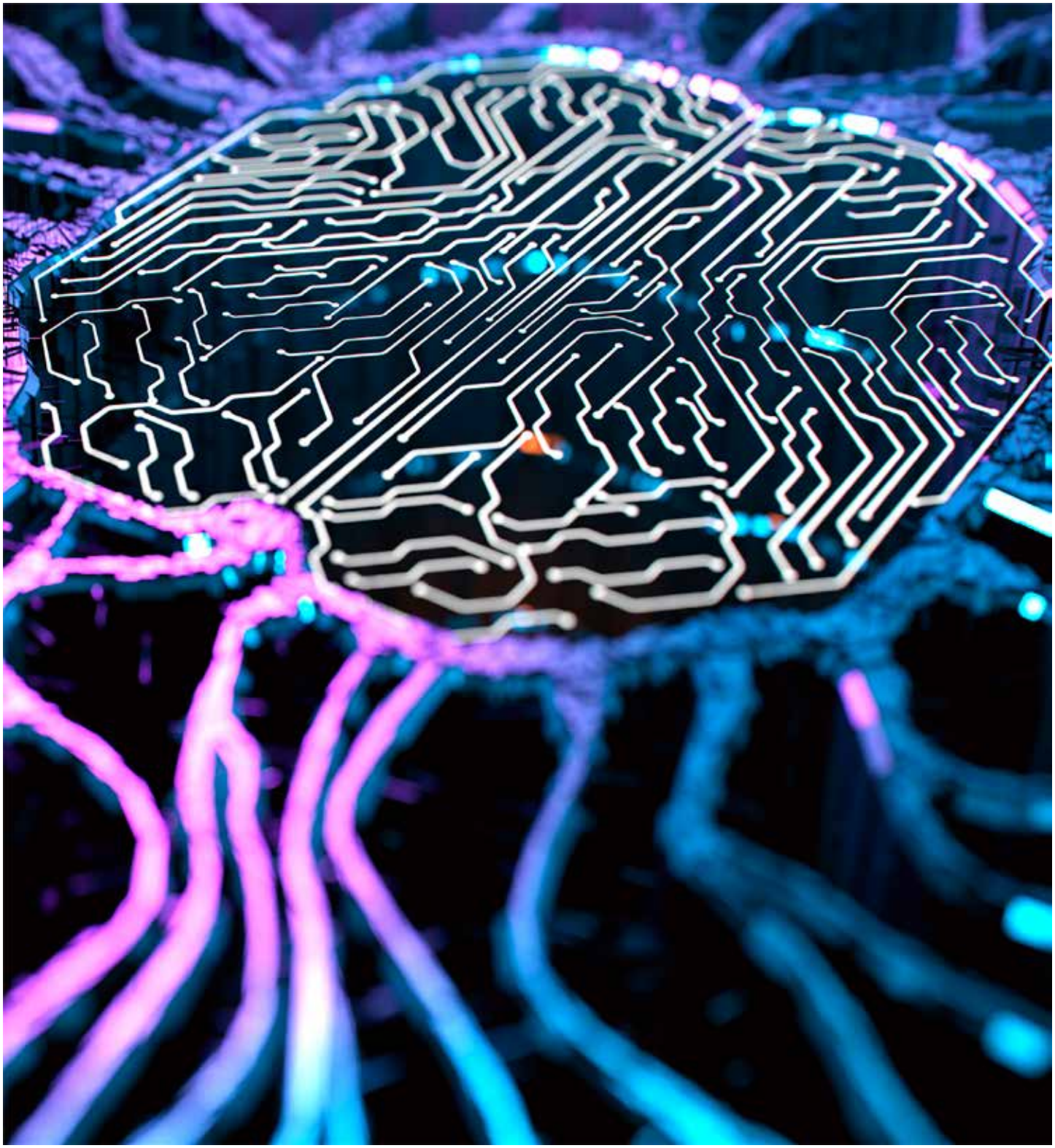
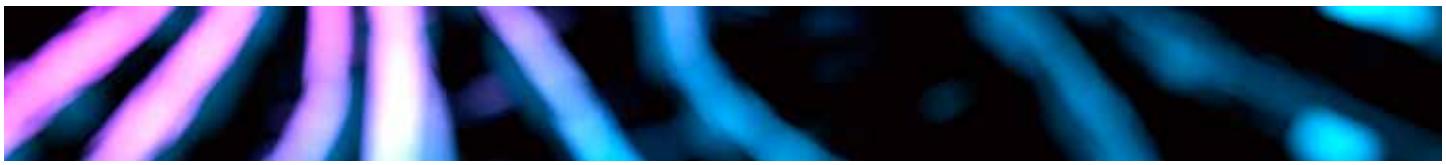
แนวทางการพัฒนาระบบ AI ที่ปลอดภัย





National Cyber Security Centre
a part of GCHQ

CISA
CYBERSECURITY & INFRASTRUCTURE SECURITY AGENCY

Australian Government
Australian Signals Directorate

ASD AUSTRALIAN SIGNALS DIRECTORATE
ACSC Australian Cyber Security Centre

Communications Security Establishment
Canadian Centre for Cyber Security
Centre de la sécurité des télécommunications
Centre canadien pour la cybersécurité

CSIRT
Equipo de Respuesta ante Incidentes de Seguridad Informática

National Cyber and Information Security Agency

Republic of Estonia
Information System Authority

Department of Justice
Federal Bureau of Investigation

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

Federal Office for Information Security

INCD
Israel National Cyber Directorate

AGENZIA PER LA CYBERSICUREZZA NAZIONALE
ACN
SAPIENTIA ADIUVAT

NISC
内閣サイバーセキュリティセンター
National center of Incident readiness and Strategy for Cybersecurity

National Cyber Security Centre

NITDA

NSM
NORWEGIAN NATIONAL CYBER SECURITY CENTRE

National Security Agency
United States of America

NASK

국가정보원
National Intelligence Service

Ministerstwo Cyfryzacji

CSA SINGAPORE
Cyber Security Agency of Singapore

เกี่ยวกับเอกสารนี้

เอกสารนี้เผยแพร่โดยศูนย์รักษาความมั่นคงปลอดภัยทางไซเบอร์แห่งชาติสหราชอาณาจักร (UK National Cyber Security Centre - NCSC) คณะกรรมาธิการความมั่นคงปลอดภัยทางไซเบอร์และโครงสร้างพื้นฐานแห่งสหรัฐอเมริกา (US Cybersecurity and Infrastructure Security Agency - CISA) และพันธมิตรระหว่างประเทศดังนี้

- คณะกรรมาธิการความมั่นคงแห่งชาติ (National Security Agency - NSA)
- สำนักงานสอบสวนกลาง (Federal Bureau of Investigation - FBI)
- ศูนย์รักษาความมั่นคงปลอดภัยทางไซเบอร์แห่งออสเตรเลีย (Australian Cyber Security Centre - ACSC) ของศูนย์อำนวยการสัญญาณข่าวกรองออสเตรเลีย (Australian Signals Directorate)
- ศูนย์รักษาความมั่นคงปลอดภัยทางไซเบอร์แห่งแคนาดา (Canadian Centre for Cyber Security - CCCS)
- ศูนย์รักษาความมั่นคงปลอดภัยทางไซเบอร์แห่งชาตินิวซีแลนด์ (New Zealand National Cyber Security Centre - NCSC-NZ)
- CSIRT แห่งรัฐบาลชิลี (Chile's Government CSIRT)
- คณะกรรมาธิการความมั่นคงปลอดภัยทางไซเบอร์และข้อมูลแห่งชาติสาธารณรัฐเช็ก (Czech Republic's National Cyber and Information Security Agency - NUKIB)
- หน่วยงานระบบสารสนเทศเอสโตเนีย (Information System Authority of Estonia (RIA) และศูนย์รักษาความมั่นคงปลอดภัยทางไซเบอร์แห่งชาติเอสโตเนีย (National Cyber Security Centre of Estonia - NCSC-EE)
- คณะกรรมาธิการความมั่นคงปลอดภัยทางไซเบอร์แห่งฝรั่งเศส (French Cybersecurity Agency - ANSSI)
- สำนักงานความมั่นคงปลอดภัยของข้อมูลแห่งสหพันธรัฐเยอรมนี (Germany's Federal Office for Information Security - BSI)
- คณะกรรมาธิการไซเบอร์แห่งชาติอิสราเอล (Israel's National Cyber Directorate - INCD)
- คณะกรรมาธิการความมั่นคงปลอดภัยทางไซเบอร์แห่งชาติอิตาลี (Italian National Cybersecurity Agency - ACN)
- ศูนย์เตรียมความพร้อมเหตุการณ์และยุทธศาสตร์เพื่อความมั่นคงปลอดภัยทางไซเบอร์แห่งชาติญี่ปุ่น (Japan's National Center of Incident Readiness and Strategy for Cybersecurity - NISC)
- สำนักเลขาธิการนโยบายวิทยาศาสตร์ เทคโนโลยี และนวัตกรรมของญี่ปุ่น สำนักงานคณะรัฐมนตรี (Japan's Secretariat of Science, Technology and Innovation Policy, Cabinet Office)
- คณะกรรมาธิการพัฒนาเทคโนโลยีสารสนเทศแห่งชาติไนจีเรีย (Nigeria's National Information Technology Development Agency - NITDA)
- ศูนย์รักษาความมั่นคงปลอดภัยทางไซเบอร์แห่งชาตินอร์เวย์ (Norwegian National Cyber Security Centre - NCSC-NO)
- กระทรวงกิจการดิจิทัลแห่งโปแลนด์ (Poland Ministry of Digital Affairs)
- สถาบันวิจัยแห่งชาติ NASK ของโปแลนด์ (Poland's NASK National Research Institute - NASK)
- หน่วยข่าวกรองแห่งชาติสาธารณรัฐเกาหลี (Republic of Korea National Intelligence Service - NIS)
- คณะกรรมาธิการความมั่นคงปลอดภัยทางไซเบอร์แห่งสิงคโปร์ (Cyber Security Agency of Singapore - CSA)

กิตติกรรมประกาศ

องค์กรต่อไปนี้มีส่วนร่วมในการพัฒนาแนวทางเหล่านี้

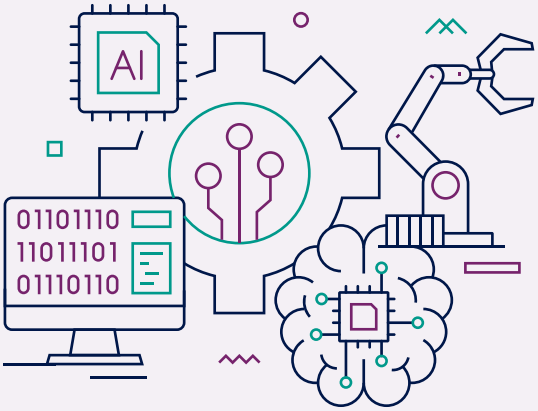
- สถาบัน Alan Turing (Alan Turing Institute)
- Anthropic
- Databricks
- ศูนย์รักษาความมั่นคงปลอดภัยและเทคโนโลยีอุบัติใหม่ (Center for Security and Emerging Technology) ของมหาวิทยาลัย Georgetown
- Google
- Google DeepMind
- IBM
- ImBue
- Microsoft
- OpenAI
- Palantir
- RAND
- Scale AI
- สถาบันวิศวกรรมซอฟต์แวร์ที่มหาวิทยาลัยคาร์เนกีเมลลอน (Software Engineering Institute at Carnegie Mellon University)
- ศูนย์สแตนฟอร์ดเพื่อความมั่นคงของ AI (Stanford Center for AI Safety)
- หลักสูตรสแตนฟอร์ดด้านภูมิรัฐศาสตร์ เทคโนโลยี และธรรมาภิบาล (Stanford Program on Geopolitics, Technology and Governance)

ข้อจำกัดความรับผิดชอบ

NCSC จัดทำข้อมูลที่ปรากฏในเอกสารนี้แบบ "ตามสภาพปัจจุบัน" และคณะผู้จัดทำเอกสารจะไม่ขอรับผิดชอบต่อความสูญเสีย การบาดเจ็บหรือความเสียหายใด ๆ ที่เกิดจากการใช้งานข้อมูลในเอกสารนี้ ยกเว้นตามที่กฎหมายกำหนด NCSC ไม่ขอรับรองหรือแนะนำองค์กร ผลิตภัณฑ์หรือบริการของบุคคลที่สามใด ๆ และหน่วยงานที่มีอำนาจที่ข้อมูลในเอกสารนี้กล่าวอ้างถึง ลิงก์และการอ้างอิงไปยังเว็บไซต์ และเอกสารของบุคคลที่สามเป็นเพียงการให้ข้อมูลเท่านั้น และไม่ได้หมายความว่าเอกสารนี้ให้การรับรองหรือแนะนำแหล่งข้อมูลเหล่านั้นมากกว่าแหล่งข้อมูลอื่น ๆ

เอกสารนี้จัดทำขึ้นตามมาตรฐาน TLP:CLEAR (https://www.first.org/tlp/)

สารบัญ



บทสรุปสำหรับผู้บริหาร

เอกสารฉบับนี้แนะนำแนวทางสำหรับผู้ให้บริการของระบบใด ๆ ที่ใช้ปัญญาประดิษฐ์ (AI) ไม่ว่าจะสร้างระบบเหล่านั้นตั้งแต่เริ่มต้น หรือใช้เครื่องมือและบริการจากผู้อื่น การนำแนวทางเหล่านี้ไปใช้จะช่วยให้ผู้ให้บริการสร้างระบบ AI ที่ทำงานได้ตามที่ตั้งใจไว้ เป็นระบบที่พร้อมใช้งานเมื่อจำเป็น และดำเนินการโดยไม่เปิดเผยข้อมูลที่ละเอียดอ่อนแก่บุคคลที่ไม่ได้รับอนุญาต

เอกสารฉบับนี้จัดทำขึ้นโดยมุ่งเป้าไปที่ผู้ให้บริการของระบบ AI ที่กำลังใช้โมเดลที่โฮสต์โดยองค์กร หรือกำลังใช้ส่วนต่อประสานโปรแกรมประยุกต์ (Application Programming Interfaces - API) จากภายนอก เราขอเรียกร้องให้ผู้มีส่วนได้ส่วนเสีย**ทุกภาคส่วน** (รวมถึงนักวิทยาศาสตร์ข้อมูล นักพัฒนา ผู้จัดการ ผู้มีอำนาจตัดสินใจ และผู้ที่รับผิดชอบต่อความเสี่ยง) ศึกษาแนวทางเหล่านี้เพื่อช่วยให้พวกเขาตัดสินใจอย่างรอบรู้ในเรื่องที่เกี่ยวกับ**การออกแบบ การพัฒนา การปรับใช้** และ**การทำงาน**ของ ระบบ AI

เกี่ยวกับแนวทางเหล่านี้

ระบบ AI มีศักยภาพที่จะสร้างประโยชน์มากมายให้กับสังคม อย่างไรก็ตาม เพื่อให้โอกาสประโยชน์สูงสุดของ AI เกิดขึ้นจริงอย่างเต็มที่ AI จำเป็นต้องได้รับการพัฒนา ปรับใช้ และดำเนินการในลักษณะที่ปลอดภัยและมีความรับผิดชอบ

ระบบ AI ต้องเผชิญกับความเสี่ยงของช่องโหว่ด้านความปลอดภัยใหม่ ๆ ที่จำเป็นต้องพิจารณาควบคู่ไปกับภัยคุกคามด้านความมั่นคงปลอดภัยทางไซเบอร์ตามมาตรฐานปกติ เมื่อการพัฒนาก้าวไปอย่างรวดเร็ว เช่นเดียวกับกรณีของ AI บ่อยครั้งที่ความปลอดภัยมักถูกมองเป็นเรื่องรอง ความปลอดภัยจำเป็นต้องเป็นข้อกำหนดหลักไม่ใช่แค่ในขั้นตอนการพัฒนา แต่ตลอดวงจรชีวิตของระบบ

ด้วยเหตุนี้ แนวทางนี้จึงถูกแบ่งออกเป็นสี่ส่วนหลัก ๆ ภายในวงจรชีวิตการพัฒนาระบบ AI นั่นคือ **การออกแบบที่ปลอดภัย การพัฒนาที่ปลอดภัย การปรับใช้ที่ปลอดภัย** และ**การทำงานและการบำรุงรักษาที่ปลอดภัย** ในแต่ละส่วนเหล่านี้ เราจะเสนอข้อควรพิจารณาและการดำเนินการที่จะช่วยลดความเสี่ยงโดยรวมต่อกระบวนการพัฒนาระบบ AI ขององค์กร

1. **การออกแบบที่ปลอดภัย**
   ในส่วนนี้จะกล่าวถึงแนวทางที่นำไปใช้กับขั้นตอนการออกแบบของวงจรชีวิตการพัฒนาระบบ AI ซึ่งจะครอบคลุมถึงเรื่องการทำความเข้าใจความเสี่ยงและการสร้างโมเดลภัยคุกคาม ตลอดจนหัวข้อเฉพาะและการประนีประนอมที่ต้องพิจารณาในการออกแบบระบบและโมเดล

2. **การพัฒนาที่ปลอดภัย**
   ในส่วนนี้จะกล่าวถึงแนวทางที่ใช้กับขั้นตอนการพัฒนาของวงจรชีวิตการพัฒนาระบบ AI รวมถึงซัพพลายเชนที่มีความปลอดภัย การจัดทำเอกสารประกอบ และการจัดการสินทรัพย์และหนี้ทางเทคนิค

3. **การปรับใช้ที่ปลอดภัย**
   ในส่วนนี้จะกล่าวถึงแนวทางที่ใช้กับขั้นตอนการปรับใช้ของวงจรชีวิตการพัฒนาระบบ AI รวมถึงการปกป้องโครงสร้างพื้นฐานและโมเดลจากการถูกโจมตี ภัยคุกคามหรือการสูญเสีย การพัฒนากระบวนการจัดการเหตุการณ์ และการทำให้แน่ใจว่าทุกอย่างจะปลอดภัยเมื่อเปิดการใช้งานระบบ

4. **การทำงานและการบำรุงรักษาที่ปลอดภัย**
   ในส่วนนี้จะกล่าวถึงแนวทางที่ใช้กับขั้นตอนการทำงานและการบำรุงรักษาที่ปลอดภัยของวงจรชีวิตการพัฒนาระบบ AI ซึ่งจะให้แนวทางในการดำเนินการที่เกี่ยวข้อง โดยเฉพาะอย่างยิ่งเมื่อมีการปรับใช้ระบบแล้ว รวมถึงการบันทึกและการเฝ้าระวังติดตามสิ่งที่เกิดขึ้นในระบบ การจัดการการอัปเดต และการแบ่งปันข้อมูลสำคัญเกี่ยวกับระบบ

แนวทางดังกล่าวเป็นไปตามแนวปฏิบัติตามหลัก ‘ความปลอดภัยโดยการตั้งค่าเริ่มต้น’ และสอดคล้องกับแนวปฏิบัติที่กำหนดไว้ใน[คำแนะนำการพัฒนาและการปรับใช้งานที่ปลอดภัย]ของ NCSC [กรอบงานพัฒนาซอฟต์แวร์ที่ปลอดภัย]ของ NIST และ ‘[หลักความปลอดภัยโดยการออกแบบ]’ ซึ่งเผยแพร่โดย CISA และ NCSC อีกทั้งหน่วยงานไซเบอร์ระหว่างประเทศ แนวทางเหล่านี้ให้ลำดับความสำคัญดังนี้

- การแสดงความเป็นเจ้าของผลลัพธ์ด้านความปลอดภัยของลูกค้า
- การเปิดรับความโปร่งใสอย่างถึงแก่นและแสดงความรับผิดชอบต่อทุกสิ่งที่ทำ
- การสร้างโครงสร้างองค์กรและความเป็นผู้นำเพื่อให้หลักความปลอดภัยโดยการออกแบบเป็นสิ่งสำคัญอันดับแรกทางธุรกิจ

คำนำ

ระบบปัญญาประดิษฐ์ (AI) มีศักยภาพที่จะสร้างประโยชน์มากมายให้กับสังคม อย่างไรก็ตาม เพื่อให้โอกาสประโยชน์สูงสุดของ AI เกิดขึ้นจริงอย่างเต็มที่ AI จำเป็นต้องได้รับการพัฒนา ปรับใช้ และดำเนินการในลักษณะที่ปลอดภัยและมีความรับผิดชอบ ความมั่นคงปลอดภัยทางไซเบอร์เป็นเงื่อนไขเบื้องต้นที่จำเป็นสำหรับความปลอดภัย ความยืดหยุ่น ความเป็นส่วนตัว ความเป็นธรรม ความมีประสิทธิภาพ และความน่าเชื่อถือของระบบ AI

อย่างไรก็ตาม ระบบ AI ต้องเผชิญกับความเสี่ยงของช่องโหว่ด้านความปลอดภัยใหม่ ๆ ที่จำเป็นต้องพิจารณาควบคู่ไปกับภัยคุกคามด้านความมั่นคงปลอดภัยทางไซเบอร์ตามมาตรฐานปกติ เมื่อการพัฒนาก้าวไปอย่างรวดเร็ว เช่นเดียวกับกรณีของ AI บ่อยครั้งที่ความปลอดภัยมักถูกมองเป็นเรื่องรอง ความปลอดภัยจำเป็นต้องเป็นข้อกำหนดหลักไม่ใช่แค่ในขั้นตอนการพัฒนา แต่ตลอดวงจรชีวิตของระบบ

เอกสารฉบับนี้แนะนำแนวทางสำหรับผู้ให้บริการ[1] ของระบบใด ๆ ที่ใช้ AI ไม่ว่าจะสร้างระบบเหล่านั้นตั้งแต่เริ่มต้นหรือใช้เครื่องมือและบริการจากผู้อื่น การนำแนวทางเหล่านี้ไปใช้จะช่วยให้ผู้ให้บริการสร้างระบบ AI ที่ทำงานได้ตามที่ตั้งใจไว้ เป็นระบบที่พร้อมใช้งานเมื่อจำเป็น และดำเนินการโดยไม่เปิดเผยข้อมูลที่ละเอียดอ่อนแก่บุคคลที่ไม่ได้รับอนุญาต

แนวทางเหล่านี้ควรได้รับการพิจารณาร่วมกับความมั่นคงปลอดภัยทางไซเบอร์ การบริหารความเสี่ยง และแนวปฏิบัติที่ดีที่สุดในการตอบสนองต่อเหตุการณ์ โดยเฉพาะอย่างยิ่ง เราขอเรียกร้องให้ผู้ให้บริการปฏิบัติตามหลัก 'ความปลอดภัยโดยการออกแบบ'[2] ที่พัฒนาโดยคณะกรรมาธิการความมั่นคงปลอดภัยทางไซเบอร์และโครงสร้างพื้นฐานแห่งสหรัฐอเมริกา (CISA) ศูนย์รักษาความมั่นคงปลอดภัยทางไซเบอร์แห่งชาติสหราชอาณาจักร (NCSC) อีกทั้งหน่วยงานไซเบอร์ระหว่างประเทศของเราทั้งหมด หลักการเหล่านี้ให้ลำดับความสำคัญดังนี้

- การแสดงความเป็นเจ้าของผลลัพธ์ด้านความปลอดภัยของลูกค้า
- การเปิดรับความโปร่งใสอย่างถึงแก่นและแสดงความรับผิดชอบต่อทุกสิ่งที่ทำ
- การสร้างโครงสร้างองค์กรและความเป็นผู้นำเพื่อให้หลักความปลอดภัยโดยการออกแบบเป็นสิ่งสำคัญอันดับแรกทางธุรกิจ

การปฏิบัติตามหลัก 'ความปลอดภัยโดยการออกแบบ' ต้องใช้ทรัพยากรจำนวนมากตลอดวงจรชีวิตของระบบ ซึ่งหมายความว่านักพัฒนาซอฟต์แวร์จะต้องลงทุนในการจัดลำดับความสำคัญของ**คุณลักษณะ กลไก** และ**การปรับใช้**ของเครื่องมือที่ใช้ปกป้องลูกค้าในทุกระดับชั้นของการออกแบบ และตลอดทุกขั้นตอนของวงจรชีวิตการพัฒนาระบบ การทำเช่นนี้จะช่วยประหยัดค่าใช้จ่ายโดยไม่ต้องออกแบบใหม่ในภายหลัง รวมถึงช่วยให้แน่ใจว่าลูกค้าและข้อมูลของพวกเขาได้รับการปกป้องในระยะสั้น

เหตุใดความปลอดภัยของระบบ AI จึงมีความแตกต่าง

ในเอกสารนี้ เราใช้คำว่า 'AI' เพื่อกล่าวอ้างถึงแอปพลิเคชันการเรียนรู้ของเครื่อง (Machine learning - ML)[3] เป็นการเฉพาะ ML ทุกประเภทได้รับการพิจารณาให้รวมอยู่ในบริบทของเอกสารนี้ เราให้นิยามของแอปพลิเคชัน ML ว่าเป็นแอปพลิเคชันที่

- เกี่ยวข้องกับส่วนประกอบของซอฟต์แวร์ (โมเดลต่าง ๆ) ที่ช่วยให้คอมพิวเตอร์รับรู้และนำบริบทมาสู่รูปแบบข้อมูลโดยที่มนุษย์ไม่ต้องเขียนโปรแกรมไว้อย่างชัดเจนทั้งหมด
- เพื่อสร้างการคาดการณ์ ให้คำแนะนำ หรือตัดสินใจบนพื้นฐานของเหตุผลทางสถิติ

เช่นเดียวกับภัยคุกคามความมั่นคงปลอดภัยทางไซเบอร์ ระบบ AI ก็สามารถเผชิญกับช่องโหว่ประเภทใหม่ ๆ ได้เช่นกัน คำว่า 'การเรียนรู้ของเครื่องที่ขัดแย้งกัน (Adversarial machine learning - AML)' อธิบายถึงการใช้ประโยชน์จากช่องโหว่ขั้นพื้นฐานในส่วนประกอบ ML รวมถึงฮาร์ดแวร์ ซอฟต์แวร์ กระบวนการเวิร์กโฟลว์ และซัพพลายเชน AML ช่วยให้ผู้โจมตีสร้างระบบการเรียนรู้ของเครื่องให้เกิดพฤติกรรมที่ไม่ได้ตั้งใจในระบบ ML ซึ่งอาจรวมถึง

- การส่งผลต่อการจัดหมวดหมู่หรือประสิทธิภาพการคาดการณ์ของโมเดล
- การอนุญาตให้ผู้ใช้ทำสิ่งที่พวกเขาไม่ควรทำหรือไม่ได้รับอนุญาตให้ทำ
- การดึงข้อมูลที่เป็นความลับหรือสำคัญออกจากโมเดล

มีหลายวิธีในการบรรลุผลเหล่านี้ เช่น การโจมตีแบบอัดฉีดคำสั่งพรอมต์เตือนในโดเมนของโมเดลภาษาขนาดใหญ่ (Large language model - LLM) หรือการจงใจทำให้ข้อมูลการฝึกระบบหรือข้อเสนอแนะที่ได้รับจากผู้ใช้เสียหาย (เรียกว่า 'การทำให้ข้อมูลเป็นพิษ')

ใครควรอ่านเอกสารนี้

เอกสารฉบับนี้จัดทำขึ้นโดยมุ่งเป้าไปที่ผู้ให้บริการของระบบ AI ไม่ว่าจะใช้โมเดลที่โฮสต์โดยองค์กร หรือใช้งานส่วนต่อประสานโปรแกรมประยุกต์ (Application Programming Interfaces - APIs) จากภายนอก เราขอเรียกร้องให้ผู้มีส่วนได้ส่วนเสีย**ทุกภาคส่วน** (รวมถึงนักวิทยาศาสตร์ข้อมูล นักพัฒนา ผู้จัดการ ผู้มีอำนาจตัดสินใจ และผู้ที่รับผิดชอบต่อความเสี่ยง) ศึกษาแนวทางเหล่านี้เพื่อช่วยให้พวกเขาตัดสินใจอย่างรอบรู้ในเรื่องที่เกี่ยวกับ**การออกแบบ การปรับใช้** และ**การทำงาน**ของระบบ AI ในการเรียนรู้ของเครื่อง

อย่างไรก็ตาม ไม่ใช่ทุกแนวทางจะสามารถนำไปใช้กับทุกองค์กรได้โดยตรง ระดับความซับซ้อนและวิธีการโจมตีจะแตกต่างกันไปขึ้นอยู่กับฝ่ายตรงข้ามใดที่พยายามโจมตีระบบ AI ดังนั้น เมื่อใช้แนวทางเหล่านี้ให้พิจารณาว่าองค์กรของคุณใช้ระบบ AI อย่างไร และภัยคุกคามใดบ้างที่องค์กรของคุณอาจเผชิญอยู่

ใครบ้างควรเป็นผู้รับผิดชอบในการพัฒนา AI ที่ปลอดภัย

ผู้คนจำนวนมากมีบทบาทในซัพพลายเชนของระบบ AI ในปัจจุบัน สามารถกล่าวได้คร่าว ๆ ว่ามีบุคคลอยู่สองกลุ่ม อันได้แก่

- 'ผู้ให้บริการ' เป็นผู้รับผิดชอบในการจัดการข้อมูล การพัฒนาอัลกอริทึม การออกแบบ การปรับใช้ และการบำรุงรักษา
- 'ผู้ใช้บริการ' คือ ผู้ให้ข้อมูลกับระบบ (Input) และผู้ได้รับผลลัพธ์จากระบบ (Output)

แม้ว่าแนวปฏิบัติระหว่างผู้ให้บริการกับผู้ใช้เช่นนี้จะถูกนำมาใช้ในแอปพลิเคชันจำนวนมาก แต่ก็กลายเป็นเรื่องที่พบได้น้อยลงเรื่อย ๆ[4] เนื่องจากผู้ให้บริการมีแนวโน้มที่จะมองหาการรวมซอฟต์แวร์ ข้อมูล โมเดล และ/หรือบริการระยะไกลจากบริษัทอื่น ๆ เพื่อไว้ในระบบของตนเอง ซัพพลายเชนที่ซับซ้อนเหล่านี้ทำให้ยากสำหรับผู้ใช้ปลายทางที่จะเข้าใจว่าความรับผิดชอบสำหรับ AI ที่ปลอดภัยอยู่ที่ไหน

ผู้ใช้ (ไม่ว่าจะเป็น 'ผู้ใช้ปลายทาง' หรือผู้ให้บริการที่รวมส่วนประกอบ AI จากภายนอกไว้ในระบบ[5]) มักจะไม่มีความรู้เชิงลึก และ/หรือความเชี่ยวชาญเพียงพอที่จะทำความเข้าใจ ประเมิน หรือจัดการกับความเสี่ยงที่เกี่ยวข้องกับระบบที่พวกเขาใช้งานอยู่ ด้วยเหตุนี้ เพื่อให้สอดคล้องตามหลัก 'ความปลอดภัยโดยการออกแบบ' **ผู้ให้บริการของส่วนประกอบ AI ควรเป็นผู้รับผิดชอบในการรับรองความปลอดภัยของผู้ใช้ที่อยู่ถัดไปในซัพพลายเชน**

ผู้ให้บริการควรเพิ่มมาตรการป้องกันด้วยการควบคุมและแก้ไขด้านความปลอดภัยในโมเดล ไปป์ไลน์ และ/หรือระบบของตน และหากมีการตั้งค่าที่ปรับได้ก็ควรตั้งค่าตัวเลือกที่ปลอดภัยที่สุดเป็นค่าเริ่มต้น ในกรณีที่ไม่สามารถลดความเสี่ยงได้ ผู้ให้บริการควรรับผิดชอบในเรื่องดังนี้

- การแจ้งให้ผู้ใช้ที่อยู่ในซัพพลายเชนทราบเพิ่มเติมเกี่ยวกับความเสี่ยงที่อาจเกิดขึ้นและความเสี่ยงใด ๆ ที่ผู้ใช้ของตนเองอาจต้องเผชิญ (ถ้ามี)
- การให้คำแนะนำเกี่ยวกับวิธีการใช้ส่วนประกอบในลักษณะที่ปลอดภัย

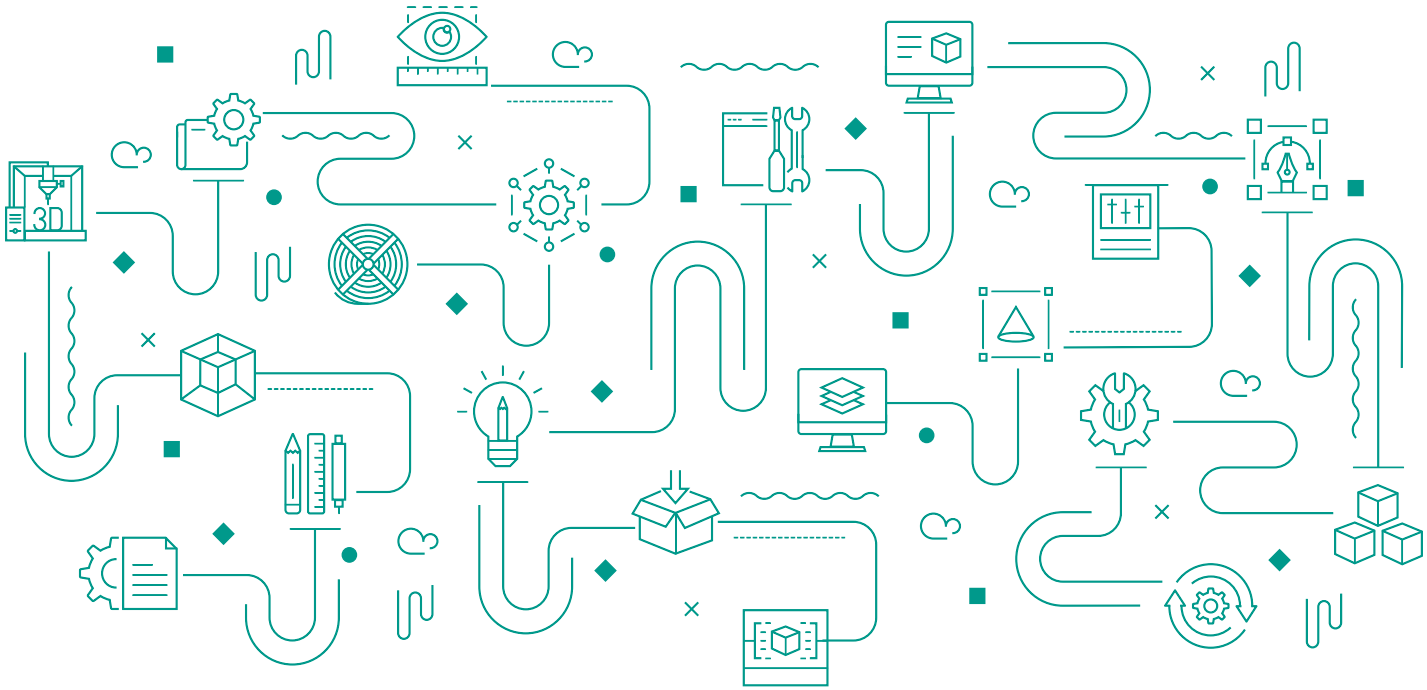
ในกรณีที่การถูกละเมิดความปลอดภัยนำไปสู่ความเสียหายทางกายภาพ ทำลายชื่อเสียง รบกวนการดำเนินธุรกิจทำให้เกิดความเสียหายอย่างมีนัยสำคัญ เกิดการรั่วไหลของข้อมูลที่ละเอียดอ่อนหรือเป็นความลับ และ/หรือมีผลกระทบทางกฎหมาย ความเสี่ยงด้านความมั่นคงปลอดภัยทางไซเบอร์ก็ควรได้รับการพิจารณาว่า**มีความสำคัญยิ่ง**

แนวทางการพัฒนาระบบ AI ที่ปลอดภัย

แนวทางนี้จึงถูกแบ่งออกเป็นสี่ส่วนหลัก ๆ ภายในวงจรชีวิตการพัฒนาระบบ AI นั่นคือ **การออกแบบที่ปลอดภัย การพัฒนาที่ปลอดภัย การปรับใช้ที่ปลอดภัย** และ**การทำงานและการบำรุงรักษาที่ปลอดภัย** ในแต่ละส่วนเหล่านี้ เราจะเสนอข้อควรพิจารณาและการดำเนินการที่จะช่วยลดความเสี่ยงโดยรวมต่อกระบวนการพัฒนาระบบ AI ขององค์กร

แนวทางที่วางไว้ในเอกสารนี้มีความสอดคล้องใกล้เคียงกับแนวปฏิบัติของวงจรชีวิตการพัฒนาซอฟต์แวร์ ซึ่งกำหนดไว้ในเอกสารดังนี้

- คำแนะนำการพัฒนาและการปรับใช้งานที่ปลอดภัยของ NCSC
- กรอบงานพัฒนาซอฟต์แวร์ที่ปลอดภัย (SSDF)[6] ของสถาบันมาตรฐานและเทคโนโลยีแห่งชาติ (NIST)



1. การออกแบบที่ปลอดภัย

ในส่วนนี้จะกล่าวถึงแนวทางที่นำไปใช้กับขั้นตอน**การออกแบบ**ของวงจรชีวิตการพัฒนาระบบ AI ซึ่งจะครอบคลุมถึงเรื่องการทำความเข้าใจความเสี่ยง และการสร้างโมเดลภัยคุกคาม ตลอดจนหัวข้อเฉพาะและการประนีประนอมที่ต้องพิจารณาในการออกแบบระบบและโมเดล

ยกระดับความตระหนักรู้แก่พนักงานเกี่ยวกับภัยคุกคามและความเสี่ยง

เจ้าของระบบและผู้นำระดับสูงเข้าใจถึงภัยคุกคามที่มีต่อการรักษาความปลอดภัยของ AI และวิธีการบรรเทาผลกระทบของภัยคุกคาม นักวิทยาศาสตร์ข้อมูลและนักพัฒนาระบบของคุณตระหนักถึงภัยคุกคามด้านความปลอดภัยและโหมดความล้มเหลวที่เกี่ยวข้อง และช่วยเหลือเจ้าของความเสี่ยงในการตัดสินใจโดยมีข้อมูลที่ครบถ้วน คุณให้คำแนะนำผู้ใช้เกี่ยวกับความเสี่ยงด้านความปลอดภัยเฉพาะที่เกี่ยวข้องกับระบบ AI (เช่น รวมไว้เป็นส่วนหนึ่งของการอบรมระบบ InfoSec มาตรฐาน) และฝึกอบรมนักพัฒนาให้ใช้เทคนิคการเขียนโค้ดด้วยวิธีที่ปลอดภัยและทำตามแนวปฏิบัติ AI ที่ปลอดภัยและมีความรับผิดชอบในการทำงาน

สร้างโมเดลจำลองที่เป็นภัยคุกคามต่อระบบของคุณ

ในฐานะส่วนหนึ่งของกระบวนการจัดการความเสี่ยง คุณจะต้องประเมินภัยคุกคามต่อระบบของคุณเป็นกระบวนการแบบองค์รวมโดยละเอียด ซึ่งรวมถึงการทำความเข้าใจผลกระทบที่อาจเกิดขึ้นกับระบบ ผู้ใช้ องค์กร และสังคมในวงกว้าง หากส่วนประกอบ AI ถูกบุกรุกโดยไม่คาดคิดหรือทำงานโดยไม่เป็นไปตามความคาดหมาย[7] กระบวนการนี้เกี่ยวข้องกับการประเมินผลกระทบจากภัยคุกคามต่อ AI โดยเฉพาะ[8] และบันทึกการตัดสินใจของคุณตามผลการประเมินนั้น

คุณรับรู้ว่าความละเอียดอ่อนและประเภทของข้อมูลที่ใช้ในระบบของคุณอาจส่งผลให้เป็นเป้าหมายดึงดูดผู้โจมตีได้ การประเมินของคุณควรพิจารณาว่าภัยคุกคามบางอย่างอาจเพิ่มสูงขึ้นในขณะที่ระบบ AI ถูกมองว่าเป็นเป้าหมายที่มีคุณค่าเพิ่มมากขึ้น และในขณะที่ AI เองก็เปิดใช้งานเวกเตอร์สำหรับการโจมตีแบบอัตโนมัติตัวใหม่

ออกแบบระบบของคุณเพื่อความปลอดภัยตลอดจนฟังก์ชันการทำงานและประสิทธิภาพ

คุณมั่นใจว่างานที่ทำอยู่ได้รับการจัดการแก้ไขอย่างเหมาะสมที่สุดโดยใช้ AI หลังจากพิจารณาเรื่องนี้แล้ว คุณประเมินว่าตัวเลือกการออกแบบเฉพาะของ AI นั้นเหมาะสมหรือไม่ คุณพิจารณาโมเดลภัยคุกคามและการลดความเสี่ยงด้านความปลอดภัยที่เกี่ยวข้องควบคู่ไปกับฟังก์ชันการทำงาน ประสบการณ์ผู้ใช้ สภาพแวดล้อมการปรับใช้ ประสิทธิภาพ การรับประกัน การกำกับดูแล ข้อกำหนดด้านจริยธรรม และกฎหมายพร้อมกับข้อพิจารณาอื่น ๆ ตัวอย่าง เช่น

- คุณพิจารณาความปลอดภัยของซัพพลายเชน เมื่อต้องเลือกว่าจะพัฒนาระบบภายในองค์กรหรือใช้ส่วนประกอบจากภายนอก เช่น
    - การที่คุณเลือกที่จะฝึกระบบโมเดลใหม่หรือใช้โมเดลเดิมที่มีอยู่ (ไม่ว่าจะมีการปรับแต่งรายละเอียดหรือไม่ก็ตาม) หรือใช้โมเดลผ่านบริการ API จากภายนอกนั้นเหมาะสมกับความต้องการของคุณ
    - การที่คุณเลือกที่จะทำงานร่วมกับผู้ให้บริการโมเดลจากภายนอก คุณต้องประเมินสถานะความปลอดภัยของผู้ให้บริการรายนั้นด้วย
    - เมื่อใช้ไลบรารีจากภายนอก คุณต้องตรวจสอบอย่างละเอียด (เช่น ให้แน่ใจว่าไลบรารีมีมาตรการควบคุมด้านความปลอดภัยที่จะป้องกันไม่ให้ระบบโหลดโมเดลที่ไม่น่าเชื่อถือโดยไม่ได้ตั้งใจ ซึ่งอาจนำไปสู่การโจมตีแบบ arbitrary code execution ได้[9])
    - คุณใช้การสแกนและการแยกแยะ/แซนด์บ็อกซ์ เมื่อนำเข้าโมเดลของบริษัทอื่นหรือใช้การเรียงบันทึกค่าน้ำหนักของโมเดล ซึ่งควรถือว่าเป็นโค้ดของบริษัทภายนอกที่ไม่น่าเชื่อถือ และอาจมีการโจมตีโดยเปิดใช้งานจากระยะไกลได้

  - หากใช้ API ภายนอก คุณใช้มาตรการควบคุมที่เหมาะสมกับข้อมูลที่สามารถส่งไปยังบริการที่อยู่นอกการควบคุมขององค์กรของคุณได้ เช่น กำหนดให้ผู้ใช้เข้าสู่ระบบและยืนยันก่อนส่งข้อมูลที่มีแนวโน้มว่าเป็นข้อมูลละเอียดอ่อน
  - คุณใช้การตรวจสอบและการคัดกรองข้อมูลและอินพุตที่จะเข้าสู่ระบบอย่างเหมาะสม ซึ่งรวมถึงเมื่อคุณรวมคำติชมจากผู้ใช้หรือข้อมูลการเรียนรู้ใหม่ลงในโมเดลของคุณ โดยรับรู้ว่าข้อมูลที่ใช้ในการฝึกระบบจะกำหนดพฤติกรรมหรือลักษณะการทำงานของระบบ

- คุณรวมการพัฒนาระบบซอฟต์แวร์ AI เข้ากับแนวทางปฏิบัติที่ดีที่สุดในการพัฒนาและการดำเนินงานที่ปลอดภัยที่คุณใช้อยู่ ซึ่งหมายถึงการทำให้แน่ใจว่าองค์ประกอบทุกส่วนของระบบ AI ถูกเขียนขึ้นในสภาพแวดล้อมที่เหมาะสมโดยใช้การปฏิบัติและภาษาการเขียนโค้ดที่จะลดหรือกำจัดประเภทของช่องโหว่ที่ทราบอยู่แล้วทุกครั้งที่เป็นไปได้

- เมื่อส่วนประกอบ AI จำเป็นต้องกระตุ้นให้ทำสิ่งต่าง ๆ เช่น แก้ไขไฟล์หรือกำหนดเอาต์พุตไปยังระบบภายนอก คุณตั้งค่าขีดจำกัดที่เหมาะสมกับการดำเนินการที่เป็นไปได้ (ซึ่งรวมถึง AI ภายนอกและระบบป้องกันความล้มเหลวที่ไม่ใช่ AI หากจำเป็น)

- ในการตัดสินใจว่าผู้ใช้โต้ตอบกับระบบอย่างไร คุณพิจารณาความเสี่ยงที่เกี่ยวข้องกับ AI เป็นเฉพาะ ตัวอย่างเช่น

  - ระบบของคุณให้ข้อมูลเอาต์พุตแก่ผู้ใช้ในลักษณะที่ใช้งานได้ โดยไม่ต้องให้รายละเอียดมากเกินความจำเป็นที่ผู้โจมตีระบบอาจนำไปใช้ประโยชน์ได้
  - หากจำเป็น ระบบของคุณกำหนดขอบเขตการป้องกันที่มีประสิทธิภาพรายรอบเอาต์พุตจากโมเดล
  - หากคุณแชร์ระบบของคุณกับลูกค้าภายนอกหรือผู้ทำงานร่วมกันผ่าน API คุณใช้มาตรการควบคุมที่เหมาะสมเพื่อป้องกันการโจมตีระบบ AI ผ่าน API นั้น
  - คุณรวมการตั้งค่าที่ปลอดภัยที่สุดเข้าในระบบตามการตั้งค่าเริ่มต้นไว้โดยอัตโนมัติ
  - คุณใช้หลักการสิทธิพิเศษน้อยที่สุดเพื่อจำกัดการเข้าถึงฟังก์ชันการทำงานของระบบเฉพาะสิ่งที่จำเป็นจริง ๆ เท่านั้น
  - คุณอธิบายให้ผู้ใช้ทราบเกี่ยวกับคุณลักษณะที่มีความเสี่ยงสูงกว่า และกำหนดให้ผู้ใช้ต้องเป็นผู้เลือกใช้งานเอง คุณยังแจ้งให้พวกเขาทราบด้วยว่าห้ามใช้งานอะไร และหากเป็นไปได้ แจ้งให้ผู้ใช้ทราบถึงทางเลือกอื่น ๆ ในการแก้ปัญหา

พิจารณาสิทธิประโยชน์ด้านความปลอดภัยและการประนีประนอมเมื่อเลือกโมเดล AI ของคุณ

โมเดล AI ที่คุณเลือกจะเกี่ยวข้องกับการปรับสมดุลข้อกำหนดต่าง ๆ ซึ่งรวมถึงตัวเลือกสถาปัตยกรรมโมเดล การกำหนดตั้งค่าข้อมูลในการฝึกระบบ อัลกอริทึมของการฝึกระบบ และไฮเปอร์พารามิเตอร์ การตัดสินใจของคุณจะขึ้นอยู่กับโมเดลภัยคุกคามของคุณและการได้รับแจ้งจากการประเมินใหม่เป็นประจำ ในขณะที่ความรู้เกี่ยวกับความปลอดภัยของ AI มีความก้าวหน้าขึ้นและความเข้าใจเกี่ยวกับอันตรายที่อาจเกิดขึ้นก็พัฒนาขึ้นตามเวลา

เมื่อเลือกโมเดล AI คุณควรคำนึงถึงปัจจัยต่าง ๆ ซึ่งอาจรวมถึงแต่ไม่จำกัดเพียงปัจจัยดังนี้

- ความซับซ้อนของโมเดลที่คุณใช้ นั่นคือ สถาปัตยกรรมที่เลือกและจำนวนส่วนประกอบพารามิเตอร์ โดยที่สถาปัตยกรรมและจำนวนพารามิเตอร์ที่เลือกของโมเดลของคุณจะส่งผลดีเพียงใดต่อปริมาณข้อมูลการฝึกระบบที่ต้องใช้ และความสามารถในการจัดการกับข้อมูลอินพุตต่าง ๆ ในขณะใช้งาน นอกเหนือไปจากปัจจัยอื่น ๆ

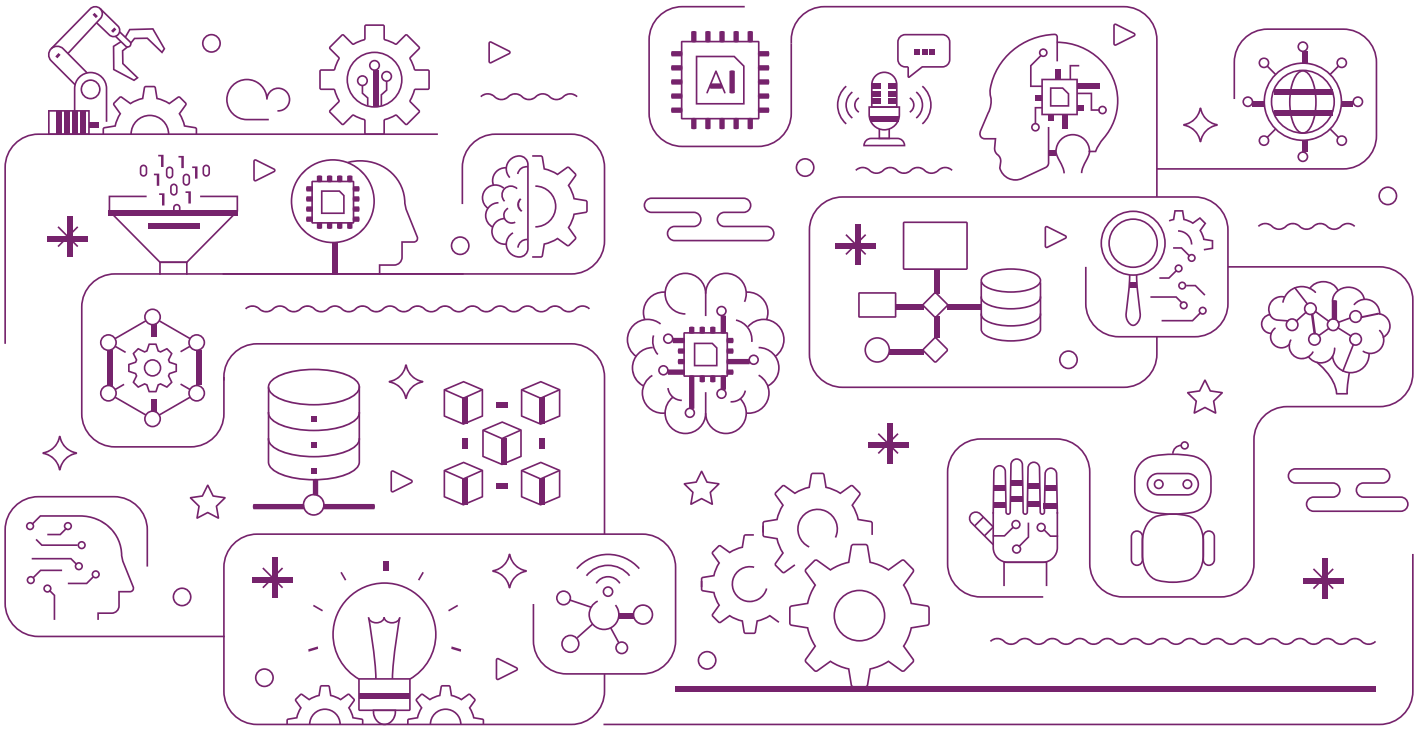
- ตรวจสอบว่าโมเดลที่คุณใช้เหมาะสมกับสิ่งที่คุณกำลังพยายามทำ และ/หรือสามารถปรับให้เข้ากับความต้องการเฉพาะของคุณหรือไม่ (เช่น โดยการปรับแต่งรายละเอียด)

- ความสามารถในการจัดตำแหน่ง ตีความ และอธิบายผลลัพธ์เอาต์พุตของโมเดลของคุณ (เช่น เพื่อการดีบัก (Debug) หรือแก้ไขปัญหา การตรวจสอบหรือการปฏิบัติตามข้อกำหนด) การใช้โมเดลที่เรียบง่ายและโปร่งใสมากขึ้นอาจดีกว่าในบางกรณี เนื่องจากโมเดลเหล่านี้ตีความได้ง่ายกว่าเมื่อเทียบกับโมเดลขนาดใหญ่และซับซ้อน

- ลักษณะเฉพาะตัวของชุดข้อมูลการฝึกระบบ รวมถึงขนาดของข้อมูล ความสมบูรณ์ คุณภาพ ความละเอียดอ่อน อายุ ความเกี่ยวข้อง และความหลากหลายของข้อมูล

- คุณค่าของการใช้โมเดลเสริมความแข็งแกร่ง (เช่นการฝึกระบบให้รู้จักป้องกันฝ่ายตรงข้ามหรือผู้โจมตี) การทำให้เป็นมาตรฐาน และ/หรือเทคนิคการเพิ่มความเป็นส่วนตัว
- คุณควรทำความเข้าใจแหล่งที่มาและซัพพลายเชนของส่วนประกอบของระบบว่ามาจากไหนและอย่างไร รวมถึงโมเดลหรือโมเดลพื้นฐาน ข้อมูลการฝึกระบบ และเครื่องมือที่เกี่ยวข้อง

สำหรับข้อมูลเพิ่มเติมเกี่ยวกับปัจจัยต่าง ๆ เหล่านี้ที่ส่งผลต่อผลลัพธ์ด้านความปลอดภัย โปรดดู 'หลักการรักษาระบบการเรียนรู้ของเครื่องให้ปลอดภัย' ของ NCSC โดยเฉพาะหัวข้อ การออกแบบเพื่อความปลอดภัย (สถาปัตยกรรมโมเดล) (Design for security (model architecture))



2. การพัฒนาที่ปลอดภัย

ในส่วนนี้จะกล่าวถึงแนวทางที่ใช้กับขั้นตอน**การพัฒนา**ของวงจรชีวิตการพัฒนาระบบ AI รวมถึงซัพพลายเชนที่มีความปลอดภัย การจัดทำเอกสารประกอบ และการจัดการสินทรัพย์และหนี้ทางเทคนิค

ทำให้ซัพพลายเชนของคุณปลอดภัย

คุณประเมินและติดตามความปลอดภัยของซัพพลายเชน AI ของคุณตลอดวงจรชีวิตของระบบ และกำหนดให้ซัพพลายเออร์ปฏิบัติตามมาตรฐานเดียวกันกับที่องค์กรของคุณใช้กับซอฟต์แวร์อื่น ๆ หากซัพพลายเออร์ไม่สามารถปฏิบัติตามมาตรฐานขององค์กรของคุณได้ คุณจะต้องปฏิบัติตามนโยบายการจัดการความเสี่ยงที่มีอยู่

ในกรณีที่ไม่ได้ผลิตเองในองค์กร คุณควรได้รับและควรรักษาส่วนประกอบฮาร์ดแวร์และซอฟต์แวร์จากแหล่งที่เชื่อถือได้ว่ามีการรักษาความปลอดภัยดีและมีเอกสารประกอบเป็นอย่างดี (เช่น โมเดล ข้อมูล ไลบรารีซอฟต์แวร์ มอดูล มิดเดิลแวร์ เฟรมเวิร์ก และ API ภายนอก) จากผู้ให้บริการเชิงพาณิชย์ที่ได้รับการยืนยัน ชุมชนโอเพ่นซอร์ส หรือนักพัฒนาบุคคลที่สามที่เชื่อถือได้อื่น ๆ เพื่อให้แน่ใจว่ามีความปลอดภัยที่แข็งแกร่งในระบบของคุณ

หากระบบที่สำคัญไม่เป็นไปตามมาตรฐานความปลอดภัย คุณก็พร้อมที่จะเปลี่ยนไปใช้โซลูชันสำรองสำหรับระบบที่มีความสำคัญต่อพันธกิจ คุณใช้ทรัพยากรเช่น คำแนะนำเกี่ยวกับซัพพลายเชน (Supply Chain Guidance) ของ NCSC และกรอบงาน เช่น ระดับซัพพลายเชนสำหรับสิ่งประดิษฐ์ซอฟต์แวร์ (Supply Chain Levels for Software Artifacts - SLSA)[10] สำหรับการติดตามการรับรองของซัพพลายเชนและวงจรชีวิตของการพัฒนาซอฟต์แวร์

ระบุ ติดตาม และปกป้องทรัพย์สินของคุณ

คุณเข้าใจถึงคุณค่าขององค์กรของคุณในส่วนของสินทรัพย์ที่เกี่ยวข้องกับ AI รวมถึงโมเดล ข้อมูล (ที่รวมคำติชมของผู้ใช้) ข้อความแจ้งเตือน ซอฟต์แวร์ เอกสาร บันทึกและการประเมิน (รวมถึงข้อมูลเกี่ยวกับสิ่งต่าง ๆ ที่อาจมีความเสี่ยงต่อความปลอดภัยหรือทำให้ระบบล้มเหลวได้) โดยตระหนักว่าจุดใดที่แสดงถึงการลงทุนครั้งใหญ่อย่างมีนัยสำคัญและการเข้าถึงจุดใดที่อาจเสี่ยงต่อการถูกโจมตี คุณจัดการบันทึกต่าง ๆ เสมือนข้อมูลที่ละเอียดอ่อน และปรับใช้มาตรการควบคุมเพื่อปกป้องบันทึกเหล่านั้นให้เป็นความลับ คงความสมบูรณ์ และสามารถเข้าถึงได้

คุณทราบว่าทรัพย์สินของคุณอยู่ที่ไหน และได้พิจารณาและยอมรับความเสี่ยงที่อาจเกิดขึ้นที่มาพร้อมกับสิ่งเหล่านั้นแล้ว คุณมีกระบวนการและเครื่องมือในการติดตาม ตรวจสอบยืนยันความถูกต้อง จัดการควบคุมเวอร์ชันต่าง ๆ และรักษาให้ทรัพย์สินของคุณปลอดภัยและสามารถคืนค่าย้อนกลับไปสู่ สถานะที่ทุกอย่างปลอดภัย หากถูกโจมตีขึ้นมา

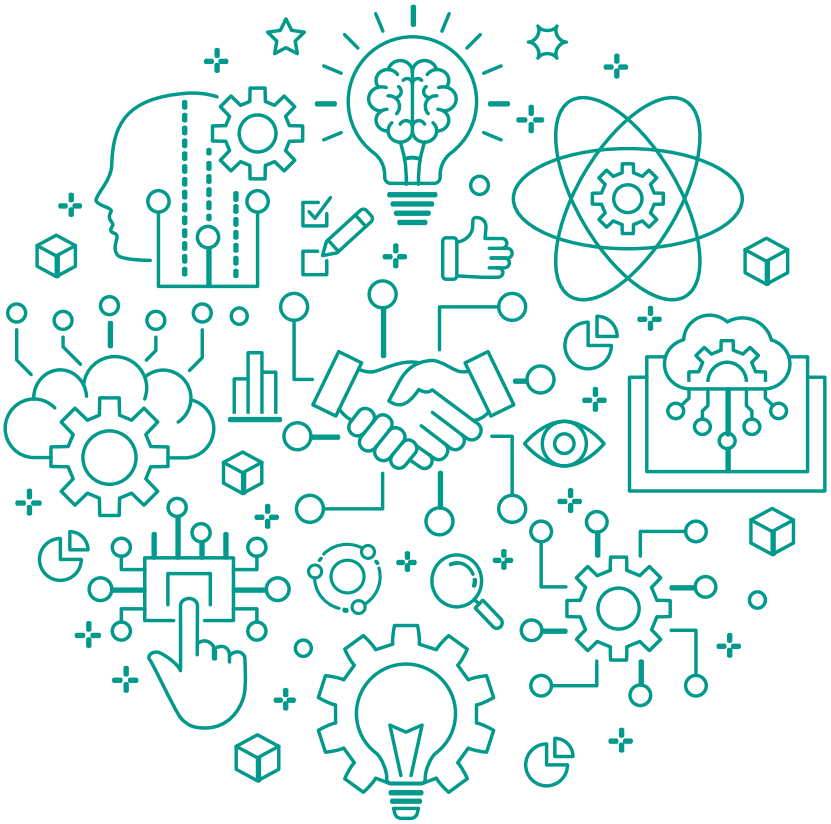
คุณมีกระบวนการและมาตรการควบคุมเพื่อจัดการข้อมูลที่ระบบ AI สามารถเข้าถึงได้ และเพื่อจัดการเนื้อหาที่สร้างโดย AI ตามความละเอียดอ่อนของระบบ (และความละเอียดอ่อนของอินพุตที่ใช้ในการสร้างระบบ)

จัดทำเอกสารสำหรับข้อมูล โมเดล และข้อความแจ้งเตือนของคุณ

คุณจัดทำเอกสารเก็บบันทึกการสร้าง การดำเนินการใช้งาน และการจัดการวงจรชีวิตของโมเดล ชุดข้อมูล และข้อความแจ้งเตือนเมตา หรือข้อความแจ้งเตือนระบบ (meta-or system-prompts) เอกสารของคุณประกอบด้วยข้อมูลที่เกี่ยวข้องกับความปลอดภัย เช่น แหล่งที่มาของข้อมูลการฝึกระบบ (รวมถึงมูลที่ผ่านการปรับแต่งให้ละเอียด และผลตอบรับจากมนุษย์หรือการปฏิบัติงานอื่น ๆ) ขอบเขตของสิ่งที่ระบบควรทำและจำกัดไม่ให้ทำ มาตรการด้านความปลอดภัย แฮชหรือลายเซ็นที่เข้ารหัสลับ ระยะเวลาที่ดำเนินการเก็บรักษาข้อมูล ความถี่ที่คุณควรตรวจสอบสิ่งต่าง ๆ และโหมดความล้มเหลวที่อาจเกิดขึ้น เครื่องมือที่เป็นประโยชน์เพื่อช่วยให้บรรลุเป้าหมายนี้ ได้แก่การ์ดโมเดล การ์ดข้อมูล และรายการวัสดุซอฟต์แวร์ (SBOM) การผลิตเอกสารที่ละเอียดครบถ้วนนั้น จะช่วยให้เกิดความโปร่งใสและช่วยให้เกิดความรับผิดชอบต่อสิ่งที่เกิดขึ้น[11]

จัดการกับหนี้ทางเทคนิคของคุณ

เช่นเดียวกับระบบซอฟต์แวร์อื่น ๆ คุณสามารถระบุ ติดตาม และจัดการ 'หนี้ทางเทคนิค' ของคุณตลอดวงจรชีวิตของระบบ AI ได้ (หนี้ทางเทคนิค คือ จุดที่การตัดสินใจทางวิศวกรรมไม่เป็นไปตามแนวปฏิบัติที่ดีที่สุด ซึ่งหมายถึงการจัดการกับวิธีแก้ปัญหาด่วนที่อาจเลือกไว้เพื่อผลประโยชน์ในระยะสั้น แต่อาจทำให้เกิดปัญหาในระยะยาวได้) หนี้ทางเทคนิคก็เหมือนกับหนี้ทางการเงิน ซึ่งไม่ใช่สิ่งที่เลวร้ายเสมอไป แต่ควรได้รับการจัดการให้เร็วที่สุดตั้งแต่เริ่มต้นกระบวนการพัฒนา[12] คุณรับรู้ว่าการทำเช่นนี้อาจมีความท้าทายในบริบทของ AI มากกว่าซอฟต์แวร์มาตรฐานทั่วไป และระดับหนี้ทางเทคนิคของคุณมีแนวโน้มที่จะสูงกว่า เนื่องจากวงจรการพัฒนาที่รวดเร็วและอาจยังไม่มีโปรโตคอลและอินเทอร์เฟซที่ได้รับการพัฒนาและยอมรับอย่างชัดเจน คุณต้องแน่ใจว่าแผนของคุณสำหรับวงจรชีวิตทั้งหมดในระบบ (รวมถึงกระบวนการในการเลิกใช้งานระบบ AI) ได้พิจารณา รับรู้ และลดความเสี่ยงต่อระบบที่คล้ายกันในอนาคต



3. การปรับใช้ที่ปลอดภัย

ในส่วนนี้จะกล่าวถึงแนวทางที่ใช้กับขั้นตอน**การพัฒนา**ของวงจรชีวิตการพัฒนาระบบ AI รวมถึงการปกป้องโครงสร้างพื้นฐานและโมเดลจากการถูกโจมตี ภัยคุกคามหรือการสูญเสีย การพัฒนากระบวนการจัดการเหตุการณ์ และการทำให้แน่ใจว่าทุกอย่างจะปลอดภัยเมื่อเปิดการใช้งานระบบ

ทำให้โครงสร้างพื้นฐานของคุณปลอดภัย

คุณปฏิบัติตามหลักการที่ดีในการทำให้โครงสร้างพื้นฐานของคุณปลอดภัย โดยใช้กับทุกส่วนของโครงสร้างพื้นฐานในวงจรชีวิตของระบบของคุณ คุณใช้การตั้งค่าการอนุญาตที่เหมาะสมสำหรับผู้ที่สามารถเข้าถึง API ต่าง ๆ โมเดล และข้อมูลของคุณ และขั้นตอนในไปป์ไลน์ที่เกี่ยวข้องกับการฝึกระบบและการประมวลผลสิ่งเหล่านั้น ในระหว่างขั้นตอนการวิจัยและการพัฒนาตลอดจนการปรับใช้งานระบบของคุณ ซึ่งรวมถึงการแยกรหัสหรือข้อมูลที่ละเอียดอ่อนออกจากกันในสภาพแวดล้อมที่แตกต่างกัน การทำเช่นนี้ยังจะช่วยลดการโจมตีความมั่นคงปลอดภัยทางไซเบอร์มาตรฐานทั่วไป ซึ่งมีจุดมุ่งหมายที่จะขโมย หรือทำลายประสิทธิภาพของโมเดล

ปกป้องโมเดลของคุณอย่างต่อเนื่อง

ผู้โจมตีอาจสามารถสร้างฟังก์ชันการทำงานของโมเดล[13] หรือข้อมูลที่ได้รับการฝึกระบบมา[14] ขึ้นมาใหม่ โดยการเข้าถึงโมเดลโดยตรง (รับรายละเอียดจากค่าน้ำหนักของโมเดล) หรือโดยทางอ้อม (โดยการสอบถามกับโมเดลผ่านแอปพลิเคชันหรือบริการที่ใช้โมเดลนั้น) ผู้โจมตียังอาจทำการรบกวนโมเดล ข้อมูล หรือข้อความแจ้งเตือนในระหว่างหรือหรือหลังการฝึกระบบแล้ว ทำให้ผลลัพธ์เอาต์พุตไม่น่าเชื่อถือ

คุณสามารถปกป้องโมเดลและข้อมูลจากการเข้าถึงทั้งทางตรงและทางอ้อมตามลำดับได้โดย

- ปรับใช้แนวปฏิบัติที่ดีที่สุดในด้านความมั่นคงปลอดภัยทางไซเบอร์ตามมาตรฐาน
- ปรับใช้มาตรการควบคุมบนอินเทอร์เฟซการสืบค้นหรือวิธีที่ผู้คนถามคำถามไปยังระบบ เพื่อตรวจจับและป้องกันความพยายามในการเข้าถึงเพื่อรับข้อมูล แก้ไขเปลี่ยนแปลงข้อมูล และขโมยข้อมูลที่เป็นความลับ

เพื่อให้แน่ใจว่าระบบอื่น ๆ สามารถยืนยันความถูกต้องของโมเดลของคุณ คุณจะต้องคำนวณและแชร์แฮชหรือลายนิ้วมือที่ใช้เข้ารหัสลับและ/หรือลายเซ็นของไฟล์โมเดล (เช่น ค่าน้ำหนักของโมเดล) และชุดข้อมูล (รวมถึงจุดตรวจสอบ) ทันทีที่โมเดลได้รับการฝึกระบบ การจัดการกุญแจการเข้าถึงระบบอย่างเหมาะสมถือเป็นเรื่องสำคัญ[15] สำหรับวิทยาการเข้ารหัสลับ (Cryptography) เสมอ

แนวปฏิบัติของคุณในการลดความเสี่ยงด้านการรักษาความลับจะขึ้นอยู่กับสิ่งที่คุณกำลังทำอยู่และโมเดลภัยคุกคามประเภทใดที่คุณพยายามป้องกัน แอปพลิเคชันบางตัว เช่น แอปพลิเคชันที่เกี่ยวข้องกับข้อมูลที่ละเอียดอ่อนอย่างยิ่ง อาจต้องมีการรับรองที่แข็งแกร่งตามทฤษฎี ซึ่งอาจทำได้ยากหรือมีค่าใช้จ่ายสูง หากเหมาะสมกับเหตุผล คุณสามารถใช้เทคโนโลยีเพิ่มความเป็นส่วนตัว (เช่น ความเป็นส่วนตัวที่แตกต่างกันหรือการเข้ารหัสแบบโฮโมมอร์ฟิก) เพื่อสำรวจหรือรับรองระดับความเสี่ยงที่เกี่ยวกับผู้บริโภคผู้ใช้ และผู้โจมตีที่สามารถเข้าถึงโมเดลและผลลัพธ์เอาต์พุตของคุณได้

พัฒนาขั้นตอนของการจัดการเหตุการณ์

เหตุการณ์ด้านความปลอดภัยที่หลีกเลี่ยงไม่ได้ส่งผลกระทบต่อระบบ AI ของคุณ แต่คุณมีแผนที่จะตอบสนอง ยกระดับ และแก้ไขสิ่งต่าง ๆ เมื่อมีเหตุการณ์เกิดขึ้น แผนของคุณคำนึงถึงฉากทัศน์ต่าง ๆ และได้รับการประเมินอย่างสม่ำเสมอ เนื่องจากทั้งระบบและการวิจัยในวงกว้างพัฒนาไปตามกาลเวลา คุณจัดเก็บทรัพยากรข้อมูลดิจิทัลที่สำคัญของบริษัทไว้ในการสำรองข้อมูลออฟไลน์หรือที่ไม่ได้เชื่อมต่อกับอินเทอร์เน็ต ผู้ที่รับผิดชอบในการจัดการปัญหาได้รับการฝึกอบรมเพื่อให้รู้จักประเมินและจัดการปัญหาต่อเหตุการณ์ที่เกี่ยวข้องกับ AI คุณมอบบันทึกการตรวจสอบที่ดีมีคุณภาพและมีคุณลักษณะหรือข้อมูลด้านความปลอดภัยอื่น ๆ ให้กับลูกค้าและผู้ใช้โดยไม่มีค่าใช้จ่ายเพิ่มเติม เพื่อให้พวกเขาสามารถรับมือกับปัญหาเหตุการณ์ใด ๆ ได้อย่างมีประสิทธิภาพ

เปิดตัวการใช้งาน AI อย่างมีความรับผิดชอบ

คุณเปิดตัวการใช้งานโมเดล แอปพลิเคชัน หรือระบบหลังจากที่ได้ประเมินความปลอดภัยที่เหมาะสมและมีประสิทธิภาพแล้ว เช่น ใช้การเปรียบเทียบ (Benchmarking) และใช้ทีมจำลองเพื่อโจมตีระบบ (Red teaming) (รวมถึงการทดสอบอื่น ๆ ที่อยู่นอกขอบเขตสำหรับแนวทางเหล่านี้ เช่น ความปลอดภัยหรือความเป็นธรรม) และคุณมีความชัดเจนในผู้ใช้ของคุณเกี่ยวกับข้อจำกัดที่ทราบหรือโหมดความล้มเหลวที่อาจเกิดขึ้น รายละเอียดของไลบรารีการทดสอบความปลอดภัยของโอเพ่นซอร์สมีอยู่ภายใต้หัวข้ออ่านข้อมูลเพิ่มเติมที่ตอนท้ายเอกสารนี้

ทำให้ผู้ใช้ทำสิ่งที่ถูกต้องได้อย่างง่ายดาย

คุณรับรู้ว่าทางเลือกการตั้งค่าหรือการกำหนดตั้งค่าใหม่ทุกรายการจะต้องได้รับการประเมินร่วมกับผลประโยชน์ทางธุรกิจที่ได้รับและความเสี่ยงด้านความปลอดภัยใด ๆ ที่ทางเลือกการตั้งค่าอาจนำมาด้วย ตามหลักการแล้ว การตั้งค่าที่ปลอดภัยที่สุดควรถูกสร้างรวมไว้ในระบบเป็นทางเลือกเดียว เมื่อจำเป็นต้องกำหนดตั้งค่า ทางเลือกการตั้งค่าเริ่มต้นควรมีความปลอดภัยในวงกว้างจากภัยคุกคามทั่วไป (นั่นก็คือ ความปลอดภัยโดยการตั้งค่า) คุณใช้มาตรการควบคุมเพื่อป้องกันไม่ให้ระบบของคุณถูกใช้หรือถูกปรับใช้ไปในลักษณะที่เป็นอันตราย

คุณให้คำแนะนำแก่ผู้ใช้เกี่ยวกับวิธีใช้โมเดลหรือระบบของคุณอย่างเหมาะสม ซึ่งเกี่ยวข้องกับการชี้ให้เห็นข้อจำกัดและโหมดความล้มเหลวที่อาจเกิดขึ้นโดยไม่คาดคิด คุณระบุให้ผู้ใช้เข้าใจอย่างชัดเจนถึงแง่มุมความปลอดภัยด้านใดที่พวกเขาต้องรับผิดชอบ และมีความโปร่งใสเกี่ยวกับเหตุผล (และวิธีการ) ที่ข้อมูลของพวกเขาอาจถูกนำไปใช้ เข้าถึง หรือจัดเก็บ (เช่น ใช้สำหรับการฝึกระบบโมเดลใหม่ หรือตรวจสอบโดยพนักงานหรือพันธมิตรคู่ค้า)

4. การทำงานและการบำรุงรักษาที่ปลอดภัย

ในส่วนนี้จะกล่าวถึงแนวทางที่นำไปใช้กับขั้นตอน**การทำงานและการบำรุงรักษาที่ปลอดภัย**ของวงจรชีวิตการพัฒนาระบบ AI ซึ่งจะให้แนวทางในการดำเนินการที่เกี่ยวข้อง โดยเฉพาะอย่างยิ่งเมื่อระบบมีการปรับใช้งานแล้ว รวมถึงการบันทึกและการเฝ้าระวังติดตามสิ่งที่เกิดขึ้นในระบบ การจัดการการอัปเดต และการแบ่งปันข้อมูลสำคัญเกี่ยวกับระบบ

ตรวจสอบพฤติกรรมของระบบของคุณ

คุณวัดผลของเอาต์พุตและประสิทธิภาพของโมเดลและระบบของคุณ เพื่อให้สามารถเห็นการเปลี่ยนแปลงพฤติกรรมแบบฉับพลันและแบบค่อยเป็นค่อยไปที่อาจส่งผลกระทบต่อความปลอดภัยได้ คุณสามารถรับรู้และระบุการบุกรุกระบบที่ไม่ได้รับอนุญาตและการละเมิดที่อาจเกิดขึ้นได้ รวมถึงการเบี่ยงเบนของข้อมูลตามธรรมชาติและเวลาที่ผ่านไป

ตรวจสอบอินพุตของระบบของคุณ

เพื่อให้เป็นไปตามข้อกำหนดด้านความเป็นส่วนตัวและการปกป้องข้อมูล คุณควรตรวจสอบและบันทึกอินพุตที่เข้าสู่ระบบของคุณ (เช่น คำขออนุมาน การสอบถาม หรือข้อความแจ้งเตือน) เพื่อช่วยให้คุณปฏิบัติตามข้อกำหนดทางกฎหมาย ตรวจสอบปัญหา และแก้ไขในกรณีที่ถูกโจมตีหรือใช้งานในทางที่ผิด ซึ่งอาจรวมถึงการตรวจจับอินพุตที่ผิดปกติ และ/หรือเป็นอินพุตจากฝ่ายตรงข้ามหรือผู้โจมตีที่เป็นอันตราย รวมถึงอินพุตที่มีจุดมุ่งหมายเพื่อใช้ประโยชน์จากขั้นตอนการเตรียมข้อมูล (เช่น การครอบตัดและการปรับเปลี่ยนขนาดรูปภาพในลักษณะที่อาจก่อให้เกิดปัญหา)

ใช้แนวปฏิบัติตามหลักความปลอดภัยโดยการออกแบบในการอัปเดต

คุณรวมการอัปเดตโดยอัตโนมัติตามค่าเริ่มต้นในทุกผลิตภัณฑ์ และใช้ขั้นตอนการอัปเดตที่ปลอดภัยและเป็นระเบียบในการจัดส่งการอัปเดตเหล่านั้น กระบวนการอัปเดตของคุณ (รวมถึงระบบการทดสอบและการประเมินผล) สะท้อนถึงความจริงที่ว่าการเปลี่ยนแปลงข้อมูล โมเดล หรือข้อความแจ้งเตือนสามารถนำไปสู่การเปลี่ยนแปลงพฤติกรรมของระบบได้ (เช่น คุณปฏิบัติต่อการอัปเดตที่สำคัญราวกับเป็นเวอร์ชันใหม่ทั้งหมด) คุณช่วยสนับสนุนให้ผู้ใช้ประเมินและตอบสนองต่อการเปลี่ยนแปลงโมเดล (เช่น โดยการเสนอการเข้าถึงเพื่อดูการเปลี่ยนแปลง และใช้ API เวอร์ชันเพื่อจัดการการปรับเปลี่ยนเหล่านั้น)

รวบรวมและแบ่งปันบทเรียนที่ได้รับ

คุณมีส่วนร่วมในชุมชนที่มีการแบ่งปันข้อมูล ทำงานร่วมกับผู้คนในระบบนิเวศของอุตสาหกรรม สถาบันการศึกษา และรัฐบาลทั่วโลกเพื่อแลกเปลี่ยนแนวปฏิบัติที่ดีที่สุดในการทำสิ่งต่าง ๆ ตามความเหมาะสม คุณรักษาช่องทางการสื่อสารแบบเปิดไว้ เพื่อรับข้อเสนอแนะเกี่ยวกับความปลอดภัยของระบบ ทั้งจากภายในและภายนอกองค์กรของคุณ รวมถึงการให้ความยินยอมแก่นักวิจัยด้านความปลอดภัยที่วิจัยและรายงานช่องโหว่ที่พวกเขาพบ เมื่อถึงเวลาจำเป็น คุณจะยกระดับปัญหาไปยังชุมชนในวงกว้าง เช่น การเผยแพร่กระดานข่าวที่มีการตอบกลับและอธิบายช่องโหว่อย่างเปิดเผย รวมถึงการแจงนับช่องโหว่ทั่วไปอย่างละเอียดและครบถ้วน คุณดำเนินการเพื่อบรรเทาและแก้ไขปัญหาอย่างรวดเร็วและเหมาะสม

อ่านข้อมูลเพิ่มเติม

การพัฒนา AI

หลักการรักษาระบบการเรียนรู้ของเครื่องให้ปลอดภัย
NCSC คำแนะนำโดยละเอียดเกี่ยวกับการพัฒนา ปรับใช้ หรือใช้งานระบบที่มีส่วนประกอบของ ML ของ NCSC

ความปลอดภัยโดยการออกแบบ - การปรับสมดุลของความเสี่ยงด้านความมั่นคงปลอดภัยทางไซเบอร์ (Secure by Design - Shifting the Balance of Cybersecurity Risk) หลักการและแนวปฏิบัติของซอฟต์แวร์ที่สร้างขึ้นตามหลักความปลอดภัยโดยการออกแบบ (Principles and Approaches for Secure by Design Software)
ซึ่งร่วมแต่งขึ้นโดย CISA และ NCSC อีกทั้งคณะกรรมาธิการอื่น ๆ แนวทางในเอกสารนี้อธิบายว่าผู้ผลิตระบบซอฟต์แวร์และ AI ควรสร้างความปลอดภัยให้กับผลิตภัณฑ์ของตนตั้งแต่ขั้นตอนการออกแบบในกระบวนการพัฒนา และส่งสินค้าที่มีความปลอดภัยถึงมือลูกค้าตั้งแต่การแกะกล่องได้อย่างไร

สรุปข้อกังวลด้านความปลอดภัยของ AI (AI Security Concerns in a Nutshell)
ผลิตโดยสำนักงานรักษาความปลอดภัยข้อมูลแห่งสหพันธรัฐเยอรมนี (BSI) เอกสารนี้ให้ภาพรวมของวิธีที่ระบบการเรียนรู้ของเครื่องสามารถถูกโจมตีได้ และแนะนำวิธีป้องกันจากการโจมตีเหล่านั้น

หลักการชี้แนะระหว่างประเทศของกระบวนการฮิโรชิม่าสำหรับองค์กรที่พัฒนาระบบ AI ขั้นสูง (Hiroshima Process International Guiding Principles for Organizations Developing Advanced AI Systems) และ หลักจรรยาบรรณระหว่างประเทศของกระบวนการฮิโรชิม่าสำหรับองค์กรที่พัฒนาระบบ AI ขั้นสูง (Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems)
เอกสารเหล่านี้จัดทำขึ้นโดยเป็นส่วนหนึ่งของกระบวนการ G7 Hiroshima AI Process ซึ่งให้คำแนะนำแก่องค์กรที่พัฒนาระบบ AI ที่ก้าวหน้าที่สุด รวมถึงโมเดลพื้นฐานที่ทันสมัยที่สุดและระบบ Generative AI ที่สร้างโดยมีเป้าหมายเพื่อส่งเสริมระบบ AI ที่ปลอดภัย มั่นคง และเชื่อถือได้ทั่วโลก

AI Verify
เป็นกรอบการทดสอบการกำกับดูแล AI และชุดเครื่องมือซอฟต์แวร์ของสิงคโปร์ (AI Governance Testing Framework and Software) ที่ตรวจสอบประสิทธิภาพของระบบ AI โดยเทียบกับชุดหลักการที่ได้รับการยอมรับในระดับสากลผ่านการทดสอบที่ได้มาตรฐาน

กรอบงานหลากมิติสำหรับแนวปฏิบัติด้านความมั่นคงปลอดภัยทางไซเบอร์ที่ดีสำหรับ AI (Multilayer Framework for Good Cybersecurity Practices for AI — ENISA) (europa.eu)
เป็นกรอบการทำงานเพื่อเป็นแนวทางให้แก่หน่วยงานผู้มีอำนาจระดับชาติและผู้มีส่วนได้ส่วนเสียด้าน AI ในขั้นตอนที่พวกเขาจำเป็นต้องปฏิบัติตามเพื่อรักษาความปลอดภัยของระบบ การดำเนินงาน และกระบวนการของ AI

ISO 5338: กระบวนการวงจรชีวิตของระบบ AI (อยู่ในระหว่างการทบทวน)
เป็นชุดของกระบวนการและแนวคิดที่เกี่ยวข้องสำหรับใช้อธิบายถึงวงจรชีวิตของระบบ AI ตามการเรียนรู้ของเครื่องและระบบการศึกษาสำนึก

แค็ตตาล็อกเกณฑ์การปฏิบัติตามข้อกำหนดของบริการ AI Cloud (AI Cloud Service Compliance Criteria Catalogue - AIC4)
แค็ตตาล็อกของ BSI ฉบับนี้ให้เกณฑ์เฉพาะของ AI ซึ่งช่วยให้สามารถประเมินความปลอดภัยของบริการ AI ตลอดวงจรชีวิตได้

อนุกรมวิธานและศัพทวิทยาของการเรียนรู้ของเครื่องที่ขัดแย้งกันโดย NIST IR 8269 (ฉบับร่าง) (NIST IR 8269 (Draft) A Taxonomy and Terminology of Adversarial Machine Learning)
เป็นชุดของกระบวนการและแนวคิดที่เกี่ยวข้องสำหรับการอธิบายวงจรชีวิตของระบบ AI ตามการเรียนรู้ของเครื่องและระบบการศึกษาสำนึก (Heuristic system)

MITRE ATLAS
เป็นฐานความรู้เกี่ยวกับกลยุทธ์ เทคนิค และกรณีศึกษาของฝ่ายที่ขัดแย้งกันสำหรับระบบการเรียนรู้ของเครื่อง (ML) ซึ่งมีโมเดลตามแบบและเชื่อมโยงกับกรอบงาน MITRE ATT&CK

ภาพรวมความเสี่ยง AI ระดับหายนะปี 2023 (An Overview of Catastrophic AI Risks - 2023)
ผลิตโดยศูนย์ความปลอดภัย AI (Center for AI Safety) เอกสารนี้ระบุขอบเขตความเสี่ยงที่เกิดจาก AI

โมเดลภาษาขนาดใหญ่: โอกาสและความเสี่ยงสำหรับอุตสาหกรรมและหน่วยงาน (Large Language Models: Opportunities and Risks for Industry and Authorities)
เอกสารที่จัดทำโดย BSI สำหรับบริษัท หน่วยงาน และนักพัฒนาที่ต้องการเรียนรู้เพิ่มเติมเกี่ยวกับโอกาสและความเสี่ยงในการพัฒนา การปรับใช้ และ/หรือการใช้โมเดลภาษาขนาดใหญ่หรือ LLM

โครงการโอเพ่นซอร์สเพื่อช่วยผู้ใช้ในการทดสอบความปลอดภัยของโมเดล AI ได้แก่

- กล่องเครื่องมือ Adversarial Robustness Toolbox (IBM)
- CleverHans โดยมหาวิทยาลัยโตรอนโต (University of Toronto)
- TextAttack โดยมหาวิทยาลัยเวอร์จิเนีย (University of Virginia)
- Prompt Bench โดยไมโครซอฟ (Microsoft)
- Counterfit โดยไมโครซอฟ (Microsoft)
- AI Verify หน่วยงานพัฒนาสื่อ Infocomm ประเทศสิงคโปร์ (Infocomm Media Development Authority, Singapore)

ความมั่นคงปลอดภัยทางไซเบอร์

เป้าหมายประสิทธิภาพความมั่นคงปลอดภัยทางไซเบอร์ของ CISA (CISA's Cybersecurity Performance Goals)
เป็นชุดการป้องกันทั่วไปที่หน่วยงานโครงสร้างพื้นฐานที่สำคัญทั้งหมดควรใช้ เพื่อลดโอกาสและผลกระทบของความเสี่ยงที่ทราบและเทคนิคของฝ่ายที่ขัดแย้งกันอย่างมีความหมาย

กรอบงาน NCSC CAF Framework
กรอบงานของการประเมินทางไซเบอร์ (The Cyber Assessment Framework - CAF) ให้คำแนะนำสำหรับองค์กรที่รับผิดชอบด้านบริการและกิจกรรมที่สำคัญอย่างยิ่ง

กรอบงานด้านความปลอดภัยในซัพพลายเชนของ MITRE (MITRE's Supply Chain Security Framework)
เป็นกรอบงานของการประเมินซัพพลายเออร์และผู้ให้บริการภายในซัพพลายเชน

การจัดการความเสี่ยง

กรอบการจัดการความเสี่ยง AI โดย NIST (NIST AI Risk Management Framework - AI RMF)
เอกสาร AI RMF นี้ สรุปวิธีจัดการความเสี่ยงทางสังคมและเทคนิคสำหรับบุคคล องค์กร และสังคมที่เกี่ยวข้องกับ AI โดยเฉพาะ

ISO 27001: ความปลอดภัยของข้อมูล ความมั่นคงปลอดภัยทางไซเบอร์ และการปกป้องความเป็นส่วนตัว
มาตรฐานนี้ให้แนวทางแก่องค์กรเกี่ยวกับการจัดตั้ง การปรับใช้ และการบำรุงรักษาระบบการจัดการความปลอดภัยของข้อมูล

ISO 31000: การจัดการความเสี่ยง
เป็นมาตรฐานสากลที่ให้แนวทางและหลักการแก่องค์กรในการบริหารความเสี่ยงภายในองค์กร

แนวทางบริหารความเสี่ยงโดย NCSC (NCSC Risk Management Guidance)
แนวทางนี้ ช่วยให้ผู้ปฏิบัติงานที่มีความเสี่ยงด้านความมั่นคงปลอดภัยทางไซเบอร์เข้าใจและจัดการความเสี่ยงด้านความมั่นคงปลอดภัยทางไซเบอร์ที่ส่งผลกระทบต่อองค์กรได้ดียิ่งขึ้น

หมายเหตุ

1. ในที่นี้หมายถึงบุคคล หน่วยงานสาธารณะ หน่วยงาน หรือองค์กรอื่น ๆ ที่พัฒนาระบบ AI (หรือที่ได้จัดให้มีการพัฒนาระบบ AI ขึ้น) และนำระบบนั้นออกสู่ตลาดหรือให้บริการภายใต้ชื่อหรือเครื่องหมายการค้าของตนเอง

2. สำหรับข้อมูลเพิ่มเติมเกี่ยวกับความปลอดภัยโดยการออกแบบ ดูเว็บเพจและแนวทางเรื่องความปลอดภัยโดยการออกแบบของ CISA การปรับสมดุลของความเสี่ยงด้านความมั่นคงปลอดภัยทางไซเบอร์: หลักการและแนวปฏิบัติสำหรับซอฟต์แวร์ที่สร้างตามหลักความปลอดภัยโดยการออกแบบ

3. ตรงข้ามกับแนวปฏิบัติ AI ที่ไม่ใช่ ML (non-ML) อย่างเช่นระบบที่ใช้กฎและคำสั่งที่กำหนดไว้ล่วงหน้า (rule-based)

4. CEPS อธิบายปฏิสัมพันธ์การพัฒนา AI ประเภทต่าง ๆ เจ็ดประเภทในสิ่งพิมพ์ของพวกเขาที่ชื่อว่า 'Reconciling the AI Value Chain with the EU's Artificial Intelligence Act'

5. ISO/IEC 22989:2022(en) ให้นิยามสิ่งนี้ว่าเป็น 'องค์ประกอบการทำงานหนึ่งที่สร้างระบบ AI ใด ๆ'

6. NIST ได้รับมอบหมายให้จัดทำแนวทาง (และดำเนินมาตรการอื่น ๆ) เพื่อส่งเสริมความก้าวหน้าของการพัฒนาและการใช้งานปัญญาประดิษฐ์ (AI) ที่ปลอดภัย มั่นคง และเชื่อถือได้ ดูความรับผิดชอบของ NIST ภายใต้คำสั่งผู้บริหาร 30 ตุลาคม 2023

7. ข้อมูลเพิ่มเติมเกี่ยวกับการสร้างโมเดลภัยคุกคามสามารถดูได้ที่ มูลนิธิ OWASP Foundation

8. ดู MITRE ATLAS การเรียนรู้ของเครื่องที่ขัดแย้งกัน 101 (Adversarial Machine Learning 101)

9. GitHub: RCE PoC สำหรับ Tensorflow โดยใช้เลเยอร์ Lambda ที่เป็นอันตราย (RCE PoC for Tensorflow using a malicious Lambda layer)

10. SLSA: การปกป้องความสมบูรณ์ของสิ่งประดิษฐ์ในซัพพลายเชนซอฟต์แวร์ใด ๆ (Safeguarding artifact integrity across any software supply chain)

11. METI (กระทรวงเศรษฐกิจ การค้า และอุตสาหกรรมของญี่ปุ่น ปี 2023) ได้เผยแพร่แนวทางที่ชื่อว่า 'Guide of Introduction of Software Bill of Materials (SBOM) for Software Management'

12. การวิจัยจาก Google: การเรียนรู้ของเครื่อง (Machine Learning): บัตรเครดิตดอกเบี้ยสูงสำหรับหนี้ทางเทคนิค (The High Interest Credit Card of Technical Debt)

13. Tramèr และคณะ ปี 2016 การขโมยโมเดลการเรียนรู้ของเครื่องผ่าน Prediction API (Stealing Machine Learning Models via Prediction APIs)

14. Boenisch ปี 2020 การโจมตีความเป็นส่วนตัวของการเรียนรู้ของเครื่อง (ตอนที่ 1) (Attacks against Machine Learning Privacy - Part 1): โมเดลการโจมตีแบบผกผันด้วย IBM-ART Framework (Model Inversion Attacks with the IBM-ART Framework)

15. ศูนย์รักษาความมั่นคงปลอดภัยทางไซเบอร์แห่งชาติ (NCSC) ปี 2020 ออกแบบและสร้าง Public Key Infrastructure ที่โฮสต์แบบส่วนตัว (Design and build a privately hosted Public Key Infrastructure)