจ้งด้วย และผู้ใช้มีสิทธิที่จะหยุดใช้งานระบบได้ ผู้บริโภคมีสิทธิที่จะร้องเรียนและได้รับคำอธิบายที่เหมาะสม

ซึ่งหากหน่วยงาน องค์กร หรือบริษัทใดๆที่ไม่ดำเนินการตามกฎหมายสามารถจะถูกปรับได้ระหว่าง 1.5% - 7% ของยอดขายขององค์กรนั้นๆจากทั่วโลก (global sales turnover)

ซึ่งคาดว่ากฎหมายจะผ่านสภายุโรปในปี 2024 นี้ และจะมีเวลา 2 ปีเพื่อให้องค์กรและบริษัทต่างๆเตรียมตัวเข้าสู่ระบบการควบคุมดูแลใหม่นี้

สหราชอาณาจักร และบทบาทการเชื่อมโยงพัฒนามาตรฐานความปลอดภัยของ AI ระหว่างประเทศ

ในเดือน พ.ย. 2023 สหราชอาณาจักรได้จัดงานประชุมสุดยอด AI ที่ปลอดภัย (AI Safety Summit) โดยมีตัวแทนจากประเทศต่างๆเข้าร่วม 27 ประเทศ รวมทั้งสหรัฐอเมริกา สหภาพยุโรป แคนาดา ออสเตรเลีย จีน ญี่ปุ่น เกาหลีใต้ หรือแม้แต่ฟิลิปปินส์ โดยได้จัดประชุมที่ Bletchley Park ซึ่งเป็นสถานที่ประวัติศาสตร์ที่อลัน ทัวริ่งนำทีมผู้เชี่ยวชาญเจาะโค้ดสื่อสารลับของฝ่ายอักษะได้สำเร็จในช่วงสงครามโลกครั้งที่ 2 หลังจากการประชุมที่แลกเปลี่ยนประเด็นด้านความปลอดภัยและความเสี่ยงต่างๆของ AI สองวันเต็ม จึงมีการเซ็นคำประกาศ Bletchley โดยมีทั้ง 27 ประเทศร่วมลงนาม โดยมีเนื้อหาหลักเกี่ยวกับการยอมรับถึงความเสี่ยงของ AI ในมิติต่างๆ โดยเฉพาะการสร้างและเผยแพร่เนื้อหาที่หลอกลวง ความมั่นคงไซเบอร์ ความเสี่ยงด้านชีวภาพ และความเสี่ยงที่จะใช้งานไปในทางลบ ทั้งที่จงใจและไม่ได้ตั้งใจ ในคำประกาศมุ่งเน้นความไม่เข้าใจเกี่ยวกับศักยภาพและความสามารถของ AI ต่างๆที่ถูกพัฒนาขึ้นอย่างรวดเร็ว มีการพูดถึง FM ที่มีความเสี่ยงในการสร้างผลลบได้อย่างมาก หากยังไม่มีความเข้าใจ หรือการประเมินและจัดการความเสี่ยงที่เพียงพอ

Rishi Sunak speaks during the closing press conference of the AI safety summit at Bletchley Park. Photograph: Toby Melville/AP

คำประกาศฯเน้นว่าความปลอดภัยเชิง AI นี้เป็นสิ่งที่ทุกฝ่ายต้องมีบทบาท ต้องมีความร่วมมือกันข้ามภาคส่วน และระหว่างประเทศ ผู้พัฒนาระบบ AI ต้องมีความรับผิดชอบต่อความปลอดภัยและการใช้งานของระบบ และควรพัฒนาความร่วมมือในการที่จะกำหนดความเสี่ยงต่างๆร่วมกัน แลกเปลี่ยนข้อมูลความเสี่ยงต่างๆอย่างเป็นวิทยาศาสตร์ เพื่อสร้างความเข้าใจผลของ AI ที่จะมีต่อสังคม รวมทั้งพัฒนานโยบายเพื่อตอบโจทย์ความเสี่ยงด้านต่างๆ โดยเฉพาะการสร้างความโปร่งใส การวัดผลที่มีตัวชี้วัดที่เหมาะสม เครื่องมือในการทดสอบระบบ และทำให้ภาครัฐและภาควิจัยมีศักยภาพเพียงพอ และมีข้อตกลงร่วมกันว่าจะไปพัฒนาแนวทางรองรับการพัฒนาความร่วมมือต่างๆ จะมีการประชุมครั้งต่อไปเพื่อติดตามความก้าวหน้าและพัฒนาให้เกิดความร่วมมือที่ชัดเจนขึ้น

ซึ่งเป็นการแสดงเจตนารมย์ที่จะพัฒนาความร่วมมือระหว่างประเทศที่ค่อนข้างกว้างขวาง และย่อมจะเป็นประโยชน์ต่อการกำหนดทิศทางของความปลอดภัยด้าน AI ระหว่างประเทศต่อไป อย่างไรก็ตาม มีเครือข่ายประชาสังคม องค์กรเทคโนโลยี

และภาควิชาการไม่น้อยที่กล่าวว่าการประชุมดังกล่าวยังขาดการมีส่วนร่วมจากภาคประชาสังคม และภาคส่วนอื่นๆ เพราะมีแต่ตัวแทนของรัฐและบริษัทขนาดใหญ่เป็นหลัก และการให้ความสำคัญอย่างมากว่าอันตรายจาก AI จะมาจาก FM หรือ LLM อาจจะคับแคบไปอย่างมาก

ในสหราชอาณาจักรเอง ก็เปิดตัวสถาบันความปลอดภัยด้าน AI ไปพร้อมกัน โดยมีภารกิจสำคัญในการจัดการความเสี่ยงที่เป็นอันตรายของ AI ต่อประโยชน์ของสังคมและประเทศ โดยมีเนื้องานสำคัญคือการพัฒนาการประเมินความเสี่ยงด้านความปลอดภัยและความมั่นคงของระบบ AI รวมถึงความผลต่อสังคมที่หลากหลาย, วิจัยด้านความปลอดภัยทาง AI โดยเฉพาะระบบธรรมาภิบาล AI การวัดผลและนวัตกรรมเพื่อความปลอดภัย, และประสานการแลกเปลี่ยนข้อมูลที่เกี่ยวข้องจากทุกภาคส่วน

บทเรียนสำหรับประเทศไทย

ความก้าวหน้าเชิงนโยบายเกี่ยวกับการกำกับดูแลความเสี่ยงของ Generative AI ในด้านต่างๆ โดยเฉพาะการสร้างและเผยแพร่ข้อมูลลวงนั้น มีความก้าวหน้าอยู่พอสมควร ประเทศไทยสามารถเรียนรู้ได้ ร่วมศึกษาและประยุกต์เข้าสู่บริบทของไทยได้ โดยเฉพาะประเด็นหลักๆเช่น

- การสร้างความรู้เท่าทันเชิง AI ให้กับประชาชนและผู้ใช้งาน
- การส่งเสริมให้เกิดการประเมินและจัดการความเสี่ยงใน AI models สำคัญๆที่เป็นภาษาไทย แนวทางการทดสอบ แล
- การระมัดระวังความเสี่ยงในเรื่องอคติต่างๆในชุดข้อมูลเทรนนิ่ง
- การส่งเสริมร่วมกับเครือข่ายระหว่างประเทศให้ผู้พัฒนาต้องปิดป้ายหรือลายน้ำหรือมีวิธีให้ประชาชน และผู้เชี่ยวชาญสามารถแยกแยะได้ว่าเนื้อหาใดสร้างจาก AI
- การกำหนดให้ platform ต่างๆต้องร่วมรับผิดชอบต่อเนื้อหาลวงที่อยู่ในระบบของตน
- การพัฒนาให้เกิด code of conduct ที่มีเนื้อหาเป็นสากลและมีกลไกการติดตามที่มีประสิทธิภาพ
- การศึกษาวิจัยอย่างมีข้อมูลประกอบชัดเจนเกี่ยวกับสถานการณ์ของการสร้างและกระจายข้อมูลลวง ทั้งที่เกี่ยวกับ AI โดยตรงหรือทางอ้อม
- การสร้างความร่วมมือทุกภาคส่วน ทั้งภาครัฐ เอกชน องค์กรเทคโนโลยี ภาควิชาการ และภาคประชาสังคมในการขับเคลื่อนความปลอดภัยเชิง AI ให้เกิดขึ้นจริง

บรรณานุกรม

Biden's AI Directive Factsheet
https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/

AI Safety Institute UK
https://www.gov.uk/government/publications/ai-safety-institute-overview/introducing-the-ai-safety-institute

Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems https://www.mofa.go.jp/files/100573473.pdf

EU Code of Conduct Pilot study https://disinfocode.eu/wp-content/uploads/2023/09/code-of-practice-on-disinformation-september-22-2023.pdf

The Bletchley Declaration
https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023

https://www.theguardian.com/technology/2023/oct/30/biden-orders-tech-firms-to-share-ai-safety-test-results-with-us-government

https://www.scientificamerican.com/article/bidens-executive-order-on-ai-is-a-good-start-experts-say-but-not-enough/

https://apnews.com/article/artificial-intelligence-chatgpt-europe-rules-906fc89d2561b200fa6eb40a06b946a5

https://www.bruegel.org/analysis/adapting-european-union-ai-act-deal-generative-artificial-intelligence

https://ec.europa.eu/commission/presscorner/detail/en/SPEECH_23_4645

https://disinfocode.eu/introduction-to-the-code/

EU Code of Conduct Pilot study https://disinfocode.eu/wp-content/uploads/2023/09/code-of-practice-on-disinformation-september-22-2023.pdf

https://www.theguardian.com/technology/2023/nov/02/top-tech-firms-to-let-governments-vet-ai-tools-sunak-says-at-safety-summit

https://www.euronews.com/next/2023/11/01/a-world-first-ai-agreement-elon-musk-and-a-kings-speech-the-key-takeaways-from-the-uk-ai-s

https://theconversation.com/bletchley-declaration-international-agreement-on-ai-safety-is-a-good-start-but-ordinary-people-need-a-say-not-just-elites-217042

The Bletchley Declaration
https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023

http://lcfi.ac.uk/news-and-events/news/2023/oct/31/ai-safety-policies/

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

https://www.technologyreview.com/2023/12/11/1084942/five-things-you-need-to-know-about-the-eus-new-ai-act/