AI Security Guidelines

แนวปฏิบัติการใช้ปัญญาประดิษฐ์อย่างมั่นคงปลอดภัย

สำนักงานคณะกรรมการการรักษาความมั่นคงปลอดภัยไซเบอร์แห่งชาติ

คำนำ

ปัญญาประดิษฐ์ (Artificial Intelligence: AI) เป็นเทคโนโลยีที่เปี่ยมด้วยศักยภาพและมีบทบาทเชิงยุทธศาสตร์ต่อการพัฒนาเศรษฐกิจและสังคมของประเทศไทย ทั้งนี้เทคโนโลยีปัญญาประดิษฐ์มาพร้อมกับความท้าทายและความเสี่ยงที่ซับซ้อน อาทิ ภัยคุกคามทางไซเบอร์ ซึ่งรวมถึงการโจมตีทางไซเบอร์แบบดั้งเดิมและการโจมตีทางไซเบอร์ที่เฉพาะเจาะจงต่อระบบปัญญาประดิษฐ์ การรั่วไหลของข้อมูลส่วนบุคคล และการบิดเบือนเชิงระบบ ซึ่งอาจส่งผลกระทบอย่างร้ายแรงต่อการพัฒนาและความมั่นคงของประเทศ

เพื่อรับมือกับความท้าทายดังกล่าว สำนักงานคณะกรรมการการรักษาความมั่นคงปลอดภัยไซเบอร์แห่งชาติ (สกมช.) โดยศูนย์วิจัยความมั่นคงปลอดภัยไซเบอร์จึงได้จัดทำแนวปฏิบัติการใช้ปัญญาประดิษฐ์อย่างมั่นคงปลอดภัย (AI Security Guidelines) โดยได้ศึกษา รวบรวมข้อมูล และวิเคราะห์จากเอกสารมาตรฐานและรายงานเชิงเทคนิคระดับสากล เช่น ISO/IEC ENISA และ OWASP รวมถึงกฎระเบียบและแนวปฏิบัติภายในประเทศ เพื่อจัดทำเป็นข้อมูลให้ผู้เกี่ยวข้องกับระบบนิเวศปัญญาประดิษฐ์ ตั้งแต่ผู้พัฒนาจนถึงผู้ใช้ อาทิเช่น ผู้บริหารระดับสูงขององค์กร ทีมพัฒนาและดำเนินงานปัญญาประดิษฐ์ นักกฎหมายและเจ้าหน้าที่คุ้มครองข้อมูลส่วนบุคคล เจ้าหน้าที่ฝ่ายความมั่นคงปลอดภัยไซเบอร์ พนักงานผู้ใช้งาน ลูกค้า เจ้าของข้อมูล ผู้ขาย ตลอดจน หน่วยงานกำกับดูแล หน่วยงานรับรองและตรวจสอบ ประชาสังคม และสาธารณชน ได้ใช้เป็นแนวทางในการพัฒนา บริหารจัดการ และใช้งานระบบปัญญาประดิษฐ์อย่างมั่นคงปลอดภัย สามารถรับมือกับความท้าทายที่เกิดขึ้นจากปัญญาประดิษฐ์ทั้งในปัจจุบันและอนาคต ส่งเสริมให้ประเทศไทยมีศักยภาพในการพัฒนาและประยุกต์ใช้เทคโนโลยีปัญญาประดิษฐ์ได้อย่างปลอดภัยและน่าเชื่อถือ อันจะนำไปสู่การขับเคลื่อนประเทศสู่อนาคตดิจิทัลที่มั่นคงและยั่งยืนในที่สุด

สำนักงานคณะกรรมการการรักษาความมั่นคงปลอดภัยไซเบอร์แห่งชาติ

กิตติกรรมประกาศ

สำนักงานคณะกรรมการการรักษาความมั่นคงปลอดภัยไซเบอร์แห่งชาติ ขอขอบคุณผู้ทรงคุณวุฒิทั้งที่เอ่ยนามและมิได้เอ่ยนามในที่นี้ ที่ได้สละเวลา ความรู้ และความสามารถ รวมทั้งอนุเคราะห์ข้อมูลและข้อเสนอแนะที่เป็นประโยชน์ต่อการจัดทำ “แนวปฏิบัติการใช้ปัญญาประดิษฐ์อย่างมั่นคงปลอดภัย (AI Security Guidelines)” ฉบับนี้จนสำเร็จ

บุคลากรที่มีส่วนร่วมในการจัดทำ “แนวปฏิบัติการใช้ปัญญาประดิษฐ์อย่างมั่นคงปลอดภัย (AI Security Guidelines)” ฉบับนี้ มาจากหลายภาคส่วน ทั้งผู้ทรงคุณวุฒิจากภาครัฐ ภาคเอกชน ภาควิชาการ ภาคประชาสังคม และอีกหลายท่านที่มิได้เอ่ยนาม ณ ที่นี้

สำนักงานคณะกรรมการการรักษาความมั่นคงปลอดภัยไซเบอร์แห่งชาติ ขอขอบคุณท่านที่ปรึกษาคณะทำงานเพื่อจัดทำแนวทางปฏิบัติการใช้ปัญญาประดิษฐ์อย่างมั่นคงปลอดภัย ดังต่อไปนี้

ดร.ศักดิ์ เสกขุนทด	ที่ปรึกษาคณะทำงาน
ดร.พีรเดช ณ น่าน	ที่ปรึกษาคณะทำงาน
ดร.มนต์ศักดิ์ โซ่เจริญธรรม	ที่ปรึกษาคณะทำงาน
ดร.ชาลี วรกุลพิพัฒน์	ที่ปรึกษาคณะทำงาน
นายวิโรจน์ อัศวรังสี	ที่ปรึกษาคณะทำงาน

ทุกท่านได้ใช้ความรู้ ความสามารถ และประสบการณ์ ในการให้ข้อมูลและข้อเสนอแนะอันเป็นประโยชน์ และอันทรงคุณค่า ทั้งในที่ประชุมและผ่านช่องทางออนไลน์ซึ่งเป็นส่วนสำคัญอย่างยิ่งในการพัฒนาแนวปฏิบัติฉบับนี้ให้สำเร็จลุล่วง

สำนักงานคณะกรรมการการรักษาความมั่นคงปลอดภัยไซเบอร์แห่งชาติ

คำแนะนำสำหรับการอ่านแนวปฏิบัติการใช้ปัญญาประดิษฐ์อย่างมั่นคงปลอดภัย

แนวปฏิบัติการใช้ปัญญาประดิษฐ์อย่างมั่นคงปลอดภัยฉบับนี้ ได้รวบรวมข้อมูลที่เป็นประโยชน์ต่อกลุ่มเป้าหมายที่หลากหลาย โดยมีเนื้อหาครอบคลุมหลากหลายมิติในการใช้งานปัญญาประดิษฐ์อย่างมั่นคงปลอดภัย ตั้งแต่เชิงนโยบายสำหรับผู้บริหารไปจนถึงเชิงเทคนิคสำหรับผู้ปฏิบัติการ ทั้งนี้สำหรับผู้อ่านที่ต้องการศึกษาข้อมูลเฉพาะที่เกี่ยวข้องกับบทบาทของตน สามารถเลือกอ่านเนื้อหาบางบทหรือบางหัวข้อได้ตามความเหมาะสม โดยมีรายละเอียดเนื้อหากับกลุ่มเป้าหมายที่เหมาะสม ดังแสดงในตารางที่ ๑

ตารางที่ ๑ แนวทางในการศึกษาข้อมูลในแต่ละบทบาท

กลุ่มเป้าหมาย	ความเกี่ยวข้อง	เนื้อหาที่เกี่ยวข้อง
๑ ภายในองค์กร		
๑.๑ ผู้บริหารระดับสูง	กำหนดทิศทางเชิงกลยุทธ์ มีความรับผิดชอบสูงสุดต่อความเสี่ยง และการบริหารจัดการจัดสรรทรัพยากร	บทที่ ๑ บทนำ บทที่ ๒ ข้อ ๒.๓ หมวดหมู่ความเสี่ยงของปัญญาประดิษฐ์ บทที่ ๔ การกำกับดูแลและการบริหารความเสี่ยงสำหรับระบบปัญญาประดิษฐ์
๑.๒ ทีมพัฒนาและดำเนินงานทางด้านปัญญาประดิษฐ์	เป็น "ผู้สร้าง" ที่ต้องนำหลักการไปปฏิบัติจริงในทางเทคนิคตลอดวงจรชีวิตของระบบปัญญาประดิษฐ์	บทที่ ๒ ข้อ ๒.๖ ภาพรวมภัยคุกคามด้านความมั่นคงปลอดภัยของปัญญาประดิษฐ์ บทที่ ๓ กรอบการรักษาความมั่นคงปลอดภัยระบบปัญญาประดิษฐ์ ภาคผนวก ข ค และ ง
๑.๓ ฝ่ายกฎหมายและเจ้าหน้าที่คุ้มครองข้อมูลส่วนบุคคล (Data Protection Officer: DPO)	ดูแลให้สอดคล้องกับกฎหมาย โดยเฉพาะพระราชบัญญัติคุ้มครองข้อมูลส่วนบุคคล และข้อบังคับต่าง ๆ	บทที่ ๒ ข้อ ๒.๓ หมวดหมู่ความเสี่ยงของปัญญาประดิษฐ์ บทที่ ๓ กรอบการรักษาความมั่นคงปลอดภัยระบบปัญญาประดิษฐ์ (เฉพาะระยะที่ ๐ และ ๖) บทที่ ๔ การกำกับดูแลและการบริหารความเสี่ยงสำหรับระบบปัญญาประดิษฐ์ (โดยเฉพาะ ๔.๑ และ ๔.๓)

กลุ่มเป้าหมาย	ความเกี่ยวข้อง	เนื้อหาที่เกี่ยวข้อง
๑.๔ ฝ่ายความมั่นคงปลอดภัยไซเบอร์	เป็น "ผู้ทวนสอบและผู้ป้องกัน" ที่ต้องเข้าใจภัยคุกคาม ทดสอบระบบ และเฝ้าระวัง	บทที่ ๒ ข้อ ๒.๖ ภาพรวมภัยคุกคามด้านความมั่นคงปลอดภัยฯ บทที่ ๓ กรอบการรักษาความมั่นคงปลอดภัยระบบปัญญาประดิษฐ์ (โดยเฉพาะระยะที่ ๑ ๓ ๔ และ ๕) ภาคผนวก ข ค ง และ จ
๑.๕ พนักงานผู้ใช้งาน	เป็นผู้ใช้ภายในที่ต้องปฏิบัติตามนโยบายและตระหนักถึงความเสี่ยง	บทที่ ๒ ข้อ ๒.๔ ประโยชน์ของปัญญาประดิษฐ์ บทที่ ๒ ข้อ ๒.๕ ความเสี่ยงจากการใช้ปัญญาประดิษฐ์ที่ไม่มั่นคงปลอดภัย ภาคผนวก ช ตัวอย่างนโยบายการใช้เทคโนโลยี Generative AI ที่ยอมรับได้ ภาคผนวก ซ ตัวอย่างนโยบายการใช้เทคโนโลยีระบบปัญญาประดิษฐ์ที่ยอมรับได้
๒. ภายนอกองค์กร		
๒.๑ ลูกค้าและผู้ใช้งานปลายทาง	เป็นผู้รับผลกระทบโดยตรงจากการทำงานของระบบปัญญาประดิษฐ์ ควรได้รับความโปร่งใสในการทำงานของระบบ	บทที่ ๑ บทนำ บทที่ ๒ ข้อ ๒.๔ ประโยชน์ของปัญญาประดิษฐ์ บทที่ ๒ ข้อ ๒.๕ ความเสี่ยงจากการใช้ปัญญาประดิษฐ์ที่ไม่มั่นคงปลอดภัย บทที่ ๔ ข้อ ๔.๔ การตรวจสอบและการรับรอง ภาคผนวก ช ตัวอย่างนโยบายการใช้เทคโนโลยี Generative AI ที่ยอมรับได้ ภาคผนวก ซ ตัวอย่างนโยบายการใช้เทคโนโลยีระบบปัญญาประดิษฐ์ที่ยอมรับได้

กลุ่มเป้าหมาย	ความเกี่ยวข้อง	เนื้อหาที่เกี่ยวข้อง
๒.๒ เจ้าของข้อมูล	ข้อมูลส่วนบุคคลถูกนำไปใช้ จึงต้องมั่นใจว่าได้รับการคุ้มครองตามสิทธิ์	บทที่ ๓ กรอบการรักษาความมั่นคงปลอดภัยของระบบปัญญาประดิษฐ์ (เฉพาะระยะที่ ๐ ๒ และ ๖) บทที่ ๔ ข้อ ๔.๓ การปฏิบัติตามกฎหมายและข้อบังคับของไทย
๒.๓ พันธมิตรและผู้ขายในห่วงโซ่อุปทาน	เป็นส่วนหนึ่งของห่วงโซ่อุปทานที่ต้องมีความมั่นคงปลอดภัย	บทที่ ๓ กรอบการรักษาความมั่นคงปลอดภัยของระบบปัญญาประดิษฐ์ (เฉพาะระยะที่ ๑ และ ๒)
๓. สังคมและหน่วยงานกำกับดูแล		
๓.๑ หน่วยงานกำกับดูแล	มุ่งเน้นการคุ้มครองสาธารณะและบังคับใช้กฎหมาย	บทที่ ๑ ข้อ ๑.๓ กลุ่มเป้าหมายและการนำไปใช้ บทที่ ๒ ข้อ ๒.๓ หมวดหมู่ความเสี่ยงของปัญญาประดิษฐ์ บทที่ ๔ ข้อ ๔.๓ การปฏิบัติตามกฎหมายและข้อบังคับของไทย
๓.๒ หน่วยงานรับรองและตรวจสอบ	ต้องการเกณฑ์การตรวจสอบที่เป็นระบบและสอดคล้องกับมาตรฐานสากล	บทที่ ๓ กรอบการรักษาความมั่นคงปลอดภัยระบบปัญญาประดิษฐ์ บทที่ ๔ ข้อ ๔.๔ การตรวจสอบและการรับรอง ภาคผนวก ข ตัวอย่างรายการตรวจสอบความมั่นคงปลอดภัยสำหรับระบบปัญญาประดิษฐ์
๓.๓ ประชาสังคมและสาธารณชน	ต้องการความเข้าใจถึงผลกระทบของระบบปัญญาประดิษฐ์ในภาพรวม	บทที่ ๑ บทนำ บทที่ ๒ ข้อ ๒.๔ ประโยชน์ของปัญญาประดิษฐ์ บทที่ ๒ ข้อ ๒.๕ ความเสี่ยงจากการใช้ปัญญาประดิษฐ์ที่ไม่มั่นคงปลอดภัย ภาคผนวก ช ตัวอย่างนโยบายการใช้เทคโนโลยี Generative AI ที่ยอมรับได้ ภาคผนวก ช ตัวอย่างนโยบายการใช้เทคโนโลยีระบบปัญญาประดิษฐ์ที่ยอมรับได้

สารบัญ

สารบัญ (ต่อ)

บทที่ ๑
บทนำ

๑.๑ ความสำคัญของความมั่นคงปลอดภัยสำหรับระบบปัญญาประดิษฐ์

ปัญญาประดิษฐ์ (Artificial Intelligence: AI) ได้กลายเป็นพลังขับเคลื่อนที่สำคัญในการปฏิรูปอุตสาหกรรมและสังคมทั่วโลก สำหรับประเทศไทย การนำมาประยุกต์ใช้ช่วยเพิ่มขีดความสามารถในการแข่งขันและยกระดับคุณภาพชีวิตได้อย่างมหาศาล อย่างไรก็ตาม การใช้งานอย่างแพร่หลายได้นำมาซึ่งความท้าทายและความเสี่ยงด้านความมั่นคงปลอดภัยไซเบอร์รูปแบบใหม่ โดยระบบปัญญาประดิษฐ์ไม่เพียงแต่เป็นเป้าหมายของการโจมตี แต่ยังถูกใช้เป็นเครื่องมือที่ซับซ้อนขึ้น ไม่ว่าจะเป็นการนำไปใช้ในทางที่ผิดเพื่อสร้างข่าวปลอม การทำ Deepfake เพื่อบิดเบือนความจริง หรือสร้างข้อความฟิชชิงที่แนบเนียนเพื่อหลอกลวงมนุษย์ ไปจนถึงการใช้เป็นเครื่องมือโจมตีทางเทคนิคเพื่อพัฒนามัลแวร์ที่สามารถหลบเลี่ยงการตรวจจับหรือใช้ค้นหาช่องโหว่ของระบบโดยอัตโนมัติ

ในบริบทสากลนานาประเทศได้ริเริ่มจัดตั้งองค์กรเฉพาะทางขึ้น เพื่อวิจัยและประเมินความปลอดภัยของปัญญาประดิษฐ์ ซึ่งนับเป็นปรากฏการณ์ที่สำคัญอย่างหนึ่งที่ทั่วโลกได้ให้ความสำคัญกับความมั่นคงปลอดภัยของระบบปัญญาประดิษฐ์อย่างจริงจัง แนวโน้มดังกล่าวสะท้อนให้เห็นถึงความจำเป็นเร่งด่วนที่ประเทศไทยต้องยกระดับความพร้อม เพื่อรับมือกับความท้าทายและความเสี่ยงที่ซับซ้อนจากเทคโนโลยีปัญญาประดิษฐ์

ประเทศไทยได้เห็นความสำคัญในการเตรียมความพร้อมในการใช้งานปัญญาประดิษฐ์อย่างเหมาะสมและสร้างสรรค์ หน่วยงานภาครัฐเริ่มมีการออกแนวปฏิบัติ คำแนะนำ และเอกสารที่เกี่ยวข้องกับการใช้งานปัญญาประดิษฐ์ เช่น ศูนย์ธรรมาภิบาลปัญญาประดิษฐ์ (AI Governance Center: AIGC) ภายใต้สำนักงานพัฒนาธุรกรรมทางอิเล็กทรอนิกส์ (สพธอ.) ได้จัดทำ “แนวทางการประยุกต์ใช้ปัญญาประดิษฐ์อย่างมีธรรมาภิบาล สำหรับผู้บริหารองค์กร” ซึ่งวางกรอบธรรมาภิบาลและหลักการเชิงนโยบายสำหรับการกำกับดูแลปัญญาประดิษฐ์ในระดับประเทศ ทั้งนี้ ประเทศไทยยังไม่มีการจัดทำเอกสารแนวปฏิบัติเชิงเทคนิคสำหรับการใช้งานปัญญาประดิษฐ์อย่างมั่นคงปลอดภัย ซึ่งจะเป็นรากฐานสำคัญในการสร้างความน่าเชื่อถือและความมั่นคงปลอดภัยของการใช้ปัญญาประดิษฐ์ ช่วยลดความเสี่ยงและผลกระทบจากภัยคุกคามทางไซเบอร์ที่เกี่ยวข้องกับปัญญาประดิษฐ์ได้

ด้วยเหตุนี้ แนวปฏิบัติฉบับนี้จึงถูกพัฒนาขึ้นบนรากฐานของมาตรฐานสากลและแนวปฏิบัติที่ดี โดยมีวัตถุประสงค์เพื่อเป็นเอกสารส่วนขยายที่มุ่งเน้นด้านการรักษาความมั่นคงปลอดภัยไซเบอร์สำหรับระบบปัญญาประดิษฐ์โดยเฉพาะ โดยมีการเชื่อมโยงหลักการเชิงนโยบายให้เป็นมาตรการควบคุมและกระบวนการที่สามารถนำไปปฏิบัติได้ตลอดวงจรชีวิตของระบบปัญญาประดิษฐ์ ครอบคลุมถึง

การป้องกัน ตรวจจับ และตอบสนองต่อภัยคุกคามทางไซเบอร์ที่มีลักษณะเฉพาะต่อเทคโนโลยีปัญญาประดิษฐ์ เพื่อเป็นเครื่องมือสำคัญในการเสริมสร้างศักยภาพขององค์กรในประเทศ ทั้งนี้ เพื่อส่งเสริมการพัฒนานวัตกรรมและการประยุกต์ใช้เทคโนโลยีปัญญาประดิษฐ์อย่างมั่นคงปลอดภัยและน่าเชื่อถือ ซึ่งจะนำไปสู่การสร้างความเชื่อมั่นและขับเคลื่อนเศรษฐกิจดิจิทัลของประเทศให้เติบโตอย่างยั่งยืนต่อไป

ทั้งนี้สิ่งสำคัญที่ต้องตระหนัก คือ ภูมิทัศน์ของเทคโนโลยีปัญญาประดิษฐ์และภัยคุกคามที่เกี่ยวข้องนั้นมีวิวัฒนาการอย่างไม่หยุดนิ่ง ดังนั้น แนวปฏิบัติฉบับนี้จึงถือเป็นเอกสารที่จะต้องมีการทบทวนและพัฒนาเป็นประจำ และอาจมีการจัดทำส่วนเพิ่มเติมในอนาคตอย่างต่อเนื่อง เพื่อให้สามารถตามทันวิวัฒนาการของปัญญาประดิษฐ์และภัยคุกคามที่เปลี่ยนแปลงอย่างรวดเร็ว พร้อมทั้งสามารถรับมือกับความท้าทายและเทคโนโลยีใหม่ได้อย่างทันท่วงที

๑.๒ วัตถุประสงค์ของแนวปฏิบัติการใช้ปัญญาประดิษฐ์อย่างมั่นคงปลอดภัย

๑) เพื่อแนะนำแนวปฏิบัติที่ดีในการดำเนินงานเพื่อรักษาความมั่นคงปลอดภัยไซเบอร์สำหรับระบบปัญญาประดิษฐ์ที่เป็นรูปธรรมและนำไปปฏิบัติได้จริงสำหรับองค์กรในประเทศไทย

๒) เพื่อยกระดับความตระหนักรู้และความเข้าใจเกี่ยวกับภัยคุกคามไซเบอร์ที่จำเพาะต่อระบบปัญญาประดิษฐ์ เช่น การวางยาข้อมูล (Data Poisoning) การหลบหลีกโมเดล (Model Evasion) และการโจมตีผ่านชุดคำสั่ง (Prompt Injection)

๓) เพื่อส่งเสริมให้เกิดการพัฒนาและประยุกต์ใช้ปัญญาประดิษฐ์อย่างมั่นคงปลอดภัยตามหลักการออกแบบอย่างมั่นคงปลอดภัย การตั้งค่าเริ่มต้นอย่างมั่นคงปลอดภัย และการป้องกันเชิงลึก

๔) เพื่อสนับสนุนและสอดประสานกับแนวทางด้านธรรมาภิบาลปัญญาประดิษฐ์ และกฎหมายที่เกี่ยวข้องของไทย เช่น พระราชบัญญัติการรักษาความมั่นคงปลอดภัยไซเบอร์ พ.ศ. ๒๕๖๒ และพระราชบัญญัติคุ้มครองข้อมูลส่วนบุคคล พ.ศ. ๒๕๖๒

๕) เพื่อยกระดับการรักษาความมั่นคงปลอดภัยไซเบอร์สำหรับปัญญาประดิษฐ์ของประเทศไทยให้ทัดเทียมกับระดับสากล

๑.๓ กลุ่มเป้าหมายและการนำไปใช้

แนวปฏิบัติฉบับนี้ถูกออกแบบมาเพื่อเป็นประโยชน์ต่อผู้มีส่วนได้ส่วนเสียในหลายระดับเพื่อส่งเสริมการพัฒนาที่มั่นคงของอุตสาหกรรมปัญญาประดิษฐ์ จำเป็นต้องมีการกำกับดูแลอย่างเป็นระบบ เนื่องจากความซับซ้อนและความไม่แน่นอนที่มีอยู่โดยธรรมชาติ รูปแบบการกำกับดูแล

แบบดั้งเดิมที่พึ่งพาทรัพยากรจากภาครัฐเพียงอย่างเดียวไม่เพียงพอที่จะรับมือกับความท้าทายในการกำกับดูแลปัญญาประดิษฐ์ ดังนั้น ผู้มีส่วนได้ส่วนเสียและทุกภาคส่วนที่เกี่ยวข้องจำเป็นต้องเข้ามามีส่วนร่วมอย่างแท้จริง

เพื่อให้สอดคล้องกับมาตรฐานสากล ISO/IEC 42001:2023 แนวปฏิบัติฉบับนี้ได้จำแนกกลุ่มเป้าหมายซึ่งเป็นผู้มีส่วนได้เสีย ออกเป็น ๓ กลุ่มหลัก ดังนี้

๑) กลุ่มผู้มีส่วนได้ส่วนเสียภายในองค์กร

๑.๑) ผู้บริหารระดับสูง เป็นผู้รับผิดชอบสูงสุดในการกำหนดนโยบาย จัดสรรทรัพยากร และยอมรับความเสี่ยงที่เกี่ยวข้องกับปัญญาประดิษฐ์

๑.๒) ทีมพัฒนาและดำเนินงานปัญญาประดิษฐ์ รวมถึงนักวิทยาศาสตร์ข้อมูล วิศวกรปัญญาประดิษฐ์ และผู้ดูแลระบบ

๑.๓) ฝ่ายกฎหมายและการปฏิบัติตามข้อกำหนด เป็นผู้ดูแลให้การพัฒนาระบบปัญญาประดิษฐ์ให้สอดคล้องกับกฎหมายและกฎระเบียบต่าง ๆ

๑.๔) ฝ่ายความมั่นคงปลอดภัยไซเบอร์ เป็นผู้รับผิดชอบในการป้องกันภัยคุกคามทางไซเบอร์ที่มุ่งเป้ามายังระบบปัญญาประดิษฐ์

๑.๕) พนักงานผู้ใช้งาน เป็นพนักงานที่นำระบบปัญญาประดิษฐ์ไปใช้เป็นเครื่องมือในการปฏิบัติงาน

๒) กลุ่มผู้มีส่วนได้ส่วนเสียภายนอกองค์กร

๒.๑) ลูกค้าและผู้ใช้งานปลายทาง เป็นกลุ่มผู้ที่ใช้บริการหรือผลิตภัณฑ์ปัญญาประดิษฐ์โดยตรง

๒.๒) เจ้าของข้อมูล หมายถึงบุคคลที่ข้อมูลของบุคคลนั้นถูกนำไปใช้ในการฝึกสอนหรือประมวลผลโดยระบบปัญญาประดิษฐ์

๒.๓) พันธมิตรทางธุรกิจ หรือบริษัทคู่ค้าที่อาจมีการเชื่อมต่อหรือแลกเปลี่ยนข้อมูลกับระบบปัญญาประดิษฐ์

๒.๔) ผู้ขายและผู้ให้บริการรวมถึงผู้ให้บริการแพลตฟอร์มคลาวด์ ผู้ขายชุดข้อมูล และผู้พัฒนาโมเดลพื้นฐาน

๓) กลุ่มสังคมและกลุ่มหน่วยงานกำกับดูแล

๓.๑) หน่วยงานกำกับดูแล เป็นหน่วยงานภาครัฐที่มีหน้าที่กำกับดูแลการใช้เทคโนโลยี เช่น สำนักงานคณะกรรมการการรักษาความมั่นคงปลอดภัยไซเบอร์แห่งชาติ (สกมช.) สำนักงานพัฒนาธุรกรรมทางอิเล็กทรอนิกส์ (สพธอ.) และ สำนักงานคณะกรรมการคุ้มครองข้อมูลส่วนบุคคล (สคส.)

๓.๒) หน่วยงานรับรองและตรวจสอบ เป็นองค์กรที่ให้บริการด้านการรับรองแก่ผู้ใช้และผู้ให้บริการ โดยอาศัยมาตรฐานการรับรองที่ชัดเจน

๓.๓) กลุ่มประชาสังคมและนักวิชาการ รวมถึงองค์กรไม่แสวงหาผลกำไร นักวิจัย และสถาบันการศึกษาที่จับตาดูผลกระทบของปัญญาประดิษฐ์ต่อสังคม

๓.๔) สาธารณะ ชุมชนและสังคมโดยรวมที่อาจได้รับผลกระทบทางอ้อมจากการนำระบบปัญญาประดิษฐ์มาใช้งาน

๑.๔ ขอบเขตของแนวปฏิบัติการใช้ปัญญาประดิษฐ์อย่างมั่นคงปลอดภัย

๑) สิ่งที่อยู่ในขอบเขต

๑.๑) การป้องกันการใช้ปัญญาประดิษฐ์ในทางที่ผิดเพื่อโจมตีระบบอื่น เช่น การสร้าง Deepfake เพื่อหลอกลวง หรือการใช้ปัญญาประดิษฐ์พัฒนามัลแวร์

๑.๒) การปฏิบัติด้านความมั่นคงปลอดภัยตลอดวงจรชีวิตของระบบปัญญาประดิษฐ์ ตั้งแต่การออกแบบ การพัฒนา การติดตั้งใช้งาน การดำเนินงานและบำรุงรักษา จนถึงการกำจัดและทำลาย โดยมุ่งเน้นที่การป้องกันระบบปัญญาประดิษฐ์จากการถูกโจมตี

๑.๓) การป้องกันภัยคุกคามที่จำเพาะต่อระบบปัญญาประดิษฐ์ เช่น การโจมตีข้อมูล การโจมตีโมเดล และการโจมตีโครงสร้างพื้นฐาน

๑.๔) การบริหารจัดการความเสี่ยงด้านความมั่นคงปลอดภัยไซเบอร์สำหรับปัญญาประดิษฐ์

๑.๕) การรักษาความมั่นคงปลอดภัยของข้อมูล โมเดล แอปพลิเคชัน และโครงสร้างพื้นฐาน

๑.๖) การประยุกต์ใช้กับระบบปัญญาประดิษฐ์ที่องค์กรพัฒนาขึ้นเอง ระบบที่นำส่วนประกอบ Open Source และโมเดลที่ฝึกไว้ล่วงหน้า จากแหล่งภายนอกมาติดตั้ง ปรับจูน หรือใช้งาน

๒) สิ่งที่ไม่อยู่ในขอบเขต

๒.๑) ประเด็นด้านจริยธรรม ความเท่าเทียม ความโปร่งใส และความเป็นธรรมของปัญญาประดิษฐ์ ซึ่งได้กล่าวถึงอย่างละเอียดแล้วในเอกสารที่จัดทำโดยศูนย์ธรรมาภิบาลปัญญาประดิษฐ์

๒.๒) ปัญญาประดิษฐ์เพื่อการป้องกันความมั่นคงปลอดภัย กล่าวคือการประยุกต์ใช้เทคโนโลยีปัญญาประดิษฐ์ในการระบุ ป้องกัน ตรวจจับ วิเคราะห์ และตอบสนองต่อภัยคุกคามในเชิงรุก ครอบคลุมสภาพแวดล้อมทางกายภาพ ดิจิทัล และแบบผสมผสาน โดยมีเป้าหมายเพื่อเพิ่มความเร็ว ความแม่นยำ ประสิทธิภาพ และความสามารถในการขยายขนาดของการปฏิบัติการด้านความมั่นคงปลอดภัยให้เหนือกว่าขีดความสามารถของมนุษย์

บทที่ ๒
หลักการพื้นฐาน มาตรฐาน และภาพรวมภัยคุกคามด้านความมั่นคงปลอดภัยระบบปัญญาประดิษฐ์

บทนี้มีวัตถุประสงค์เพื่อวางหลักการพื้นฐานทางทฤษฎีและเชื่อมโยงแนวปฏิบัติในแนวปฏิบัติฉบับนี้เข้ากับมาตรฐานสากลและกฎระเบียบที่เกี่ยวข้อง เพื่อให้องค์กรสามารถนำไปปรับใช้ได้อย่างสอดคล้องกับหลักการที่เป็นที่ยอมรับในระดับโลกและบริบทของประเทศไทย

๒.๑ ความหมายและความแตกต่างระหว่างปัญญาประดิษฐ์กับระบบปัญญาประดิษฐ์

ปัญญาประดิษฐ์ คือ สาขาหนึ่งของวิทยาการคอมพิวเตอร์ที่สร้างระบบและซอฟต์แวร์ที่สามารถทำงานได้ในลักษณะเดียวกันกับความสามารถเฉพาะของมนุษย์ ปัญญาประดิษฐ์ช่วยให้เครื่องจักรสามารถเรียนรู้จากประสบการณ์ ปรับตัวเข้ากับข้อมูลใหม่ ใช้ข้อมูลอัลกอริทึม และพลังการประมวลผลเพื่อตีความสถานการณ์ที่ซับซ้อนและตัดสินใจ โดยใช้การมีส่วนร่วมของมนุษย์น้อยที่สุด ปัญญาประดิษฐ์สามารถเข้าใจภาษา จดจำรูปแบบ แก้ปัญหา และแม้กระทั่งแสดงความคิดสร้างสรรค์ได้ ซึ่งบ่อยครั้งที่ทำได้ด้วยความรวดเร็วและเหนือกว่าความสามารถของมนุษย์เป็นอย่างมาก

ระบบปัญญาประดิษฐ์ แนวปฏิบัติฉบับนี้นำเสนอตัวอย่างคำนิยามของระบบปัญญาประดิษฐ์จากแหล่งที่น่าเชื่อถือ ดังนี้

๑) The Organisation for Economic Co-operation and Development (OECD) นิยามของระบบปัญญาประดิษฐ์ คือระบบที่ทำงานบนเครื่องจักร โดยมีวัตถุประสงค์ที่ชัดเจนหรือโดยนัย จะทำการอนุมานจากข้อมูลที่ได้รับเพื่อสร้างผลลัพธ์ เช่น การคาดการณ์เนื้อหา คำแนะนำหรือการตัดสินใจ ที่สามารถมีอิทธิพลต่อสภาพแวดล้อมทางกายภาพหรือเสมือนจริงได้ ระบบปัญญาประดิษฐ์แต่ละระบบมีความแตกต่างกันในด้านระดับความเป็นอิสระ และความสามารถในการปรับตัวหลังจากการนำไปใช้งาน

๒) สหภาพยุโรป ในกฎหมาย EU AI Act ได้ให้นิยามระบบปัญญาประดิษฐ์ว่าเป็นระบบที่ทำงานบนเครื่องจักรซึ่งถูกออกแบบมาให้ทำงานด้วยระดับความเป็นอิสระที่แตกต่างกัน และอาจแสดงความสามารถในการปรับตัวหลังจากการนำไปใช้งาน และสำหรับวัตถุประสงค์ที่ชัดเจนหรือโดยนัย จะทำการอนุมานจากข้อมูลที่ได้รับ เพื่อสร้างผลลัพธ์ เช่น การคาดการณ์เนื้อหา คำแนะนำหรือการตัดสินใจ ที่สามารถมีอิทธิพลต่อสภาพแวดล้อมทางกายภาพหรือเสมือนจริงได้

คำจำกัดความสำคัญแสดงให้เห็นถึงความแตกต่างระหว่างคำว่า ปัญญาประดิษฐ์และระบบปัญญาประดิษฐ์ คือ ปัญญาประดิษฐ์มี “ความสามารถ” ในการดำเนินกิจกรรมที่มนุษย์ทำได้ ในขณะที่ระบบปัญญาประดิษฐ์เป็น "โซลูชันทางวิศวกรรม" ที่นำความสามารถนั้นมาใช้งาน

๒.๒ ประเภทของปัญญาประดิษฐ์ตามการใช้งาน

ปัญญาประดิษฐ์สามารถจำแนกเป็นกลุ่มใหญ่ตามประเภทการใช้งานที่พบได้บ่อย มีดังต่อไปนี้

๑) การประมวลผลภาษาธรรมชาติ (Natural Language Processing: NLP) มุ่งเน้นให้เครื่องจักรสามารถเข้าใจ ตีความ และสร้างภาษามนุษย์ได้ ตัวอย่างการใช้งาน เช่น แชทบอทและผู้ช่วยเสมือนการแปลภาษา เป็นต้น

๒) คอมพิวเตอร์วิทัศน์ (Computer Vision: CV) ช่วยให้เครื่องจักรสามารถ "มองเห็น" และตีความข้อมูลภาพจากรูปภาพ วิดีโอ หรือฉากในโลกแห่งความเป็นจริง ตัวอย่างการใช้งาน เช่น การจดจำรูปภาพ การวิเคราะห์วิดีโอ เป็นต้น

๓) หุ่นยนต์และระบบอัตโนมัติ (Robotics & Autonomous Systems) ผสมผสานปัญญาประดิษฐ์เข้ากับฮาร์ดแวร์เพื่อสร้างเครื่องจักรที่สามารถทำงานได้ด้วยตนเองโดยอัตโนมัติหรือกึ่งอัตโนมัติ ตัวอย่างการใช้งาน เช่น หุ่นยนต์บริการ ยานยนต์อัตโนมัติ เป็นต้น

๔) การวิเคราะห์เชิงพยากรณ์ (Predictive Analytics) ใช้ข้อมูลในอดีตเพื่อคาดการณ์ผลลัพธ์หรือแนวโน้มในอนาคต ตัวอย่างการใช้งาน เช่น การพยากรณ์อากาศ การคาดการณ์ทางการเงิน

๕) ปัญญาประดิษฐ์แบบรู้สร้าง (Generative AI หรือ GenAI) เป็นปัญญาประดิษฐ์ที่มีความสามารถในการสร้างข้อมูลใหม่ เช่น ข้อความ รูปภาพ เสียง หรือวีดีโอ ที่ไม่เคยมีมาก่อนด้วยตัวเอง

๖) ปัญญาประดิษฐ์แบบตัวแทน (Agentic AI) เป็นปัญญาประดิษฐ์ที่สามารถดำเนินการและตัดสินใจได้ด้วยตัวเอง สามารถเข้าใจเป้าหมายที่ชัดเจน สามารถเรียนรู้ ปรับตัว ตัดสินใจ และดำเนินการด้วยตัวเองเพื่อให้บรรลุเป้าหมายที่ตั้งไว้ได้ คุณสมบัติสำคัญที่ทำให้ Agentic AI แตกต่างคือความสามารถในการเป็น “ผู้กระทำ” ที่สามารถเชื่อมต่อและสั่งการระบบอื่น ๆ เพื่อปฏิบัติภารกิจให้สำเร็จลุล่วงได้โดยอัตโนมัติ

๗) ปัญญาประดิษฐ์เชิงกายภาพ (Physical AI) มีความแตกต่างจากปัญญาประดิษฐ์ทั่วไปที่มีการทำงานอยู่ในโลกดิจิทัล ปัญญาประดิษฐ์เชิงกายภาพมีการเชื่อมต่อกับระบบที่อยู่ในโลกกายภาพ เช่น หุ่นยนต์และระบบอัตโนมัติ ทำให้สามารถรับรู้และตอบสนองในสภาพแวดล้อมกายภาพได้

๒.๓ หมวดหมู่ความเสี่ยงของปัญญาประดิษฐ์

แนวปฏิบัติฉบับนี้เสนอแนะกรอบการจำแนกประเภทความเสี่ยงสำหรับระบบปัญญาประดิษฐ์ โดยอ้างอิงตามแนวปฏิบัติของกฎหมายว่าด้วยปัญญาประดิษฐ์แห่งสหภาพยุโรป (EU AI Act) ซึ่งเป็นกฎหมายต้นแบบที่ได้รับการยอมรับในระดับสากล

ระดับที่ ๑ ความเสี่ยงที่ยอมรับไม่ได้

คำนิยาม ระบบปัญญาประดิษฐ์ที่ถูกจัดประเภทว่ามีคุณลักษณะการทำงานที่ก่อให้เกิดภัยคุกคามต่อความปลอดภัย สิทธิขั้นพื้นฐาน และคุณค่าที่เป็นที่ยอมรับในสังคมสากล

ตัวอย่าง

๑) ระบบการให้คะแนนทางสังคม การใช้ระบบปัญญาประดิษฐ์ โดยหน่วยงานภาครัฐเพื่อประเมินหรือจัดลำดับความน่าเชื่อถือของบุคคลธรรมดา ซึ่งนำไปสู่การปฏิบัติที่ไม่เป็นธรรมหรือสร้างความเสียหาย

๒) ระบบที่มุ่งบงการพฤติกรรมมนุษย์โดยใช้เทคนิคจิตใต้สำนึก การใช้เทคนิคจิตใต้สำนึก หรือการใช้ประโยชน์จากกลุ่มเปราะบาง เช่น เด็ก ผู้สูงอายุ หรือผู้พิการ เพื่อชักจูงการตัดสินใจที่อาจก่อให้เกิดอันตราย

๓) ระบบบ่งชี้ตัวตนทางชีวมิติแบบเรียลไทม์ การใช้ระบบปัญญาประดิษฐ์เพื่อวิเคราะห์ข้อมูลชีวมิติจากระยะไกลในพื้นที่สาธารณะแบบเรียลไทม์ เพื่อวัตถุประสงค์ในการบังคับใช้กฎหมาย ยกเว้นกรณีที่ได้รับอนุญาตอย่างเข้มงวดสูงสุดตามกฎหมาย

ข้อเสนอแนะ องค์กรพึงกำหนดนโยบายและมาตรการควบคุมภายในที่ชัดเจนในการ **ห้าม** การริเริ่ม พัฒนา จัดหา หรือนำระบบปัญญาประดิษฐ์ที่เข้าข่ายในกลุ่มนี้มาใช้งานภายในองค์กรโดยสิ้นเชิง

หมายเหตุ อาจมีข้อยกเว้นบางประการเพื่อวัตถุประสงค์ในการบังคับใช้กฎหมาย โดยระบบการระบุตัวตนทางชีวมิติระยะไกลแบบ "เรียลไทม์" จะได้รับอนุญาตให้ใช้ในคดีร้ายแรงจำนวนจำกัด ในขณะที่ระบบการระบุตัวตนทางชีวมิติระยะไกลแบบ "ย้อนหลัง" ซึ่งเป็นการระบุตัวตนที่เกิดขึ้นหลังจากเวลาผ่านไปพอสมควร จะได้รับอนุญาตให้ใช้เพื่อดำเนินคดีในอาชญากรรมร้ายแรงและต้องได้รับการอนุมัติจากศาลก่อนเท่านั้น

ระดับที่ ๒ ความเสี่ยงสูง

คำนิยาม ระบบปัญญาประดิษฐ์ที่แม้จะได้รับอนุญาตให้ใช้งานได้ แต่มีศักยภาพในการส่งผลกระทบเชิงลบอย่างมีนัยสำคัญต่อสุขภาพ ความปลอดภัย หรือสิทธิขั้นพื้นฐานของบุคคล และจำเป็นต้องอยู่ภายใต้การกำกับดูแลที่เข้มงวด

ตัวอย่าง

๑) ระบบที่ใช้ในโครงสร้างพื้นฐานที่สำคัญ เช่น การบริหารจัดการโครงข่ายพลังงาน การควบคุมระบบจราจร

๒) ระบบที่ใช้ในการบริหารทรัพยากรมนุษย์ เช่น การคัดกรองใบสมัคร การประเมินผลการปฏิบัติงาน การตัดสินใจเลิกจ้าง

๓) ระบบที่เกี่ยวข้องกับผลิตภัณฑ์ด้านความปลอดภัย เช่น ส่วนประกอบด้านความปลอดภัยในยานยนต์ อุปกรณ์การแพทย์

๔) ระบบที่ใช้ในการประเมินสินเชื่อและการเข้าถึงบริการทางการเงิน

๕) ระบบที่ใช้ในการบังคับใช้กฎหมายและการพิจารณาคดี

ข้อเสนอแนะ โครงการที่เข้าข่ายความเสี่ยงสูงจะต้องแสดงให้เห็นถึงความสามารถในการปฏิบัติตามพันธกรณี ๗ ข้อ ตลอดวงจรชีวิตของระบบปัญญาประดิษฐ์ ดังต่อไปนี้

๑) การบริหารจัดการความเสี่ยง

๒) การกำกับดูแลข้อมูล

๓) การจัดทำเอกสารทางเทคนิค

๔) การเก็บบันทึกการทำงาน

๕) การให้ข้อมูลและความโปร่งใส

๖) การกำกับดูแลโดยมนุษย์

๗) การรับประกันความแม่นยำ ความทนทาน และความมั่นคงปลอดภัยทางไซเบอร์

ระดับที่ ๓ ความเสี่ยงจำกัด

คำนิยาม ระบบปัญญาประดิษฐ์ที่ความเสี่ยงหลักเกิดจากการขาดความโปร่งใส ซึ่งอาจทำให้ผู้ใช้ไม่ตระหนักว่ากำลังมีปฏิสัมพันธ์กับระบบปัญญาประดิษฐ์

ตัวอย่าง

๑) ระบบสนทนาอัตโนมัติ (Chatbots)

๒) ระบบที่สร้างเนื้อหาประเภท Deepfake

๓) ระบบที่สร้างเนื้อหาอื่น ๆ จากปัญญาประดิษฐ์ (AI-generated Content)

ข้อเสนอแนะ องค์กรต้องดำเนินการตาม**พันธกรณีด้านความโปร่งใส** โดยต้องเปิดเผยข้อมูลให้ผู้ใช้งานรับทราบอย่างชัดเจน

ระดับที่ ๔ ความเสี่ยงน้อยที่สุด

คำนิยาม ระบบปัญญาประดิษฐ์ประเภทอื่น ๆ ที่ไม่เข้าข่ายทั้งสามระดับข้างต้น และมีความเสี่ยงต่ำ

ตัวอย่าง

๑) ระบบกรองอีเมลขยะ

๒) ปัญญาประดิษฐ์ที่ใช้ในวิดีโอเกม

๓) ระบบบริหารจัดการสินค้าคงคลัง

ข้อเสนอแนะ องค์กรพึงส่งเสริมให้การพัฒนาและการใช้งานเป็นไปตาม**หลักจรรยาบรรณโดยสมัครใจ**

ตารางที่ ๑ สรุปหมวดหมู่ความเสี่ยงของปัญญาประดิษฐ์และการดำเนินการตามแนวทางของ EU AI Act

ระดับความเสี่ยง	คำอธิบาย	สิ่งที่ต้องดำเนินการ
ยอมรับไม่ได้ (Unacceptable)	เป็นภัยคุกคามต่อสิทธิมนุษยชน	ห้ามทำโดยเด็ดขาด นอกจากมีข้อยกเว้นตามกฎหมาย
สูง (High)	อาจส่งผลกระทบต่อความปลอดภัยและสิทธิขั้นพื้นฐาน	ทำได้แต่ต้องปฏิบัติตามพันธกรณี ๗ ข้ออย่างเข้มงวด
จำกัด (Limited)	เสี่ยงต่อความไม่โปร่งใส	ทำได้แต่ต้องแจ้งให้ผู้ใช้ทราบ
น้อยที่สุด (Minimal)	ความเสี่ยงต่ำมาก	ทำได้โดยอิสระ (ส่งเสริมให้ดำเนินการตามหลักจรรยาบรรณ)

๒.๔ ประโยชน์ของปัญญาประดิษฐ์

ความสามารถของปัญญาประดิษฐ์ในการประมวลผลข้อมูล การเรียนรู้จากรูปแบบ และการทำงานอัตโนมัติ นำไปสู่ความก้าวหน้าอย่างมีนัยสำคัญในด้านความสามารถในการผลิต การดูแลสุขภาพ ความปลอดภัย ความมั่นคงปลอดภัย การบริการลูกค้า การวิจัยทางวิทยาศาสตร์ และการเข้าถึง ทำให้ปัญญาประดิษฐ์กลายเป็นพลังแห่งการเปลี่ยนแปลงในสังคมสมัยใหม่ ตัวอย่างเช่น

๑) เพิ่มประสิทธิภาพการทำงานซ้ำ ๆ โดยอัตโนมัติ เช่น การป้อนข้อมูล การบริการลูกค้า ช่วยประหยัดเวลาและลดความผิดพลาดของมนุษย์

๒) การแก้ปัญหาขั้นสูง โดยการวิเคราะห์ชุดข้อมูลขนาดใหญ่ ค้นหาข้อมูลเชิงลึกเพื่อช่วยในสาขาต่าง ๆ เช่น การวิจัยทางการแพทย์ การตรวจหาโรค และการสร้างแบบจำลองสภาพภูมิอากาศ

๓) เพิ่มการเข้าถึง โดยเป็นกลไกสำคัญของเครื่องมือสำหรับผู้พิการ เช่น การแปลงเสียงพูดเป็นข้อความ และลดข้อจำกัดในความเข้าใจภาษาผ่านการแปลภาษาแบบเรียลไทม์

๔) เร่งสร้างนวัตกรรม โดยขับเคลื่อนความก้าวหน้าในด้านต่าง ๆ เช่น การค้นพบยา ระบบอัตโนมัติ และการศึกษาเฉพาะบุคคล

๕) ปรับปรุงความปลอดภัย โดยตรวจสอบและคาดการณ์ความเสี่ยงในภาคส่วนต่าง ๆ เช่น ในกระบวนการผลิต และการรับมือภัยพิบัติ

๖) ลดต้นทุน ปรับการใช้ทรัพยากรให้เหมาะสมที่สุดในด้านโลจิสติกส์ การจัดการพลังงาน และการดูแลสุขภาพ ช่วยลดค่าใช้จ่ายในการดำเนินงาน

๒.๕ ความเสี่ยงจากการใช้ปัญญาประดิษฐ์ที่ไม่มั่นคงปลอดภัย

การใช้ปัญญาประดิษฐ์ที่ไม่มั่นคงปลอดภัยก่อให้เกิดความเสี่ยงในหลายแง่มุม ทั้งอันตรายทางเทคนิคที่ซ่อนอยู่ รวมถึงความท้าทายทางสังคมและจริยธรรม ในมุมมองทางเทคนิค อคติในข้อมูลที่ใช้ฝึกฝนอาจทำให้ระบบปัญญาประดิษฐ์ตัดสินใจในลักษณะเลือกปฏิบัติ เช่น การปฏิบัติอย่างไม่เป็นธรรมต่อกลุ่มคนบางกลุ่มในการรับสมัครงานหรือการประเมินสินเชื่อ ความไม่โปร่งใสของอัลกอริทึมอาจทำให้การตรวจสอบย้อนกลับการตัดสินใจที่ผิดพลาดทำได้ยาก ซึ่งอาจเป็นอันตรายถึงชีวิตโดยตรงเมื่อนำไปใช้ในสาขาที่สำคัญอย่างการวินิจฉัยทางการแพทย์และการขับขี่อัตโนมัติ ในระดับสังคม การใช้ปัญญาประดิษฐ์ในทางที่ผิดอาจนำไปสู่การรั่วไหลของข้อมูลส่วนบุคคล โดยข้อมูลส่วนตัวจำนวนมากถูกรวบรวมและใช้อย่างผิดกฎหมาย หรือแม้กระทั่งใช้เพื่อการฉ้อโกงแบบเจาะจง ยิ่งไปกว่านั้น ช่องโหว่ในระบบปัญญาประดิษฐ์อาจถูกแฮกเกอร์ใช้เป็นประโยชน์ ก่อให้เกิดเหตุการณ์ด้านความมั่นคงปลอดภัยไซเบอร์ที่เป็นภัยคุกคามต่อบุคคล องค์กร ตลอดจนความปลอดภัยสาธารณะ ตัวอย่างเช่น การนำปัญญาประดิษฐ์ไปใช้สร้าง Deepfake เพื่อบ่อนทำลายชื่อเสียง หรือใช้เพื่อเร่งกระบวนการพัฒนามัลแวร์ที่ซับซ้อน ซึ่งเป็นการนำเทคโนโลยีไปใช้ในทางที่ผิดโดยตรง

๒.๖ ภาพรวมภัยคุกคามด้านความมั่นคงปลอดภัยของปัญญาประดิษฐ์

ภัยคุกคามต่อระบบปัญญาประดิษฐ์ส่วนใหญ่สามารถจัดประเภทภายใต้การโจมตีโมเดลการเรียนรู้ของเครื่องด้วยข้อมูลปนเปื้อน (Adversarial Machine Learning: AML) ซึ่งหมายถึงการจงใจสร้างข้อมูลนำเข้าเพื่อใช้ประโยชน์จากช่องโหว่ของโมเดล และทำให้โมเดลมีพฤติกรรมที่ไม่พึงประสงค์ การจำแนกประเภทภัยคุกคามในแนวปฏิบัติฉบับนี้อ้างอิงตามมาตรฐานและแนวปฏิบัติสากล เช่น OWASP AI Exchange และ ISO/IEC 23894:2023 โดยแบ่งออกเป็น ๒ กลุ่มหลัก ดังนี้

กลุ่มที่ ๑ ภัยคุกคามภายใต้การโจมตีโมเดลการเรียนรู้ของเครื่องด้วยข้อมูลปนเปื้อน

ภัยคุกคามในกลุ่มนี้เป็นการโจมตีที่มุ่งเป้าไปที่ตรรกะการทำงานของปัญญาประดิษฐ์โดยตรง เพื่อบิดเบือน หรือใช้ประโยชน์จากช่องโหว่ของโมเดล

๑) การปลอมปนหรือการวางยาข้อมูล (Data Poisoning)

๑.๑) คำอธิบาย เป็นการแทรกแซงชุดข้อมูลฝึกสอนด้วยข้อมูลที่ถูกสร้างขึ้นอย่างมุ่งร้าย เพื่อบ่อนทำลายความสมบูรณ์ของโมเดลตั้งแต่ขั้นตอนการฝึกสอน

๑.๒) เกิดขึ้นได้อย่างไร ผู้โจมตีแทรกข้อมูลที่มีเงื่อนไขแฝงเข้าไปในชุดข้อมูล ซึ่งเมื่อโมเดลได้เรียนรู้ข้อมูลดังกล่าวไปแล้ว จะส่งผลให้เกิดพฤติกรรมที่ผิดพลาดตามที่ผู้โจมตีกำหนด หรือเป็นการสร้างประตูหลังในโมเดลเมื่อถูกกระตุ้นด้วยเงื่อนไขนั้น ๆ ในขั้นตอนการใช้งานจริง

๒) การละเมิดความเป็นส่วนตัวของข้อมูล (Data Privacy Breach)

๒.๑) คำอธิบาย โมเดลปัญญาประดิษฐ์อาจจดจำข้อมูลที่ละเอียดอ่อนจากชุดข้อมูลฝึกสอนโดยไม่ได้เจตนา และอาจเปิดเผยข้อมูลดังกล่าวออกมาได้ในขั้นตอนการอนุมาน

๒.๒) เกิดขึ้นได้อย่างไร ผู้โจมตีใช้เทคนิคการโจมตีแบบอนุมาน (Inference Attacks) เช่น Model Inversion เพื่อพยายามสร้างข้อมูลฝึกสอนบางส่วนขึ้นมาใหม่ หรือใช้ Membership Inference เพื่อตรวจสอบว่าข้อมูลของบุคคลใดบุคคลหนึ่งเคยถูกใช้ในการฝึกโมเดลหรือไม่ ซึ่งเป็นการยืนยันความสัมพันธ์ระหว่างบุคคลกับข้อมูลที่ละเอียดอ่อน

๓) การหลบหลีก (Evasion)

๓.๑) คำอธิบาย เป็นการสร้างข้อมูลนำเข้าเชิงปรปักษ์ในขั้นตอนการอนุมาน เพื่อทำให้โมเดลเกิดการจำแนกประเภทที่ผิดพลาด

๓.๒) เกิดขึ้นได้อย่างไร ผู้โจมตีสร้าง "ตัวอย่างเชิงปรปักษ์" (Adversarial Example) โดยการเจือปนข้อมูลนำเข้าเล็กน้อย ซึ่งโดยปกติมนุษย์ไม่สามารถสังเกตเห็นได้ แต่เพียงพอที่จะหลอกให้โมเดลตัดสินใจผิดพลาดได้

๔) การโจมตีผ่านชุดคำสั่ง (Prompt Injection)

๔.๑) คำอธิบาย เป็นการสร้างชุดคำสั่ง (Prompt) ที่มีเจตนาหลีกเลี่ยงหรือทำลาย

มาตรการควบคุมความปลอดภัย (Safety Guardrails) ของโมเดลภาษาขนาดใหญ่ (Large Language Model: LLM) เพื่อให้โมเดลปฏิบัติงานนอกเหนือขอบเขตที่ได้รับอนุญาต

๔.๒) เกิดขึ้นได้อย่างไร การโจมตีมีทั้งรูปแบบทางตรง (ผู้ใช้สั่งให้โมเดลเพิกเฉยต่อข้อกำหนดเดิม) และทางอ้อม (การซ่อนคำสั่งอันตรายไว้ในแหล่งข้อมูลภายนอกที่โมเดลต้องดึงมาประมวลผล)

๕) การสกัดโมเดล (Model Extraction)

๕.๑) คำอธิบาย เป็นกระบวนการสร้างแบบจำลองทดแทน (Substitute Model) ที่มีความสามารถเทียบเคียงกับโมเดลเป้าหมาย ผ่านการสืบค้นข้อมูลจากส่วนต่อประสานโปรแกรมประยุกต์ (API) ซึ่งถือเป็นการลักลอบขโมยทรัพย์สินทางปัญญา

๕.๒) เกิดขึ้นได้อย่างไร ผู้โจมตีส่งคำสั่ง (Queries) จำนวนมากไปยัง API ของโมเดลเป้าหมาย และใช้คู่ข้อมูลนำเข้า-ผลลัพธ์ ที่ได้ในการฝึกสอนโมเดลของตนเองให้ทำงานลอกเลียนแบบ

๖) การวางยาโมเดล (Model Poisoning)

๖.๑) คำอธิบาย เป็นการโจมตีที่ซับซ้อนโดยการแทรกแซงกระบวนการฝึกสอนโดยตรงเพื่อปรับเปลี่ยนพารามิเตอร์หรือสถาปัตยกรรมของโมเดลให้เกิดช่องโหว่ตามที่ต้องการ

๖.๒) เกิดขึ้นได้อย่างไร ผู้โจมตีจะเข้าไปปรับเปลี่ยนน้ำหนักหรือโครงสร้างของโมเดลในระหว่างกระบวนการฝึก ซึ่งมักเกิดขึ้นในสถาปัตยกรรมที่มีการฝึกสอนแบบกระจายศูนย์ เช่น Federated Learning

๗) การปฏิเสธการให้บริการ (Denial of Service)

๗.๑) คำอธิบาย เป็นการโจมตีที่มุ่งเป้าไปที่การใช้ทรัพยากรประมวลผลของโมเดลอย่างสิ้นเปลือง (Resource Exhaustion) เพื่อลดทอนความพร้อมใช้งานของระบบ

๗.๒) เกิดขึ้นได้อย่างไร ผู้โจมตีส่งข้อมูลนำเข้าที่ได้รับการออกแบบมาเพื่อกระตุ้นให้โมเดลต้องใช้ทรัพยากรในการประมวลผลสูงกว่าปกติอย่างมีนัยสำคัญ ส่งผลให้ระบบทำงานช้าลงหรือไม่สามารถให้บริการได้

กลุ่มที่ ๒ ภัยคุกคามที่ไม่ใช่การโจมตีโมเดลการเรียนรู้ของเครื่องด้วยข้อมูลที่ปนเปื้อน

ภัยคุกคามในกลุ่มนี้เป็นการโจมตีโครงสร้างพื้นฐานและส่วนประกอบแวดล้อมของระบบปัญญาประดิษฐ์โดยใช้เทคนิคความปลอดภัยทางไซเบอร์แบบดั้งเดิม

๑) การโจมตีห่วงโซ่อุปทาน (AI Supply Chain Attacks)

๑.๑) คำอธิบาย เป็นการโจมตีที่ใช้ประโยชน์จากช่องโหว่ในส่วนประกอบของบุคคลที่สาม (Third-Party Components) ซึ่งเป็นส่วนหนึ่งของวงจรชีวิตการพัฒนาระบบปัญญาประดิษฐ์

๑.๒) เกิดขึ้นได้อย่างไร เกิดจากการที่นักพัฒนานำส่วนประกอบ เช่น ไลบรารี ซอฟต์แวร์ หรือโมเดลที่ฝึกไว้ล่วงหน้าจากแหล่งที่ไม่น่าเชื่อถือมาใช้ ซึ่งส่วนประกอบเหล่านั้น อาจถูกฝังมัลแวร์หรือมีช่องโหว่แฝงอยู่

๒) การโจมตีโครงสร้างพื้นฐาน (Attacks On Supporting Infrastructure)

๒.๑) คำอธิบาย เป็นการใช้ประโยชน์จากช่องโหว่ในโครงสร้างพื้นฐานเทคโนโลยีสารสนเทศแบบดั้งเดิมที่รองรับการทำงานของระบบปัญญาประดิษฐ์

๒.๒) เกิดขึ้นได้อย่างไร เกิดจากการใช้เทคนิคการโจมตีแบบดั้งเดิม เช่น SQL Injection, Cross-Site Scripting (XSS) หรือการกำหนดค่าการควบคุมการเข้าถึงที่ผิดพลาด (Broken Access Control)

๓) การขโมยโมเดลโดยตรง (Direct Model Theft)

๓.๑) คำอธิบาย เป็นการเข้าถึงระบบโดยไม่ได้รับอนุญาตเพื่อลักลอบนำไฟล์โมเดลซึ่งเป็นทรัพย์สินทางปัญญาออกไป

๓.๒) เกิดขึ้นได้อย่างไร เกิดจากมาตรการป้องกันของระบบที่ไม่รัดกุม เช่น การตั้งค่าสิทธิ์การเข้าถึงไฟล์ที่ไม่เหมาะสม ทำให้ผู้โจมตีสามารถดาวน์โหลดไฟล์โมเดลออกไปได้โดยตรง

ตารางที่ ๒ สรุปผลกระทบหลักของภัยคุกคามแต่ละประเภทต่อความลับ (Confidentiality) ความถูกต้องสมบูรณ์ (Integrity) และ ความพร้อมใช้งาน (Availability)

ประเภทภัยคุกคาม	ผลกระทบต่อ C-I-A	เหตุผล
๑. Data Poisoning	Integrity	เป็นการบ่อนทำลายความถูกต้องของโมเดลตั้งแต่กระบวนการฝึกสอน ทำให้ผลลัพธ์ขาดความน่าเชื่อถือ
๒. Data Privacy Breaches	Confidentiality	มีเป้าหมายเพื่อเปิดเผยข้อมูลที่ละเอียดอ่อนจากชุดข้อมูลฝึกสอน ซึ่งเป็นการละเมิดความลับของข้อมูลโดยตรง
๓. Evasion	Integrity	ทำให้โมเดลจำแนกประเภทผิดพลาด ณ ขณะใช้งาน ซึ่งเป็นการลดทอนความถูกต้องและความน่าเชื่อถือของผลลัพธ์
๔. Prompt Injection	Integrity, Confidentiality	การสั่งงานนอกขอบเขตกระทบต่อความถูกต้อง (I) และอาจนำไปสู่การเปิดเผยข้อมูลที่เป็นความลับ (C)
๕. Model Extraction	Confidentiality	เป็นการลอกเลียนแบบเพื่อขโมยทรัพย์สินทางปัญญา ซึ่งถือเป็นข้อมูลที่เป็นความลับขององค์กร
๖. Model Poisoning	Integrity	มีเป้าหมายเพื่อบิดเบือนการทำงานและผลลัพธ์ของโมเดลโดยตรง ทำให้ขาดความถูกต้องสมบูรณ์
๗. Denial of Service (DoS)	Availability	มุ่งเป้าให้ระบบหยุดทำงานหรือทำงานช้าลง ทำให้ผู้ใช้ที่ได้รับอนุญาตไม่สามารถเข้าถึงบริการได้
๘. AI Supply Chain	Confidentiality, Integrity, Availability	ส่วนประกอบที่มีช่องโหว่สามารถนำไปสู่การขโมยข้อมูล (C) การบิดเบือนการทำงาน (I) หรือทำให้ระบบล่ม (A)
๙. Infrastructure Attacks	Confidentiality, Integrity, Availability	ช่องโหว่ในระบบแวดล้อมอาจนำไปสู่การเข้าถึงข้อมูลโดยไม่ได้รับอนุญาต (C) การแก้ไขผลลัพธ์ (I) และการทำให้ระบบหยุดชะงัก (A)
๑๐. Direct Model Theft	Confidentiality	เป็นการขโมยไฟล์โมเดลโดยตรงซึ่งเป็นการละเมิดทรัพย์สินทางปัญญาที่เป็นความลับขององค์กร

นอกจากนี้ ในอนาคตอาจมีภัยคุกคามทางไซเบอร์ใหม่ที่เปลี่ยนแปลงไปกับการพัฒนาของปัญญาประดิษฐ์

ตารางที่ ๓ แสดงตัวอย่างภัยคุกคามทางไซเบอร์ยุคใหม่พร้อมแนวทางการควบคุมเบื้องต้น

ประเภทภัยคุกคาม	ตัวอย่าง	แนวทางควบคุมเบื้องต้น
๑.ภัยคุกคามจาก Generative AI		
๑.๑ การเจาะจงโทนข้อความ (Vibe Hacking)	การสร้างข้อความ อีเมลที่ปรับโทน น้ำเสียง และบริบทเลียนแบบผู้บริหาร เพื่อทำ Social Engineering	- การตรวจจับพฤติกรรมผิดปกติ (Behavioral Anomaly Detection) - ระบบตรวจสอบความแท้จริงของเนื้อหา (AI Content Authenticity เช่น C2PA Watermarking) - การสร้างความตระหนักรู้ให้กับพนักงานและผู้บริหาร
๑.๒ มัลแวร์เปลี่ยนแปลงลายมือชื่อ (Polymorphic Signature Malware)	มัลแวร์ที่ปรับเปลี่ยนโคดหรือลักษณะของตัวเองโดยอัตโนมัติด้วย Generative AI เพื่อหลบเลี่ยงการตรวจจับโดยระบบ Antivirus	- การตรวจจับแบบพฤติกรรม - การป้องกันแบบ Zero-Trust Endpoint Protection - การล่าภัยคุกคามอย่างต่อเนื่อง
๑.๓ Deepfake เช่าใช้ (Deepfake-as-a-Service: DFaaS)	การเช่าหรือจ้างทำวิดีโอหรือเสียงปลอม เช่น วิดีโอ CEO สั่งโอนเงิน	- การยืนยันตัวตนหลายปัจจัย (Multi-Factor Authentication: MFA) - การตรวจสอบ Deepfake ด้วย AI Forensics - นโยบายห้ามยืนยันธุรกรรมผ่านช่องทางเดียว (Dual-Channel Verification Policy)
๒.ภัยคุกคามจาก Agentic AI		
๒.๑ การโจมตีหลายขั้นตอนหรือหลายเพย์โหลด (Multi-Step /	Agentic AI วางแผนโจมตีต่อเนื่อง เช่น สแกนระบบ เปิด Backdoor ขโมยข้อมูล ปล่อย Ransomware	- การตรวจสอบแบบ Kill Chain Monitoring - การจำลองภัยคุกคาม - การใช้ระบบอัตโนมัติในการตอบสนอง

ประเภทภัยคุกคาม	ตัวอย่าง	แนวทางควบคุมเบื้องต้น
Multi-Payload Attacks)		เหตุการณ์ (Security Orchestration Automation and Response: SOAR)
๒.๒ ผู้โจมตีที่ใช้ปัญญาประดิษฐ์ (AI-Enabled Attacker)	Agentic AI ทำหน้าที่เป็นผู้โจมตีอัตโนมัติ โจมตีระบบได้ตลอดเวลาโดยไม่มีเวลาพักและเรียนรู้จากความล้มเหลว	- การป้องกันด้วย AI ต่อ AI (Autonomous Defense Agents) - การทดสอบเจาะระบบต่อเนื่อง - การใช้กับดักดิจิทัล (Cyber Deception) เช่น Honeypot/Decoy Systems
๒.๓ การโจมตีห่วงโซ่อุปทานอัตโนมัติ (Autonomous Supply Chain Attack)	Agentic AI โจมตีผ่าน Library องค์ประกอบโอเพนซอร์ส หรือ CI/CD Pipeline	- การใช้บัญชีรายการส่วนประกอบซอฟต์แวร์ที่ถูกกำกับและควบคุมเวอร์ชัน - การจัดการความเสี่ยงของผู้ขาย - การป้องกัน CI/CD Pipeline (Code Signing, Build Integrity)
๓.ภัยคุกคามจาก Physical AI		
๓.๑ การโจมตีแบบฝูงโดรนหรือหุ่นยนต์แบบกระจายศูนย์ (DDoS Drone/Robot Swarm)	การควบคุมโดรนหรือหุ่นยนต์จำนวนมากเพื่อโจมตีโครงสร้างพื้นฐาน เช่น สนามบิน โรงไฟฟ้า ศูนย์ข้อมูล	- การกำหนดเขตควบคุม (Geofencing & Remote ID) - ระบบตรวจจับการบุกรุกทางกายภาพ - การเตรียมระบบสำรองสำหรับระบบที่สำคัญ
๓.๒ โดรนขนาดแมลงสำหรับสอดแนม (Insect-Size	โดรนจิ๋วใช้ดักฟังหรือเก็บภาพเอกสารสำคัญ	- ระบบต่อต้านโดรน (Counter-UAV Systems: RF, Radar, Acoustic) - การกำหนดพื้นที่ป้องกัน (Shielded/Restricted Zones)

ประเภทภัยคุกคาม	ตัวอย่าง	แนวทางควบคุมเบื้องต้น
UAV Reconnaissance)		- การสร้างการรับรู้ด้านความปลอดภัยทางกายภาพ
๓.๓ การโจมตีทางกายภาพอัตโนมัติ (Autonomous Kinetic Attack)	รถไร้คนขับถูกควบคุมให้ชนหรือหุ่นยนต์อุตสาหกรรมถูกสั่งทำลายสายการผลิต	- การอัปเดตซอฟต์แวร์ที่ปลอดภัย (Secure OTA Updates) - กลไกหยุดฉุกเฉิน (AI Safety Kill Switch) - การตรวจสอบระบบควบคุมอุตสาหกรรม (ICS Security Monitoring)

๒.๗ มาตรฐานและกรอบการดำเนินงานสากล

การปฏิบัติตามมาตรฐานสากลช่วยให้องค์กรมีแนวทางที่ชัดเจนในการบริหารจัดการความมั่นคงปลอดภัยปัญญาประดิษฐ์และสร้างความน่าเชื่อถือในระดับนานาชาติ สำหรับการรักษาความมั่นคงปลอดภัยของปัญญาประดิษฐ์มีมาตรฐานและกรอบการดำเนินงานสากลที่เกี่ยวข้องดังต่อไปนี้

๑) มาตรฐานจาก ISO/IEC

๑.๑) ISO/IEC 42001:2023 (Artificial Intelligence — Management System) เป็นมาตรฐานสากลฉบับแรกของโลกสำหรับระบบการจัดการปัญญาประดิษฐ์ โดยกำหนดกรอบการดำเนินงานเพื่อให้องค์กรสามารถพัฒนาและใช้งานปัญญาประดิษฐ์อย่างมีความรับผิดชอบ ครอบคลุมทั้งด้านความเสี่ยง จริยธรรม และความโปร่งใส

๑.๒) ISO/IEC 23894:2023 (Guidance on AI Risk Management) ให้คำแนะนำเกี่ยวกับการบริหารจัดการความเสี่ยงที่เกี่ยวข้องกับปัญญาประดิษฐ์โดยเฉพาะ ซึ่งสามารถนำไปปรับใช้ร่วมกับกรอบการบริหารความเสี่ยงขององค์กรได้

๑.๓) ISO/IEC 27000 Series ชุดมาตรฐานด้านการจัดการความมั่นคงปลอดภัยของสารสนเทศ โดยเฉพาะ ISO/IEC 27001:2022 ยังคงเป็นหลักการพื้นฐานสำคัญในการรักษาความปลอดภัยของโครงสร้างพื้นฐานที่รองรับการทำงานของระบบปัญญาประดิษฐ์

๑.๔) ISO/IEC 5338:2023 กระบวนการวงจรชีวิตของระบบปัญญาประดิษฐ์ ระบุกระบวนการต่าง ๆ ที่สนับสนุนการนิยาม การควบคุม การจัดการ การดำเนินการ และการปรับปรุงระบบปัญญาประดิษฐ์ ในแต่ละขั้นตอนของวงจรชีวิตของระบบปัญญาประดิษฐ์ กระบวนการเหล่านี้ยังสามารถนำไปใช้ภายในองค์กรหรือโครงการในระหว่างการพัฒนาหรือการจัดหาระบบปัญญาประดิษฐ์ได้เช่นกัน

๒) กรอบการดำเนินงานระดับภูมิภาค

กรอบการดำเนินงานจาก European Union Agency for Cybersecurity (ENISA) ได้เสนอกรอบการดำเนินงานแบบหลายชั้นเพื่อส่งเสริมแนวปฏิบัติที่ดีด้านความมั่นคงปลอดภัยไซเบอร์สำหรับปัญญาประดิษฐ์ ซึ่งประกอบด้วย ๓ ชั้น

๒.๑) หลักการพื้นฐานความมั่นคงปลอดภัยไซเบอร์ หลักการพื้นฐานที่ต้องใช้กับทุกสภาพแวดล้อมของระบบเทคโนโลยีสารสนเทศและการสื่อสารที่มีปัญญาประดิษฐ์

๒.๒) ความมั่นคงปลอดภัยที่จำเพาะต่อปัญญาประดิษฐ์ แนวปฏิบัติที่จัดการกับความท้าทายเฉพาะของปัญญาประดิษฐ์

๒.๓) ความมั่นคงปลอดภัยปัญญาประดิษฐ์ในแต่ละภาคส่วน แนวปฏิบัติเพิ่มเติมสำหรับภาคส่วนที่มีความเสี่ยงสูง เช่น การเงิน สาธารณสุข

๒.๘ กฎหมาย กฎระเบียบ และแนวปฏิบัติที่เกี่ยวข้องในประเทศไทย

การประยุกต์ใช้ปัญญาประดิษฐ์ในประเทศไทยต้องอยู่ภายใต้กรอบของกฎหมาย กฎระเบียบ และแนวปฏิบัติที่เกี่ยวข้อง ดังนี้

๑) พระราชบัญญัติการรักษาความมั่นคงปลอดภัยไซเบอร์ พ.ศ. ๒๕๖๒

กฎหมายฉบับนี้เป็นกฎหมายหลักในการกำกับดูแลความมั่นคงปลอดภัยไซเบอร์ของประเทศ หากนำระบบปัญญาประดิษฐ์ไปใช้ในหน่วยงานที่เป็นโครงสร้างพื้นฐานสำคัญทางสารสนเทศ (Critical Information Infrastructure - CII) ระบบดังกล่าวจะต้องผ่านมาตรฐานและมาตรการด้านความปลอดภัยตามที่ สกมช. กำหนด

๒) พระราชบัญญัติคุ้มครองข้อมูลส่วนบุคคล พ.ศ. ๒๕๖๒ (Personal Data Protection Act: PDPA)

กฎหมายฉบับนี้มีความสำคัญอย่างยิ่งต่อการพัฒนาและใช้งานปัญญาประดิษฐ์ เนื่องจากระบบปัญญาประดิษฐ์ โดยเฉพาะ Machine Learning ต้องใช้ข้อมูลจำนวนมากในการฝึกสอน ประเด็นที่ต้องพิจารณาอย่างเคร่งครัด ได้แก่

๒.๑) ฐานทางกฎหมายในการประมวลผล การรวบรวมและใช้ข้อมูลส่วนบุคคลเพื่อฝึกสอนปัญญาประดิษฐ์ ต้องมีฐานทางกฎหมายรองรับ ซึ่งส่วนใหญ่มักจะเป็น "ความยินยอม" จากเจ้าของข้อมูล

๒.๒) สิทธิ์ของเจ้าของข้อมูล องค์กรต้องมีช่องทางให้เจ้าของข้อมูลสามารถใช้สิทธิ์ของตนได้ เช่น สิทธิ์ในการเพิกถอนความยินยอม หรือสิทธิ์ในการคัดค้านการประมวลผล

๒.๓) มาตรการรักษาความมั่นคงปลอดภัยของข้อมูล องค์กรมีหน้าที่ต้องจัดให้มีมาตรการรักษาความมั่นคงปลอดภัยของข้อมูลส่วนบุคคลที่นำมาใช้กับปัญญาประดิษฐ์อย่างเหมาะสม เพื่อป้องกันการรั่วไหลหรือการเข้าถึงโดยไม่ได้รับอนุญาต

๓) แนวทางการประยุกต์ใช้ปัญญาประดิษฐ์อย่างมีธรรมาภิบาล สำหรับผู้บริหารองค์กร (AI Governance Guideline For Executives)

เอกสาร "แนวทางการประยุกต์ใช้ปัญญาประดิษฐ์อย่างมีธรรมาภิบาล สำหรับผู้บริหารองค์กร" ของ สำนักงานพัฒนาธุรกรรมทางอิเล็กทรอนิกส์ (สพธอ.) ถือเป็นกรอบธรรมาภิบาลปัญญาประดิษฐ์ระดับประเทศ ซึ่งครอบคลุมหลักการสำคัญหลายมิติ ไม่ว่าจะเป็นการกำกับดูแล ความโปร่งใส ความเป็นธรรม และยังได้รวมหลักการด้านความน่าเชื่อถือ ความมั่นคงปลอดภัย และความเป็นส่วนตัว ไว้เป็นหนึ่งในแกนหลักด้วย

ตารางที่ ๔ สรุปมาตรฐานและกรอบการดำเนินงานด้านความมั่นคงปลอดภัยปัญญาประดิษฐ์ ที่สำคัญจากหน่วยงานชั้นนำทั้งในและต่างประเทศ ซึ่งถูกใช้เป็นเอกสารอ้างอิงหลักในการพัฒนาแนวปฏิบัติฉบับนี้

ตารางที่ ๔ มาตรฐานและกรอบการดำเนินงานด้านความมั่นคงปลอดภัยปัญญาประดิษฐ์

ลำดับ	กลุ่มเอกสาร	ชื่อเอกสาร	หน่วยงาน	การนำมาปรับใช้ในแนวปฏิบัติ	
A	กรอบธรรมาภิบาลและกฎหมายของประเทศไทย				
A.๑		แนวทางการประยุกต์ใช้ Generative AI อย่างมีธรรมาภิบาลสำหรับองค์กร	สำนักงานพัฒนาธุรกรรมทางอิเล็กทรอนิกส์ (สพธอ.)	ให้แนวทางด้านธรรมาภิบาลสำหรับ Generative AI โดยเฉพาะ ครอบคลุมความเสี่ยงด้านเนื้อหาที่ไม่ถูกต้อง ความเป็นส่วนตัว และทรัพย์สินทางปัญญา	ใช้เป็นกรอบธรรมาภิบาลหลักของไทย ที่แนวปฏิบัติฉบับนี้จะเข้าไป "เติมเต็มในมิติความมั่นคงปลอดภัยไซเบอร์" โดยนำความเสี่ยงที่ระบุไว้มาต่อยอดเพื่อกำหนดมาตรการป้องกันทางเทคนิค
A.๒		แนวทางการประยุกต์ใช้ปัญญาประดิษฐ์อย่างมีธรรมาภิบาล สำหรับผู้บริหารองค์กร	สำนักงานพัฒนาธุรกรรมทางอิเล็กทรอนิกส์ (สพธอ.)	วางกรอบการกำกับดูแลปัญญาประดิษฐ์ในระดับองค์กร ๓ ด้าน ๑) โครงสร้างการกำกับดูแล ๒) กลยุทธ์ และ ๓) การปฏิบัติงาน	นำโครงสร้าง ๓ ด้านมาเป็นแกนหลักในการจัดทำแนวทางการกำกับดูแลปัญญาประดิษฐ์ เพื่อให้แนวปฏิบัติสอดคล้องกับแนวทางระดับชาติ และช่วยให้องค์กรนำไปปรับใช้ได้ง่าย
A.๓	พ.ร.บ. การรักษาความมั่นคงปลอดภัยไซเบอร์ พ.ศ. ๒๕๖๒		สำนักงานคณะกรรมการการรักษาความมั่นคงปลอดภัยไซเบอร์แห่งชาติ (สกมช.)	กำหนดกรอบการรักษาความมั่นคงปลอดภัยไซเบอร์ระดับชาติ โดยเฉพาะสำหรับหน่วยงานโครงสร้างพื้นฐานสำคัญทางสารสนเทศ	ใช้อ้างอิงเป็นข้อบังคับทางกฎหมาย เพื่อกำหนดระดับความเข้มข้นของมาตรการสำหรับระบบปัญญาประดิษฐ์ ที่ใช้งานในภาคส่วนที่มีความสำคัญอย่างยิ่งยวด

ลำดับ	กลุ่มเอกสาร	ชื่อเอกสาร	หน่วยงาน	การนำมาปรับใช้ในแนวปฏิบัติ	
A.๔	พ.ร.บ. คุ้มครองข้อมูลส่วนบุคคล พ.ศ. ๒๕๖๒		สำนักงานคณะกรรมการคุ้มครองข้อมูลส่วนบุคคล (สคส.)	กำหนดหลักเกณฑ์และเงื่อนไขในการเก็บรวบรวม ใช้ หรือเปิดเผยข้อมูลส่วนบุคคล รวมถึงมาตรการรักษาความปลอดภัยของข้อมูล	ใช้เป็นข้อกำหนดพื้นฐานในการจัดการข้อมูลตลอดวงจรชีวิตของระบบปัญญาประดิษฐ์ โดยเฉพาะการออกแบบเรื่องความยินยอม และการป้องกันข้อมูลรั่วไหล
B	มาตรฐานและกรอบการดำเนินงานสากล				
B.๑		ISO/IEC 42001:2023 (AI — Management system)	ISO/IEC	เป็นมาตรฐานสากลฉบับแรกสำหรับระบบการจัดการปัญญาประดิษฐ์ กำหนดกรอบการดำเนินงานที่รับผิดชอบและมีธรรมาภิบาล	นำมาใช้เป็นกรอบอ้างอิงหลักในการกำกับดูแลปัญญาประดิษฐ์ เพื่อแนะนำองค์กรในการสร้างระบบการจัดการ ปัญญาประดิษฐ์ที่ครบวงจรและพร้อมสำหรับการตรวจสอบ
B.๒		ISO/IEC 23894:2023 (Guidance on AI Risk Management)	ISO/IEC	ให้คำแนะนำเชิงลึกเกี่ยวกับการบริหารจัดการความเสี่ยงที่เกี่ยวข้องกับปัญญาประดิษฐ์โดยเฉพาะ	ใช้เป็นแนวทางหลักในการพัฒนากระบวนการประเมินความเสี่ยงปัญญาประดิษฐ์ในระยะการออกแบบ
B.๓		ISO/IEC 27001:2022 (Information Security Management)	ISO/IEC	เป็นมาตรฐานหลักสำหรับการจัดตั้งและบริหารจัดการระบบความมั่นคงปลอดภัยของสารสนเทศ (ISMS) ในองค์กร	ใช้เป็นมาตรฐาน สำหรับการรักษาความปลอดภัยของโครงสร้างพื้นฐานที่รองรับระบบปัญญาประดิษฐ์

ลำดับ	กลุ่มเอกสาร	ชื่อเอกสาร	หน่วยงาน	การนำมาปรับใช้ในแนวปฏิบัติ	
B.๔		ISO/IEC 5338:2023 AI System Life Cycle Processes	ISO/IEC	ระบุกระบวนการต่าง ๆ ที่สนับสนุน การนิยาม การควบคุม การจัดการ การดำเนินการ และการปรับปรุงระบบปัญญาประดิษฐ์ ในแต่ละขั้นตอนของวงจรชีวิตของระบบปัญญาประดิษฐ์ กระบวนการเหล่านี้ยังสามารถนำไปใช้ภายในองค์กรหรือโครงการในระหว่างการพัฒนาหรือการจัดหาระบบปัญญาประดิษฐ์ได้เช่นกัน	
B.๕		Multilayer Framework for Good Cybersecurity Practices for AI	ENISA	นำเสนอกรอบการป้องกัน ๓ ชั้น ได้แก่ ๑) การรักษาความมั่นคงปลอดภัยไซเบอร์ขั้นพื้นฐาน ๒) การรักษาความมั่นคงปลอดภัยไซเบอร์ที่จำเพาะต่อปัญญาประดิษฐ์ ๓) การรักษาความมั่นคงปลอดภัยไซเบอร์ที่จำเพาะต่อแต่ละภาคส่วนสำหรับปัญญาประดิษฐ์	ใช้เพื่อเสริมแนวคิดการป้องกันเชิงลึกและเน้นย้ำความสำคัญของการปรับมาตรการให้เข้ากับบริบทของแต่ละอุตสาหกรรม

ลำดับ	กลุ่มเอกสาร	ชื่อเอกสาร	หน่วยงาน	การนำมาปรับใช้ในแนวปฏิบัติ	
B.๖		OWASP Top 10 for Large Language Model Applications	OWASP Foundation	ระบุ ๑๐ ความเสี่ยงด้านความมั่นคงปลอดภัยที่สำคัญที่สุดสำหรับแอปพลิเคชันที่ใช้ LLM เช่น Prompt Injection และ Insecure Output Handling	นำรายการความเสี่ยง ๑๐ อันดับแรกมาใช้เป็นกรณีศึกษาที่ชัดเจน เพื่อให้ผู้พัฒนาตระหนักถึงช่องโหว่ที่พบบ่อยและวิธีการป้องกัน
B.๗		OWASP AI Exchange	OWASP Foundation	เป็นส่วนหนึ่งของโครงการ OWASP AI Security and Privacy Guide ทำหน้าที่เป็นคลังทรัพยากรส่วนกลางที่ขับเคลื่อนโดยชุมชน เพื่อรวบรวมเครื่องมือ ชุดข้อมูล และองค์ความรู้สำหรับความมั่นคงปลอดภัยของปัญญาประดิษฐ์โดยเฉพาะ	การนำมาปรับใช้ในแนวปฏิบัติมีลักษณะดังนี้ ๑. เชื่อมโยงทฤษฎีสู่การปฏิบัติ ๒. สนับสนุนวงจรการพัฒนาที่ปลอดภัย ๓. สร้างมาตรฐานการทดสอบที่เป็นรูปธรรม

เนื้อหาที่นำเสนอในแนวปฏิบัติฉบับนี้มีการพัฒนาและเรียบเรียงให้เติมเต็มช่องว่างของกฎหมาย กฎระเบียบ และแนวปฏิบัติที่เกี่ยวกับการใช้ปัญญาประดิษฐ์อย่างมั่นคงปลอดภัยในบริบทของประเทศไทย โดยมีความสอดคล้องกับกฎหมาย กรอบการดำเนินงาน และมาตรฐานสากล โดยสามารถสรุปแนวทางการนำข้อมูลในแนวปฏิบัติฉบับนี้มาปรับใช้ได้ ดังนี้

๑) แนวปฏิบัติฉบับนี้จะปิดช่องว่างที่สำคัญโดยการเชื่อมโยง “ธรรมาภิบาล” จากแนวทางของสำนักงานพัฒนาธุรกรรมทางอิเล็กทรอนิกส์ (สพธอ.) เข้ากับ "ความมั่นคงปลอดภัยไซเบอร์" จากมาตรฐานสากล ทำให้เกิดเป็นกรอบการดำเนินงานที่ครบวงจรสำหรับประเทศไทย

๒) เสริมความแข็งแกร่งโดยการอ้างอิงมาตรฐานระดับโลก เช่น ISO/IEC และ ITU มาตรฐาน รวมถึงมาตรฐาน แนวปฏิบัติ และกรอบการดำเนินงานระดับภูมิภาค เช่น ENISA ซึ่งจะช่วยเพิ่มความน่าเชื่อถือและการยอมรับในระดับสากลให้กับแนวปฏิบัตินี้ และช่วยยกระดับขีดความสามารถในการแข่งขันขององค์กรไทย

๓) นำหลักการสำคัญมาใช้ เช่น การออกแบบอย่างมั่นคงปลอดภัย (Secure By Design) การป้องกันเชิงลึก (Defense-In-Depth) และกรอบการทำงานตลอดวงจรชีวิตของระบบปัญญาประดิษฐ์ จะถูกนำมาใช้เป็นแกนกลางของเอกสาร เพื่อสร้างแนวทางที่เป็นระบบปฏิบัติได้จริงและยั่งยืน

บทที่ ๓
กรอบการรักษาความมั่นคงปลอดภัยระบบปัญญาประดิษฐ์

ระบบปัญญาประดิษฐ์ถือเป็นระบบสารสนเทศระบบหนึ่ง การรักษาความมั่นคงปลอดภัยของระบบปัญญาประดิษฐ์จึงควรเริ่มต้นจากหลักการพื้นฐานในการรักษาความมั่นคงปลอดภัยของระบบสารสนเทศ ซึ่งมีหลักการพื้นฐานที่จำเป็น ได้แก่ หลักการออกแบบอย่างมั่นคงปลอดภัย หลักการตั้งค่าเริ่มต้นอย่างมั่นคงปลอดภัย และหลักการป้องกันเชิงลึก ในบริบทของระบบปัญญาประดิษฐ์ การรักษาความมั่นคงปลอดภัยเฉพาะโมเดลหรือแอปพลิเคชันที่ใช้งานนั้นไม่เพียงพอ ควรมีการรักษาความมั่นคงปลอดภัยที่ครอบคลุมตลอดวงจรชีวิตของระบบปัญญาประดิษฐ์ ดังนั้น บทนี้จึงแนะนำหลักการพื้นฐานในการรักษาความมั่นคงปลอดภัยและกรอบวงจรชีวิตของระบบปัญญาประดิษฐ์

๓.๑ หลักการพื้นฐานในการรักษาความมั่นคงปลอดภัย

๑) หลักการออกแบบอย่างมั่นคงปลอดภัย

ผู้พัฒนาระบบปัญญาประดิษฐ์ควรกำหนดให้ความมั่นคงปลอดภัยเป็นข้อกำหนดหลักตั้งแต่ขั้นตอนการออกแบบ ไม่ใช่สิ่งที่นำมาเพิ่มหลังจากได้มีการพัฒนาและใช้งานแล้ว

หลักการออกแบบอย่างมั่นคงปลอดภัย เป็นหลักการที่กำหนดให้มิติด้านความมั่นคงปลอดภัยไซเบอร์เป็นองค์ประกอบหลักและเป็นข้อกำหนดที่ต้องพิจารณา ตั้งแต่ระยะเริ่มต้นของวงจรการพัฒนาระบบ (System Development Lifecycle: SDLC) แทนที่จะเป็นเพียงส่วนเสริมที่ถูกนำมาพิจารณาในภายหลัง กระบวนทัศน์นี้เน้นการป้องกันเชิงรุก โดยการคาดการณ์และลดทอนความเสี่ยงที่อาจเกิดขึ้นในสถาปัตยกรรมของระบบ การลดความเสี่ยงผลกระทบการจากภัยคุกคามตั้งแต่ระยะการออกแบบเป็นสิ่งที่สำคัญ เพราะช่องโหว่ความมั่นคงปลอดภัยที่มีตั้งแต่ระยะการออกแบบจะส่งผลกระทบต่อเนื่องไปยังระยะต่าง ๆ ตลอดวงจรชีวิตของระบบปัญญาประดิษฐ์

การออกแบบระบบปัญญาประดิษฐ์อย่างมั่นคงปลอดภัย ควรมีการคำนึงถึงและดำเนินการอย่างน้อย ดังนี้

๑.๑) การสร้างแบบจำลองภัยคุกคามและการประเมินความเสี่ยง ควรประเมินภัยคุกคามที่มีต่อระบบปัญญาประดิษฐ์อย่างรอบด้านเพื่อระบุความเสี่ยงที่อาจเกิดขึ้นได้ ซึ่งรวมถึงภัยคุกคามที่มีต่อระบบสารสนเทศแบบดั้งเดิมทั่วไปและภัยคุกคามที่เฉพาะเจาะจงต่อระบบปัญญาประดิษฐ์ เช่น Data Poisoning, Model Inversion, Adversarial Attack และ Model Theft

๑.๒) กระบวนการรวบรวมและจัดเตรียมข้อมูล ควรออกแบบระบบให้มีการเก็บรวบรวม ใช้และจัดเก็บเฉพาะข้อมูลที่จำเป็นต่อการใช้งานเท่านั้น เพื่อลดพื้นที่ในการโจมตีและลดผลกระทบที่อาจเกิดขึ้นจากการละเมิดข้อมูล ควรมีการประยุกต์ใช้เทคโนโลยียกระดับความเป็นส่วนตัว (Privacy Enhancing Technologies: PETs) เช่น การทำข้อมูลนิรนาม การทำนามแฝงตั้งแต่ขั้นตอนการออกแบบเพื่อคุ้มครองข้อมูลส่วนบุคคล และการสร้างกลไกตรวจสอบความถูกต้องและแหล่งที่มาของข้อมูลเพื่อลดความเสี่ยงจากการโจมตีแบบ Data Poisoning

๑.๓) การพัฒนาและฝึกสอนโมเดล การคัดเลือกสถาปัตยกรรมโมเดลที่มีคุณสมบัติทนทานต่อการโจมตีเชิงปรปักษ์ และการบูรณาการเทคนิคการฝึกสอนเชิงปรปักษ์ เข้าเป็นส่วนหนึ่งของกระบวนการพัฒนา เพื่อสร้างภูมิต้านทานให้โมเดลตั้งแต่ต้น

๒) หลักการตั้งค่าเริ่มต้นอย่างปลอดภัย

องค์กร ผู้พัฒนา และผู้ให้บริการระบบปัญญาประดิษฐ์ต้องรับผิดชอบต่อความมั่นคงปลอดภัยของลูกค้า และตั้งค่าเริ่มต้นของระบบให้มั่นคงปลอดภัยที่สุด

หลักการตั้งค่าเริ่มต้นอย่างปลอดภัย มุ่งเน้นการตั้งค่าเริ่มต้นของระบบให้มีระดับความมั่นคงปลอดภัยสูงสุด เพื่อลดพื้นที่การโจมตี และลดความเสี่ยงและผลกระทบที่อาจเกิดขึ้นจากภัยคุกคาม ผู้ใช้งานสามารถปรับแก้การตั้งค่าการรักษาความมั่นคงปลอดภัยให้มีความเหมาะสมกับความจำเป็นในการใช้งานระบบปัญญาประดิษฐ์อย่างมั่นคงปลอดภัยและมีประสิทธิภาพ

การตั้งค่าเริ่มต้นระบบปัญญาประดิษฐ์อย่างมั่นคงปลอดภัย ควรมีการคำนึงถึงและดำเนินการอย่างน้อย ดังนี้

๒.๑) การควบคุมการเข้าถึง (Access Control) โดยใช้หลักการสิทธิ์เข้าถึงที่น้อยที่สุด อนุญาตสิทธิ์ กับผู้ที่ได้รับอนุญาตเท่าที่จำเป็น ที่สามารถดำเนินการในการเข้าถึงและใช้งานข้อมูลและระบบปัญญาประดิษฐ์

๒.๒) การพิสูจน์ตัวตน (Authentication) ควรมีการใช้งานการพิสูจน์แบบหลายปัจจัยสำหรับบัญชีผู้ดูแลระบบที่มีสิทธิ์ในการบริหารจัดการระบบปัญญาประดิษฐ์ เพื่อลดความเสี่ยงจากการเข้าถึงระบบโดยไม่ได้รับอนุญาตจากผู้ไม่ประสงค์ดี ซึ่งอาจมีการโจรกรรมข้อมูลพิสูจน์ตัวตน เช่น รหัสผ่าน

๒.๓) การตรวจสอบและแก้ไขข้อมูลนำเข้า (Input Sanitization) ควรมีการตรวจสอบข้อมูลนำเข้าสู่ระบบและแก้ไขข้อมูลนำเข้าให้อยู่ในรูปแบบและขอบเขตที่ระบบใช้งานได้อย่างมั่นคงปลอดภัย เพื่อป้องกันการโจมตี เช่น Prompt Injection และ Data Manipulation

๒.๔) กระบวนการติดตั้งและใช้งาน การกำหนดค่าเริ่มต้นของส่วนต่อประสานโปรแกรมประยุกต์ (Application Programming Interface: API) ให้บังคับใช้กลไกการพิสูจน์ตัวตนและ

การอนุญาตสิทธิ์อย่างเข้มงวด และเปิดใช้งานกลไกจำกัดอัตราการร้องขอ (Rate Limiting) เพื่อป้องกันการโจมตีแบบ Model Extraction

๒.๕) การบันทึกล็อกและการเฝ้าระวัง ควรมีการจัดเก็บข้อมูลล็อก ข้อมูลนำเข้า ข้อมูลนำออกของโมเดล และการดำเนินการด้วยสิทธิ์ผู้ดูแลระบบ เพื่อเป็นข้อมูลในการตรวจสอบร่องรอยภัยคุกคามและการโจมตีที่อาจเกิดขึ้น เช่น Data Exfiltration และ Adversarial Attack

๓) หลักการป้องกันเชิงลึก

หลักการป้องกันเชิงลึก เป็นกลยุทธ์ด้านความมั่นคงปลอดภัยที่ตั้งอยู่บนสมมติฐานว่าจะไม่มีมาตรการป้องกันใดเพียงมาตรการเดียวที่สามารถป้องกันภัยคุกคามได้สมบูรณ์ จึงใช้การวางมาตรการควบคุมความปลอดภัยที่หลากหลายและทำงานส่งเสริมกัน ในลักษณะหลายชั้น เพื่อสร้างอุปสรรคหลายชั้นต่อผู้บุกรุก หากมาตรการป้องกันในชั้นใดชั้นหนึ่งล้มเหลว มาตรการในชั้นถัดไปจะยังคงทำหน้าที่ชะลอหรือยับยั้งการโจมตีได้

กลยุทธ์นี้ถูกนำมาใช้โดยการวางกลไกควบคุมในส่วนประกอบต่าง ๆ ของระบบนิเวศปัญญาประดิษฐ์ ดังนี้

๓.๑) ระดับเครือข่าย การใช้ Web Application Firewall (WAF) เพื่อคัดกรองการรับส่งข้อมูลที่มุ่งร้ายต่อ API Endpoint และการแบ่งส่วนเครือข่ายเพื่อจำกัดขอบเขตความเสียหาย

๓.๒) ระดับแอปพลิเคชัน การบังคับใช้นโยบายการพิสูจน์ตัวตนและการอนุญาตสิทธิ์ที่เข้มงวด และการใช้เทคนิค Input Sanitization เพื่อตรวจสอบและขจัดข้อมูลนำเข้าที่อาจเป็นอันตรายก่อนส่งไปยังโมเดล เพื่อป้องกันการโจมตีแบบ Prompt Injection

๓.๓) ระดับโมเดล การใช้โมเดลที่ผ่านกระบวนการ Adversarial Training ควบคู่กับการติดตั้งกลไกตรวจสอบผลลัพธ์ หรือ Guardrails เพื่อคัดกรองเนื้อหาที่ไม่เหมาะสมหรือเป็นอันตราย

๓.๔) ระดับข้อมูล การบังคับใช้นโยบายการเข้ารหัสลับข้อมูล ทั้งในสภาวะพัก และระหว่างการส่งผ่าน และการควบคุมการเข้าถึงชุดข้อมูลฝึกสอนอย่างรัดกุม

๓.๕) ระดับการปฏิบัติการและการเฝ้าระวัง การจัดให้มีระบบบันทึกเหตุการณ์และการเฝ้าระวังความผิดปกติเพื่อตรวจจับพฤติกรรมการใช้งานที่น่าสงสัย ซึ่งอาจบ่งชี้ถึงความพยายามในการโจมตีระบบ

๓.๒ กรอบวงจรชีวิตของระบบปัญญาประดิษฐ์

แนวปฏิบัตินี้ใช้กรอบวงจรชีวิตของระบบปัญญาประดิษฐ์เป็นโครงหลักในการรักษาความมั่นคงปลอดภัยไซเบอร์ของระบบปัญญาประดิษฐ์ กรอบวงจรชีวิตของระบบปัญญาประดิษฐ์นี้ได้ปรับปรุงมาจากมาตรฐาน ISO/IEC 22989:2022 โดยการปรับกรอบการทำงาน จากมุมมองเชิงหน้าที่ของ ISO ซึ่งอธิบายว่า “ต้องทำอะไรบ้าง” ไปสู่มุมมองเชิงบูรณาการความมั่นคงปลอดภัย ที่เน้นย้ำว่า “ต้องทำ

อย่างไรให้มั่นคงปลอดภัย” ในแต่ละขั้นตอน กรอบวงจรชีวิตของระบบปัญญาประดิษฐ์ แนวปฏิบัตินี้จึงถือเป็นแนวปฏิบัติเพื่อการนำไปใช้จริงที่มีความจำเพาะเจาะจงด้านความมั่นคงปลอดภัยไซเบอร์ซึ่งสอดคล้องกับหลักการพื้นฐานในการรักษาความมั่นคงปลอดภัย และแนวคิด Shift-Left Security สมัยใหม่

เหตุผลสนับสนุนการนำเสนอขั้นตอนที่แตกต่างจากมาตรฐาน ISO/IEC 22989:2022 มี ดังนี้

๑) การเปลี่ยนกระบวนทัศน์จากกรอบการทำงานเชิงหน้าที่สู่กรอบการทำงานที่เน้นความมั่นคงปลอดภัย

๑.๑) ISO/IEC 22989:2022 มีวัตถุประสงค์เพื่อสร้างคำจำกัดความและกรอบอ้างอิงที่เป็นกลาง สำหรับวงจรชีวิตของระบบปัญญาประดิษฐ์ โดยเน้นการอธิบายขั้นตอนการทำงานเชิงหน้าที่และเชิงตรรกะ ตั้งแต่การเกิดแนวคิดไปจนถึงการสิ้นสุดการใช้งาน ซึ่งเป็นประโยชน์อย่างยิ่งในเชิงของการกำกับดูแล และการบริหารจัดการโครงการ

๑.๒) กรอบวงจรชีวิตของระบบปัญญาประดิษฐ์ที่นำเสนอ มีวัตถุประสงค์เพื่อเป็นแนวปฏิบัติที่มุ่งเน้นการลงมือทำ โดยการบูรณาการข้อพิจารณาด้านความมั่นคงปลอดภัยไซเบอร์เข้าไปเป็นส่วนหนึ่งของทุกขั้นตอนอย่างชัดเจน การเติมคำว่า "อย่างมั่นคงปลอดภัย" ในทุกระยะเป็นการส่งสัญญาณที่ชัดเจนว่าความมั่นคงปลอดภัยไม่ใช่กิจกรรมเสริม แต่เป็นส่วนหนึ่งของกระบวนการหลัก

๒) การเน้นย้ำความมั่นคงปลอดภัยเชิงรุก

กรอบวงจรชีวิตของระบบปัญญาประดิษฐ์ ที่นำเสนอได้แยกขั้นตอน “Design and Development” ของ ISO ออกเป็น “ระยะที่ ๑ การออกแบบอย่างมั่นคงปลอดภัย” และ “ระยะที่ ๒ การพัฒนาอย่างมั่นคงปลอดภัย” ซึ่งมีนัยสำคัญทางวิชาการ คือ

๒.๑) การออกแบบอย่างมั่นคงปลอดภัย เน้นกิจกรรมความมั่นคงปลอดภัยในระดับสถาปัตยกรรม เช่น การทำแบบจำลองภัยคุกคามการกำหนดนโยบายการกำกับดูแลข้อมูล และการเลือกใช้สถาปัตยกรรมโมเดลที่ทนทานต่อการโจมตี

๒.๒) การพัฒนาอย่างมั่นคงปลอดภัย เน้นกิจกรรมความมั่นคงปลอดภัยในระดับการลงมือปฏิบัติ เช่น การเขียนโคดที่ปลอดภัย การตรวจสอบช่องโหว่ของไลบรารีที่ใช้ และการใช้เทคนิค Adversarial Training ในระหว่างการฝึกสอนโมเดล

การแยกสองส่วนนี้ออกจากกันสอดคล้องกับหลักการ “Shift-left” ในงานด้านความมั่นคงปลอดภัยไซเบอร์ ที่มุ่งเน้นการค้นหาและแก้ไขปัญหาความมั่นคงปลอดภัยให้เร็วที่สุดเท่าที่จะทำได้ในวงจรการพัฒนา เพื่อลดต้นทุนและความเสียหายในระยะยาว

๓) ความสอดคล้องกับกรอบการพัฒนาที่มั่นคงปลอดภัยสมัยใหม่

โครงสร้างของกรอบที่นำเสนอมีความสอดคล้องอย่างยิ่งกับกรอบการดำเนินงานด้านการพัฒนาซอฟต์แวร์อย่างมั่นคงปลอดภัยที่เป็นที่ยอมรับในระดับสากล เช่น OWASP Software Assurance

Maturity Model (SAMM) ซึ่งสนับสนุนและผลักดันให้บูรณาการกิจกรรมความมั่นคงปลอดภัยเข้าไปในทุกขั้นตอนของ SDLC กรอบวงจรชีวิตของระบบปัญญาประดิษฐ์นี้จึงเป็นการนำกรอบการทำงานเหล่านั้นมาปรับใช้ให้จำเพาะเจาะจงกับบริบทของปัญญาประดิษฐ์

๔) การสร้างความชัดเจนและเพิ่มการนำไปปฏิบัติได้จริง

การใช้ชื่อขั้นตอนที่ชัดเจน เช่น การทวนสอบด้านความมั่นคงปลอดภัย และการนำไปใช้งานอย่างมั่นคงปลอดภัย ช่วยสร้างความชัดเจนให้กับทีมพัฒนา ทีมความมั่นคงปลอดภัย และผู้ตรวจสอบว่าในแต่ละระยะมีกิจกรรมด้านความมั่นคงปลอดภัยที่ต้องดำเนินการและทวนสอบ ซึ่งทำให้แนวทางนี้สามารถนำไปปรับใช้เป็นรายการตรวจสอบหรือ เกณฑ์การประเมินได้ง่ายกว่ากรอบการทำงานของ ISO ที่มีความเป็นนามธรรมสูงกว่า

การเปรียบเทียบและการเชื่อมโยงแนวคิดระหว่าง ISO/IEC 22989:2022 และ กรอบวงจรชีวิตของระบบปัญญาประดิษฐ์ที่นำเสนอแสดงใน ตารางที่ ๕

ตารางที่ ๕ การเปรียบเทียบและการเชื่อมโยงแนวคิด

ISO/IEC 22989:2022	กรอบวงจรชีวิตของระบบปัญญาประดิษฐ์ ที่นำเสนอ	คำอธิบายการเชื่อมโยง
๑. Inception	**ระยะที่ ๐** ขั้นตอนแนวคิด (Concept)	เป็นขั้นตอนเดียวกัน คือการระบุความต้องการและเป้าหมาย
๒. Design and Development	**ระยะที่ ๑** การออกแบบอย่างมั่นคงปลอดภัย (Secure Design)	แยกขั้นตอนเดียวของ ISO เพื่อเน้นย้ำกิจกรรมความปลอดภัยเชิงรุก (Shift-Left)
	ระยะที่ ๒ การพัฒนาอย่างมั่นคงปลอดภัย (Secure Development)	
๓. Verification and Validation	**ระยะที่ ๓** การทวนสอบด้านความมั่นคงปลอดภัย (Secure Verification)	เป็นขั้นตอนเดียวกัน โดยเน้นการทดสอบด้านความมั่นคงปลอดภัยโดยเฉพาะ
๔. Deployment	**ระยะที่ ๔** การนำไปใช้งานอย่างมั่นคงปลอดภัย (Secure Deployment)	เป็นขั้นตอนเดียวกัน โดยเน้นการตั้งค่าที่ปลอดภัยและการส่งมอบที่รัดกุม
๕. Operation and Monitoring	**ระยะที่ ๕** การดำเนินงานและการบำรุงรักษาอย่างมั่นคงปลอดภัย	รวม ๒ ขั้นตอนของ ISO เข้าด้วยกัน เนื่องจากในทางปฏิบัติ การเฝ้าระวังและ

ISO/IEC 22989:2022	กรอบวงจรชีวิตของระบบปัญญาประดิษฐ์ ที่นำเสนอ	คำอธิบายการเชื่อมโยง
๖. Re-Evaluation	(Secure Operation & Maintenance)	ประเมินซ้ำเป็นส่วนหนึ่งของการดำเนินงานและบำรุงรักษาอย่างต่อเนื่อง
๗. Retirement	**ระยะที่ ๖** การกำจัดและทำลาย (Disposal)	เป็นขั้นตอนเดียวกัน โดยใช้คำว่า "Disposal" เพื่อเน้นย้ำถึงการทำลายข้อมูลอย่างปลอดภัย

เอกสารอ้างอิง:

๑) ISO/IEC 22989:2022 Information technology — Artificial intelligence — Artificial intelligence concepts and terminology

๒) ISO/IEC 27002:2022 8.27 Secure system architecture and engineering principles

๓) ISO/IEC 42001:2023 A.6 AI system life cycle

๔) OWASP-AI-EXCHANGE 3. Development-time threats

กรอบวงจรชีวิตของระบบปัญญาประดิษฐ์ประกอบด้วยระยะที่ ๐ ถึง ระยะที่ ๖ ดังต่อไปนี้

๓.๒.๑ ระยะที่ ๐ ขั้นตอนแนวคิด

ขั้นตอนแนวคิดถือเป็นระยะที่มีความสำคัญสูงสุดในวงจรชีวิตของระบบปัญญาประดิษฐ์ที่มั่นคงปลอดภัย เนื่องจากเป็นขั้นตอนของการวางรากฐานด้านความมั่นคงปลอดภัยและการกำกับดูแลก่อนที่จะมีการลงทุนหรือลงมือพัฒนาใด ๆ วัตถุประสงค์หลักของระยะนี้คือ ทำความเข้าใจบริบททางธุรกิจ กรอบทางกฎหมายที่บังคับใช้ ระบุสินทรัพย์ที่สำคัญ และประเมินความเสี่ยงที่อาจเกิดขึ้นกับระบบปัญญาประดิษฐ์อย่างเป็นระบบและครอบคลุม โดยแนะนำให้พิจารณาดำเนินการอย่างน้อย ดังต่อไปนี้

๑) การพิจารณาบริบททางกฎหมายและข้อบังคับ

ก่อนการประเมินความเสี่ยงทางเทคนิค องค์กรต้องระบุและทำความเข้าใจภูมิทัศน์ทางกฎหมายและข้อบังคับที่ระบบปัญญาประดิษฐ์จะต้องเข้าไปดำเนินการ ซึ่งข้อกำหนดเหล่านี้ถือเป็นความเสี่ยงด้านการปฏิบัติตามข้อกำหนด และเป็นปัจจัยสำคัญในการออกแบบระบบ

๑.๑) กฎหมายและกฎระเบียบภายในประเทศไทย

๑.๑.๑) พระราชบัญญัติการรักษาความมั่นคงปลอดภัยไซเบอร์ พ.ศ. ๒๕๖๒

หากระบบปัญญาประดิษฐ์ถูกนำไปใช้ในหน่วยงานที่เป็นโครงสร้างพื้นฐานสำคัญทางสารสนเทศ องค์กรจะต้องปฏิบัติตามมาตรฐานขั้นต่ำและกระบวนการที่กำหนดอย่างเข้มงวด

๑.๑.๒) กฎระเบียบเฉพาะทาง พิจารณากฎระเบียบของหน่วยงานกำกับดูแลในแต่ละอุตสาหกรรม เช่น ธนาคารแห่งประเทศไทยสำหรับ FinTech หรือสำนักงานคณะกรรมการกำกับหลักทรัพย์และตลาดหลักทรัพย์ (ก.ล.ต.) สำหรับการใช้ปัญญาประดิษฐ์ในการซื้อขายสินทรัพย์ดิจิทัล

๑.๒) ข้อกำหนดของคู่สัญญาและกฎหมายระหว่างประเทศ

๑.๒.๑) ข้อตกลงในสัญญา สัญญาทางธุรกิจกับคู่ค้าอาจมีข้อกำหนดเฉพาะเกี่ยวกับระดับความมั่นคงปลอดภัย การจัดการข้อมูล หรือมาตรฐานด้านจริยธรรมของปัญญาประดิษฐ์ที่องค์กรต้องปฏิบัติตาม

๑.๒.๒) General Data Protection Regulation (GDPR) หากระบบปัญญาประดิษฐ์มีการประมวลผลข้อมูลส่วนบุคคลของพลเมืองในสหภาพยุโรป หรือมีคู่สัญญาที่ดำเนินธุรกิจในสหภาพยุโรป องค์กรจำเป็นต้องปฏิบัติตามข้อกำหนดของ GDPR ซึ่งมีความเข้มงวดสูง

๑.๒.๓) EU AI Act แม้จะเป็นกฎหมายของ EU แต่มีผลกระทบข้ามพรมแดนองค์กรควรพิจารณาแนวทางการจำแนกความเสี่ยง ดังแสดงในหัวข้อ ๒.๓ เพื่อเตรียมความพร้อมและสร้างความได้เปรียบในการแข่งขัน

๒) การทำความเข้าใจองค์ประกอบของความเสี่ยง สินทรัพย์ ภัยคุกคาม และช่องโหว่

ก่อนที่จะทำการประเมินความเสี่ยง สิ่งสำคัญคือต้องเข้าใจองค์ประกอบพื้นฐานของความเสี่ยงในบริบทของความมั่นคงปลอดภัยปัญญาประดิษฐ์ ซึ่งประกอบด้วย

๒.๑) สินทรัพย์ คือสิ่งใดก็ตามที่มีคุณค่าต่อองค์กรซึ่งเกี่ยวข้องกับระบบปัญญาประดิษฐ์สินทรัพย์เหล่านี้อาจเป็นได้ทั้งสิ่งที่จับต้องได้และจับต้องไม่ได้ ตัวอย่างเช่น

๒.๑.๑) โมเดลปัญญาประดิษฐ์ที่ฝึกสอนแล้ว ถือเป็นทรัพย์สินทางปัญญาที่มีมูลค่าสูงอาจเป็นเป้าหมายของการโจมตี Model Extraction

๒.๑.๒) ชุดข้อมูลฝึกสอน โดยเฉพาะอย่างยิ่งหากมีข้อมูลที่ละเอียดอ่อน เช่น ข้อมูลส่วนบุคคล ข้อมูลทางการเงิน หรือความลับทางการค้า การรั่วไหลของข้อมูลนี้อาจนำไปสู่การละเมิดความเป็นส่วนตัวอย่างรุนแรง

๒.๑.๓) ชื่อเสียงและความน่าเชื่อถือขององค์กร การที่ระบบปัญญาประดิษฐ์ตัดสินใจผิดพลาดอย่างร้ายแรง มีอคติ หรือสร้างเนื้อหาที่เป็นอันตราย อาจทำลายความไว้วางใจของลูกค้าและสาธารณชน

๒.๒) ภัยคุกคาม คือเหตุการณ์หรือผู้กระทำที่มีศักยภาพในการสร้างความเสียหายต่อสินทรัพย์

๒.๒.๑) ผู้ไม่หวังดีภายนอก แฮกเกอร์ที่พยายามโจมตีระบบเพื่อผลประโยชน์ทางการเงินหรือเพื่อสร้างชื่อเสียง

๒.๒.๒) การโจมตีเชิงปรปักษ์ เป็นภัยคุกคามเชิงเทคนิคที่หลอกลวงหรือบิดเบือนการทำงานของโมเดลปัญญาประดิษฐ์โดยเฉพาะ

๒.๒.๓) ผู้ใช้งานภายในที่ขาดความตระหนัก พนักงานที่อาจสร้าง Prompt ที่นำไปสู่การเปิดเผยข้อมูลที่ละเอียดอ่อนโดยไม่ได้ตั้งใจ

๒.๓) ช่องโหว่ คือจุดอ่อนหรือข้อบกพร่องในระบบ กระบวนการ หรือมาตรการควบคุมที่เปิดโอกาสให้ภัยคุกคามสามารถสร้างความเสียหายต่อสินทรัพย์ได้สำเร็จ ตัวอย่างในบริบทปัญญาประดิษฐ์ เช่น

๒.๓.๑) การขาดการฝึกสอนโมเดลให้ทนทาน ทำให้โมเดลไม่มีภูมิต้านทานและอ่อนไหวต่อการโจมตีแบบ Evasion Attacks

๒.๓.๒) การขาดการตรวจสอบและคัดกรองข้อมูลนำเข้า เปิดช่องให้เกิดการโจมตีแบบ Prompt Injection ในระบบ LLM

๒.๓.๓) API ที่ไม่มีการป้องกัน การเปิดเผย API ของโมเดลโดยไม่มีการพิสูจน์ตัวตนหรือการจำกัดอัตราการเรียกใช้ทำให้ง่ายต่อการโจมตีแบบ Model Extraction

๓) กระบวนการประเมินความเสี่ยงสำหรับระบบปัญญาประดิษฐ์

กิจกรรมหลักในระยะนี้คือการประเมินความเสี่ยงอย่างเป็นระบบ เพื่อระบุและทำความเข้าใจ ภูมิทัศน์ของภัยคุกคามที่จำเพาะเจาะจงกับระบบปัญญาประดิษฐ์ที่กำลังจะพัฒนา โดยอ้างอิงกรอบการดำเนินงานจาก ISO/IEC 23894:2023 ซึ่งประกอบด้วยขั้นตอนดังนี้

๓.๑) การระบุความเสี่ยง

เป็นกระบวนการเชิงรุกในการจำแนกและจัดทำรายการภัยคุกคามและช่องโหว่ที่อาจส่งผลกระทบต่อสินทรัพย์ของระบบปัญญาประดิษฐ์ โดยพิจารณาจากแหล่งข้อมูลต่าง ๆ เช่น ISO/IEC 23894:2023 Annex B และ OWASP AI Security Top 10 ตัวอย่างระบบปัญญาประดิษฐ์ประเมินสินเชื่อ

๓.๑.๑) สินทรัพย์ ความถูกต้องของการอนุมัติสินเชื่อ ข้อมูลทางการเงินของผู้สมัคร ชื่อเสียงของสถาบันการเงิน

๓.๑.๒) การระบุความเสี่ยง ผู้ไม่หวังดีอาจใช้เทคนิค Data Poisoning โดยการแทรกข้อมูลผู้สมัครปลอมที่มีลักษณะดีแต่แฝงความสัมพันธ์ที่ผิดปกติเข้าไปในชุดข้อมูลฝึกสอน (อ้างอิง OWASP-AI-EXCHANGE 3. Development-Time Threats) เพื่อสร้าง Backdoor ทำให้โมเดลมีแนวโน้มที่จะอนุมัติสินเชื่อให้กับผู้สมัครที่มีลักษณะคล้ายกันในอนาคต แม้ว่าจะมีคุณสมบัติไม่เพียงพอก็ตาม

๓.๒) การวิเคราะห์ความเสี่ยง

๓.๒.๑) เป็นขั้นตอนการประเมินโอกาสการเกิด และผลกระทบ ของความเสี่ยงที่ระบุไว้เพื่อทำความเข้าใจระดับความรุนแรงของแต่ละความเสี่ยง โดยมีตัวอย่าง ดังนี้

๑) การวิเคราะห์โอกาสเกิด อาจอยู่ในระดับ “ปานกลาง” เนื่องจากการโจมตีประเภทนี้ต้องอาศัยการเข้าถึงหรือมีอิทธิพลต่อกระบวนการรวบรวมข้อมูลซึ่งอาจทำได้ยาก แต่หากทำสำเร็จจะตรวจจับได้ยากมาก

๒) การวิเคราะห์ผลกระทบ อยู่ในระดับ “สูง” เนื่องจากหากการโจมตีสำเร็จจะนำไปสู่ความเสียหายทางการเงินโดยตรง (หนี้เสีย) ความเสียหายจากการถูกลงโทษตามกฎหมาย เช่น ค่าปรับจากกฎหมายที่เกี่ยวข้อง การละเมิดกฎเกณฑ์ของหน่วยงานกำกับดูแล และการสูญเสียความน่าเชื่อถือขององค์กร (อ้างอิง ISO/IEC 42001:2023 A.5 Assessing impacts of AI systems)

๓.๓) การประเมินระดับความเสี่ยง

เป็นขั้นตอนสุดท้ายในการเปรียบเทียบผลลัพธ์จากการวิเคราะห์ความเสี่ยงกับเกณฑ์ ความเสี่ยงที่องค์กรยอมรับได้ เพื่อจัดลำดับความสำคัญและตัดสินใจว่าความเสี่ยงใดที่ต้องมีมาตรการจัดการ โดยมีตัวอย่างเช่น การประเมินระดับความเสี่ยงจากผลการวิเคราะห์ (โอกาสเกิด “ปานกลาง” x ผลกระทบ “สูง”) ทำให้ความเสี่ยงโดยรวมอยู่ในระดับ “สูง” ซึ่งเกินกว่าระดับความเสี่ยงที่องค์กรยอมรับได้ ดังนั้น ความเสี่ยงนี้จึงต้องได้รับการจัดการ โดยต้องกำหนดมาตรการควบคุมในระยะที่ ๑ เช่น การออกแบบกระบวนการตรวจสอบความถูกต้องและแหล่งที่มาของข้อมูลอย่างเข้มงวด

๔) ผลลัพธ์ที่คาดหวัง

เมื่อสิ้นสุดระยะที่ ๐ องค์กรควรมีเอกสารและข้อกำหนดที่ชัดเจน ซึ่งเป็นผลลัพธ์จากการประเมินความเสี่ยง ได้แก่

๔.๑) สรุปข้อกำหนดทางกฎหมายและข้อบังคับที่เกี่ยวข้อง

๔.๒) ทะเบียนความเสี่ยง เอกสารที่รวบรวมความเสี่ยงทั้งหมดของระบบปัญญาประดิษฐ์ที่ระบุและประเมินไว้

๔.๓) ข้อกำหนดด้านความมั่นคงปลอดภัยเบื้องต้น ข้อกำหนดที่ต้องถูกนำไปพิจารณาในขั้นตอนการออกแบบ เช่น "ระบบต้องมีกลไกป้องกันการโจมตีแบบ Prompt Injection" หรือ "ข้อมูลส่วนบุคคลทั้งหมดในชุดข้อมูลต้องผ่านกระบวนการทำข้อมูลนิรนาม"

๔.๔) เอกสารประเมินผลกระทบของระบบปัญญาประดิษฐ์ตามแนวทางของ ISO/IEC 42001:2023 เพื่อทำความเข้าใจผลกระทบในวงกว้างต่อบุคคลและสังคม

เอกสารอ้างอิง:

๑) ISO/IEC 23894:2023 Annex B Risk sources, Annex C Risk management and AI system life cycle

๒) ISO/IEC 42001:2023 6.1 Actions to address risks and opportunities, A.5 Assessing impacts of AI systems

๓) OWASP-AI-EXCHANGE 3. Development-time threats, 2. Threats through use, 4. Runtime application security threats

๓.๒.๒ ระยะที่ ๑ การออกแบบอย่างมั่นคงปลอดภัย

ระยะการออกแบบอย่างมั่นคงปลอดภัยเป็นขั้นตอนการแปรเปลี่ยนความเสี่ยงที่ระบุไว้ในระยะแนวคิด ให้กลายเป็นข้อกำหนดด้านความมั่นคงปลอดภัยและสถาปัตยกรรมของระบบที่จับต้องได้

วัตถุประสงค์หลักคือการสร้าง "พิมพ์เขียวที่มั่นคงปลอดภัย" สำหรับระบบปัญญาประดิษฐ์ ก่อนที่จะเริ่มกระบวนการพัฒนาใด ๆ การแก้ไขช่องโหว่ในระยะนี้มีต้นทุนต่ำกว่าการแก้ไขในระยะหลังการพัฒนาอย่างมีนัยสำคัญ ซึ่งสอดคล้องกับหลักการออกแบบอย่างมั่นคงปลอดภัย

กิจกรรมหลักในระยะนี้ประกอบด้วยการสร้างแบบจำลองภัยคุกคามและการออกแบบสถาปัตยกรรมที่มั่นคงปลอดภัย

๑) การสร้างแบบจำลองภัยคุกคามสำหรับระบบปัญญาประดิษฐ์

การสร้างแบบจำลองภัยคุกคามเป็นกระบวนการเชิงรุกที่เป็นระบบ เพื่อวิเคราะห์ ระบุ และจัดลำดับความสำคัญของภัยคุกคามที่อาจเกิดขึ้นกับระบบปัญญาประดิษฐ์ การทำแบบจำลองภัยคุกคามในบริบทของระบบปัญญาประดิษฐ์ จะต้องพิจารณาพื้นผิวการโจมตี ที่เป็นเอกลักษณ์ของเฉพาะเจาะจงต่อวงจรชีวิตของระบบปัญญาประดิษฐ์ทั้งหมด ตั้งแต่ข้อมูลไปจนถึงโมเดลและ API

กระบวนการนี้สามารถใช้กรอบการวิเคราะห์ภัยคุกคาม เช่น STRIDE (Spoofing, Tampering, Repudiation, Information Disclosure, Denial of Service, Elevation of Privilege) มาประยุกต์ใช้กับส่วนประกอบต่าง ๆ ของปัญญาประดิษฐ์ได้ ดังนี้

๑.๑) ตัวอย่างการประยุกต์ใช้แบบจำลองภัยคุกคามกับระบบปัญญาประดิษฐ์ ระบบแนะนำผลิตภัณฑ์ ส่วนประกอบ Pipeline ข้อมูล

๑.๑.๑) ภัยคุกคาม ผู้ไม่หวังดีหรือคู่แข่งอาจแทรกข้อมูลการซื้อหรือรีวิวปลอมเข้ามาในชุดข้อมูลฝึกสอน เพื่อบิดเบือนให้โมเดลแนะนำผลิตภัณฑ์ของตนเองมากกว่าของคู่แข่ง

๑.๑.๒) มาตรการควบคุมเชิงออกแบบ ออกแบบระบบการตรวจสอบข้อมูล

เพื่อตรวจสอบแหล่งที่มาและความสมบูรณ์ของข้อมูล กำหนดให้มีขั้นตอนการตรวจสอบความถูกต้องของข้อมูลและการคัดกรองข้อมูลที่เข้มงวดก่อนนำข้อมูลเข้าสู่กระบวนการฝึกสอน (อ้างอิง OWASP-AI-EXCHANGE #DEVDATAPROTECT)

๑.๒) ส่วนประกอบโมเดลปัญญาประดิษฐ์

๑.๒.๑) ภัยคุกคาม โมเดลอาจจดจำและเปิดเผยข้อมูลส่วนบุคคลของลูกค้าที่อยู่ในชุดข้อมูลฝึกสอนโดยไม่ได้ตั้งใจ ผ่านผลลัพธ์การแนะนำผลิตภัณฑ์

๑.๒.๒) มาตรการควบคุมเชิงออกแบบ กำหนดในสถาปัตยกรรมให้ต้องใช้เทคนิค Privacy-Preserving Machine Learning เช่น Differential Privacy ในระหว่างกระบวนการฝึกสอน เพื่อลดความเสี่ยงที่โมเดลจะจดจำข้อมูลที่ละเอียดอ่อน

๑.๓) ส่วนประกอบ API สำหรับให้บริการ

๑.๓.๑) ภัยคุกคาม ผู้โจมตีส่งคำร้อง (Query) ไปยัง API จำนวนมหาศาลเพื่อพยายามทำวิศวกรรมย้อนกลับและขโมยโมเดลหรือทำให้ระบบล่ม

๑.๓.๒) มาตรการควบคุมเชิงออกแบบ ออกแบบ API ให้บังคับใช้กลไกการยืนยันตัวตน การอนุญาตสิทธิ์ และการจำกัดอัตราการเรียกใช้งานอย่างเข้มงวด

๒) การออกแบบสถาปัตยกรรมที่มั่นคงปลอดภัย

จากผลลัพธ์ของการสร้างแบบจำลองภัยคุกคาม ขั้นตอนถัดมาคือการออกแบบสถาปัตยกรรมของระบบปัญญาประดิษฐ์ที่มีความยืดหยุ่นและทนทานต่อการโจมตี โดยอาศัยหลักการทางวิศวกรรมและสถาปัตยกรรมที่มั่นคงปลอดภัย (อ้างอิง ISO/IEC 27002:2022 A8.27) ซึ่งมีหลักการสำคัญ ดังนี้

๒.๑) การออกแบบกลไกป้องกันผลลัพธ์ที่เป็นอันตราย

สำหรับระบบ Generative AI ต้องมีการออกแบบกลไกควบคุม เช่น การกรองผลลัพธ์เพื่อป้องกันไม่ให้ระบบถูกใช้เป็นเครื่องมือในการสร้างเนื้อหาที่เป็นอันตรายตามที่ผู้ใช้ร้องขอ เช่น โคดของมัลแวร์ หรือข้อความที่ใช้ในการหลอกลวง

๒.๒) การแยกส่วน

๒.๒.๑) หลักการ การแยกส่วนประกอบต่าง ๆ ของระบบออกจากกัน เพื่อจำกัดขอบเขตของความเสียหายหากส่วนใดส่วนหนึ่งถูกโจมตี

๒.๒.๒) ตัวอย่างในสถาปัตยกรรมปัญญาประดิษฐ์ ออกแบบให้สภาวะแวดล้อมสำหรับการฝึกสอนโมเดล ซึ่งมีข้อมูลดิบที่ละเอียดอ่อน แยกขาดจากสภาวะแวดล้อมสำหรับการให้บริการโมเดลที่ต้องเชื่อมต่อกับภายนอก การบุกรุกที่ Inference API จะต้องไม่สามารถเข้าถึงสภาวะแวดล้อมสำหรับการฝึกสอนโมเดลได้

๒.๓) การตรวจจับ

๒.๓.๑) หลักการ การมีกลไกสำหรับตรวจจับกิจกรรมที่น่าสงสัยหรือเป็นอันตราย

๒.๓.๒) ตัวอย่างในสถาปัตยกรรมปัญญาประดิษฐ์ ออกแบบระบบเฝ้าระวังที่ไม่ได้ดูแค่การใช้ CPU และ Memory แต่สามารถตรวจจับความเบี่ยงเบนของข้อมูล หรือพฤติกรรมของโมเดลที่ผิดปกติ ซึ่งอาจเป็นสัญญาณของการโจมตีแบบ Data Poisoning ที่ค่อย ๆ เกิดขึ้น นอกจากนี้ยังต้องออกแบบให้สามารถตรวจจับรูปแบบการส่งคำร้องที่ผิดปกติซึ่งอาจบ่งชี้ถึงความพยายามทำ Model Extraction

๒.๔) กลไกป้องกันความเสียหายเมื่อระบบล้มเหลว

๒.๔.๑) หลักการ ออกแบบให้ระบบเข้าสู่สถานะที่ปลอดภัยที่สุดเมื่อเกิดความล้มเหลวหรือตรวจพบการโจมตี

๒.๔.๒) ตัวอย่างในสถาปัตยกรรมปัญญาประดิษฐ์ สำหรับ LLM ที่ใช้ภายในองค์กร หากระบบตรวจจับได้ว่ามี Prompt Injection ที่พยายามเข้าถึงข้อมูลลับ กลไกป้องกันความเสียหายเมื่อระบบล้มเหลวที่ออกแบบไว้คือการตัดการทำงานและส่งคืนคำตอบมาตรฐานที่ปลอดภัย เช่น "ขออภัย ไม่สามารถดำเนินการตามคำขอนี้ได้" แทนที่จะพยายามประมวลผลคำสั่งที่เป็นอันตรายนั้นต่อไป

๒.๕) ระบบสำรอง

๒.๕.๑) หลักการ การมีส่วนประกอบสำรองเพื่อรับประกันความพร้อมใช้งานและความถูกต้องสมบูรณ์

๒.๕.๒) ตัวอย่างในสถาปัตยกรรมปัญญาประดิษฐ์ สำหรับระบบปัญญาประดิษฐ์วินิจฉัยโรคที่มีความสำคัญสูง การออกแบบอาจกำหนดให้ใช้โมเดลแบบรวมกลุ่ม (Ensemble Models) ที่ประกอบด้วยโมเดลหลาย ๆ ตัวที่มีสถาปัตยกรรมหรือถูกฝึกสอนบนชุดข้อมูลที่แตกต่างกันเล็กน้อย หากโมเดลตัวหนึ่งถูกหลอกลวงโดยการโจมตีเชิงปรปักษ์ที่จำเพาะเจาะจงโมเดลตัวอื่น ๆ จะยังสามารถให้ผลลัพธ์ที่ถูกต้องเพื่อใช้ในการตรวจสอบและยืนยันซึ่งกันและกันได้ เป็นการป้องกัน ความล้มเหลวที่จุดเดี่ยว (Single Point of Failure)

๒.๖) ความสามารถในการฟื้นตัว

๒.๖.๑) หลักการ การออกแบบให้ระบบสามารถกู้คืนกลับสู่สภาวะปกติที่เชื่อถือได้หลังจากเกิดเหตุการณ์ด้านความมั่นคงปลอดภัย เช่น ข้อมูลหรือโมเดลถูกบิดเบือนหรือทำลาย

๒.๖.๒) ตัวอย่างในสถาปัตยกรรมปัญญาประดิษฐ์ จัดทำกระบวนการ สำรองข้อมูลที่ครอบคลุมทั้ง ชุดข้อมูลที่ใช้ฝึกสอน และโมเดลที่ผ่านการตรวจสอบแล้วอย่างสม่ำเสมอ ในกรณีที่ตรวจพบว่าข้อมูลหรือโมเดลถูกโจมตีจนเสียหาย เช่น จากการโจมตีแบบ Data Poisoning ขั้นรุนแรง สถาปัตยกรรม

ต้องรองรับ การกู้คืนระบบกลับสู่สถานะที่เชื่อถือได้ล่าสุดได้อย่างรวดเร็ว เพื่อลดผลกระทบต่อการดำเนินงาน

๓) ข้อพิจารณาเพิ่มเติมสำหรับระบบปัญญาประดิษฐ์ที่จัดหาจากภายนอก

ในกรณีที่องค์กรจัดหาหรือใช้บริการแพลตฟอร์มระบบปัญญาประดิษฐ์จากผู้ให้บริการภายนอกซึ่งองค์กรไม่ได้เป็นผู้พัฒนาโดยตรง การดำเนินการจะเปลี่ยนจากการออกแบบสถาปัตยกรรมเองไปเน้นที่กระบวนการกำกับดูแลและทวนสอบความน่าเชื่อถือของผู้ให้บริการ โดยมีกิจกรรมที่ต้องพิจารณาดังนี้

๓.๑) การประเมินความน่าเชื่อถือของผู้ให้บริการ

๓.๑.๑) ทวนสอบว่าผู้ให้บริการได้รับการรับรองตามมาตรฐานสากลหรือไม่ เช่น ISO/IEC 27001:2022 ISO/IEC 42001:2023 หรือ SOC 2

๓.๑.๒) ร้องขอเอกสารที่เกี่ยวข้อง เช่น รายงานผลการทดสอบเจาะระบบหรือนโยบายการจัดการความมั่นคงปลอดภัยของระบบปัญญาประดิษฐ์ของผู้ให้บริการ

๓.๒) การสร้างแบบจำลองภัยคุกคามในส่วนที่เชื่อมต่อ

๓.๒.๑) แม้จะไม่สามารถสร้างแบบจำลองภัยคุกคามของตัวแพลตฟอร์มได้ แต่องค์กรต้องสร้างแบบจำลองภัยคุกคามในส่วนของการเชื่อมต่อระหว่างระบบขององค์กรและผู้ให้บริการเพื่อประเมินความเสี่ยง เช่น การรั่วไหลของข้อมูลระหว่างการส่งผ่านหรือการใช้ API Key ในทางที่ผิด

๓.๓) ข้อตกลงทางสัญญาและกฎหมาย

๓.๓.๑) ตรวจสอบให้แน่ใจว่าสัญญาระบุถึงความรับผิดชอบของผู้ให้บริการอย่างชัดเจนกรณีเกิดเหตุการณ์ด้านความมั่นคงปลอดภัย

๓.๓.๒) ทบทวนนโยบายความเป็นส่วนตัวของผู้ให้บริการ เพื่อทำความเข้าใจว่าข้อมูลที่องค์กรส่งไปจะถูกนำไปใช้เพื่อวัตถุประสงค์อื่น เช่น การนำไปฝึกสอนโมเดลต่อหรือไม่ และข้อมูลถูกจัดเก็บในเขตอำนาจกฎหมายใด

๔) ผลลัพธ์ที่คาดหวัง

เมื่อสิ้นสุดระยะที่ ๑ องค์กรควรมีเอกสารประกอบการออกแบบที่ชัดเจน ได้แก่

๔.๑) เอกสารแบบจำลองภัยคุกคาม ระบุภัยคุกคาม ช่องโหว่ และมาตรการควบคุมที่จำเป็นสำหรับระบบปัญญาประดิษฐ์

๔.๒) แผนภาพและข้อกำหนดสถาปัตยกรรมที่มั่นคงปลอดภัย แสดงโครงสร้างของระบบและกลไกความปลอดภัยที่ออกแบบไว้

๔.๓) ข้อกำหนดด้านความมั่นคงปลอดภัยโดยละเอียด ซึ่งจะถูกส่งมอบให้ทีมพัฒนาเพื่อนำไปปฏิบัติในระยะที่ ๒ ต่อไป

เอกสารอ้างอิง:

๑) ISO/IEC 23894:2023 - AI — Guidance on risk management

๒) ISO/IEC 27002:2022 8.27 Secure system architecture and engineering principles

๓) ISO/IEC 42001:2023 A.6 AI system life cycle

๔) OWASP-AI-EXCHANGE 3. Development-time threats #DEVDATAPROTECT, #DEVSECURITY

๓.๒.๓ ระยะที่ ๒ การพัฒนาอย่างมั่นคงปลอดภัย

ระยะการพัฒนาอย่างมั่นคงปลอดภัย คือขั้นตอนการแปรเปลี่ยนพิมพ์เขียวเชิงทฤษฎีและข้อกำหนดด้านความปลอดภัย จากระยะการออกแบบในระยะที่ ๑ ให้กลายเป็นผลลัพธ์เชิงปฏิบัติที่มั่นคงปลอดภัย

วัตถุประสงค์หลักของระยะนี้ คือการนำสถาปัตยกรรมที่ออกแบบไว้มาสร้างเป็นโคดฝึกสอนโมเดล และประกอบส่วนต่าง ๆ ขึ้นเป็นระบบต้นแบบ โดยยังคงรักษาและบังคับใช้มาตรการควบคุมความมั่นคงปลอดภัยอย่างเคร่งครัดตลอดกระบวนการ

กิจกรรมหลักในระยะนี้มุ่งเน้นไปที่ความมั่นคงปลอดภัยของห่วงโซ่อุปทาน การคุ้มครองสินทรัพย์ระหว่างการพัฒนา และการประยุกต์ใช้แนวปฏิบัติการเขียนโคดที่ปลอดภัย

๑) การบริหารจัดการความมั่นคงปลอดภัยของห่วงโซ่อุปทานปัญญาประดิษฐ์

ไม่ว่าองค์กรจะพัฒนาโมเดลขึ้นเองทั้งหมด หรือนำโมเดลโอเพนซอร์สจากภายนอกมาปรับใช้ก็ตาม การกระทำดังกล่าวถือเป็นส่วนหนึ่งของห่วงโซ่อุปทานปัญญาประดิษฐ์ ซึ่งแนวปฏิบัติในส่วนนี้ยังคงมีผลบังคับใช้อย่างสมบูรณ์ ห่วงโซ่อุปทานของปัญญาประดิษฐ์มีความซับซ้อนและหลากหลายกว่าซอฟต์แวร์ทั่วไป โดยประกอบด้วย ข้อมูลโมเดลที่ฝึกไว้ล่วงหน้า (Pre-Trained Models) ไลบรารีซอฟต์แวร์ และแพลตฟอร์มคลาวด์ ช่องโหว่เพียงจุดเดียวในห่วงโซ่อุปทานนี้อาจส่งผลกระทบต่อความมั่นคงปลอดภัยของทั้งระบบ (อ้างอิง OWASP-AI-EXCHANGE #SUPPLYCHAINMANAGE)

๑.๑) การประเมินส่วนประกอบโอเพนซอร์ส

๑.๑.๑) บริบทของปัญญาประดิษฐ์ การพัฒนาปัญญาประดิษฐ์พึ่งพาไลบรารีโอเพนซอร์สจำนวนมาก เช่น TensorFlow PyTorch Hugging Face Transformers ช่องโหว่ในไลบรารีเหล่านี้ เช่น ช่องโหว่ในฟังก์ชันการโหลดข้อมูลหรือการประมวลผลโมเดล สามารถถูกใช้เป็นช่องทางโจมตีระบบได้โดยตรง

๑.๑.๒) มาตรการควบคุม

▪ ใช้เครื่องมือ Software Composition Analysis (SCA) เพื่อสแกนหาช่องโหว่ (CVEs) ที่เป็นที่รู้จักในไลบรารีและส่วนประกอบที่เกี่ยวข้องทั้งหมด

▪ จัดให้มีความสามารถในการจำแนกและติดตามส่วนประกอบซอฟต์แวร์ทั้งหมดที่ใช้ในระบบ เพื่อให้สามารถบริหารจัดการและตอบสนองต่อช่องโหว่ที่ถูกค้นพบใหม่ได้อย่างมีประสิทธิภาพ โดยกระบวนการดังกล่าวควรสร้างผลลัพธ์ที่มีโครงสร้างชัดเจนและเป็นที่ยอมรับ เพื่อให้สามารถนำไปใช้แลกเปลี่ยนข้อมูลระหว่างเครื่องมือและระบบต่าง ๆ ได้ เพื่อให้สามารถทำงานร่วมกันได้ เช่น รูปแบบที่สอดคล้องกับแนวทางของ ISO/IEC 5962 (SPDX) หรือ ECMA-424 (CycloneDX) หรือแนวทางที่องค์กรได้นำมากำหนดและประยุกต์ใช้ในองค์กร

๑.๒) การทวนสอบโมเดลที่ฝึกไว้ล่วงหน้า

๑.๒.๑) บริบทของปัญญาประดิษฐ์ การใช้โมเดลที่ฝึกไว้ล่วงหน้าจากแหล่งภายนอก เช่น Model Zoos มีความเสี่ยงที่โมเดลเหล่านั้นอาจถูกฝัง Backdoor หรือมัลแวร์ โดยใช้เทคนิค Model Poisoning ซึ่งจะทำงานเมื่อได้รับข้อมูลนำเข้าที่จำเพาะเจาะจง

๑.๒.๒) มาตรการควบคุม

- ดาวน์โหลดโมเดลจากแหล่งที่น่าเชื่อถือและผ่านการรับรองเท่านั้น
- ตรวจสอบค่าผลรวมตรวจสอบ (Checksum) ของไฟล์โมเดลเพื่อให้แน่ใจว่าไม่มีการเปลี่ยนแปลง
- ใช้เครื่องมือสแกนโมเดลเพื่อตรวจหาสิ่งผิดปกติหรือโคดอันตรายที่อาจแฝงอยู่ก่อนนำมาใช้งาน

๒) การคุ้มครองสินทรัพย์ในกระบวนการพัฒนา

สินทรัพย์ของปัญญาประดิษฐ์ไม่ได้จำกัดอยู่แค่ซอร์สโคด แต่ยังรวมถึงองค์ประกอบอื่น ๆ ที่มีความสำคัญและมูลค่าสูง ซึ่งต้องได้รับการคุ้มครองอย่างเหมาะสมตลอดกระบวนการพัฒนา (อ้างอิง ISO/IEC 27002:2022 #Asset_management)

๒.๑) ชุดข้อมูลฝึกสอนและข้อมูลทดสอบ

๒.๑.๑) ความเสี่ยง การรั่วไหลของข้อมูลฝึกสอนอาจละเมิดความเป็นส่วนตัวของเจ้าของข้อมูลอย่างรุนแรง

๒.๑.๒) มาตรการควบคุม บังคับใช้นโยบายการเข้ารหัสลับข้อมูลขณะจัดเก็บ กำหนดสิทธิ์การเข้าถึงข้อมูลอย่างรัดกุมด้วยหลักการกำหนดสิทธิ์เข้าถึงให้น้อยที่สุด และใช้เทคนิคการปิดข้อมูลหรือการทำข้อมูลนิรนาม กับข้อมูลในสภาพแวดล้อมที่ไม่ใช่การใช้งานจริง (อ้างอิง OWASP #DEVDATAPROTECT)

๒.๒) โมเดลปัญญาประดิษฐ์

๒.๒.๑) ความเสี่ยง โมเดลที่ฝึกสอนเสร็จแล้วเป็นทรัพย์สินทางปัญญาที่มีมูลค่าสูงและอาจเป็นเป้าหมายของการลักลอบนำออกไป

๒.๒.๒) มาตรการควบคุม จัดเก็บโมเดลในที่ปลอดภัยและมีการควบคุมการเข้าถึงที่เข้มงวด ใช้ระบบควบคุมเวอร์ชันสำหรับโมเดล เช่น Data Version Control (DVC) หรือ Git Large File Storage (LFS) เพื่อติดตามการเปลี่ยนแปลงและสามารถกู้คืนเวอร์ชันที่ปลอดภัยได้ และมีมาตรการปกป้องที่เหมาะสม เช่น การเข้ารหัสลับ โมเดลเมื่อไม่ได้ใช้งาน

๒.๓) การควบคุมเวอร์ชันของสินทรัพย์ปัญญาประดิษฐ์

๒.๓.๑) ความเสี่ยง การเปลี่ยนแปลงของชุดข้อมูลฝึกสอนหรือโมเดลโดยไม่มีการควบคุมเวอร์ชัน ทำให้ไม่สามารถตรวจสอบย้อนกลับ หรือกู้คืนเวอร์ชันที่ปลอดภัยได้เมื่อเกิดปัญหา มาตรการควบคุม

๒.๓.๒) ซอร์สโคด ใช้ระบบควบคุมเวอร์ชันมาตรฐาน เช่น Git

๒.๓.๓) ชุดข้อมูลและโมเดล ใช้เครื่องมือที่ออกแบบมาเพื่อควบคุมเวอร์ชันของไฟล์ขนาดใหญ่และชุดข้อมูลโดยเฉพาะ เช่น DVC หรือ Git LFS เพื่อให้สามารถติดตามการเปลี่ยนแปลงของข้อมูลที่ใช้ฝึกสอนและโมเดลที่ได้ในแต่ละเวอร์ชันได้อย่างเป็นระบบ

๒.๔) Prompt และไฟล์การกำหนดค่า

๒.๔.๑) ความเสี่ยง สำหรับ LLM System Prompts หรือชุดคำสั่งที่ออกแบบมาอย่างดีถือเป็นความลับทางการค้า การรั่วไหลอาจทำให้คู่แข่งลอกเลียนแบบความสามารถของระบบได้

๒.๔.๒) มาตรการควบคุม จัดการ Prompts และ API Keys เสมือนเป็นข้อมูลลับจัดเก็บในระบบจัดการข้อมูลลับโดยเฉพาะ และห้ามเก็บเป็นข้อความธรรมดา (Plain Text) ในซอร์สโคดเด็ดขาด

๓) แนวปฏิบัติการเขียนโคดและการแบ่งแยกสภาพแวดล้อม

๓.๑) การเขียนโคดอย่างปลอดภัยสำหรับแอปพลิเคชันปัญญาประดิษฐ์

๓.๑.๑) บริบทของปัญญาประดิษฐ์ นอกเหนือจากช่องโหว่ทั่วไป โคดที่ใช้ในแอปพลิเคชันปัญญาประดิษฐ์ต้องจัดการกับความเสี่ยงเฉพาะทาง

๓.๑.๒) มาตรการควบคุม

- การตรวจสอบความถูกต้องของข้อมูลนำเข้า โคดที่รับ Prompt จากผู้ใช้ต้องผ่านการตรวจสอบและคัดกรองอย่างเข้มงวด เพื่อป้องกันการโจมตีแบบ Prompt Injection
- การจัดการข้อผิดพลาดอย่างปลอดภัย หลีกเลี่ยงการแสดงข้อมูลทางเทคนิคที่ละเอียดอ่อนเกี่ยวกับสถาปัตยกรรมของโมเดลหรือไลบรารีที่ใช้ ซึ่งอาจเป็นประโยชน์ต่อผู้โจมตี

▪ ใช้เครื่องมือ Static Application Security Testing (SAST) สแกนหาช่องโหว่ในโคดที่เขียนขึ้นเองอย่างสม่ำเสมอ (อ้างอิง ISO/IEC 27002 #Application_security)

๓.๒) การแบ่งแยกสภาพแวดล้อม

๓.๒.๑) บริบทของปัญญาประดิษฐ์ การแยกระหว่างสภาพแวดล้อมการพัฒนา การทดสอบ และการใช้งานจริงเป็นสิ่งจำเป็นอย่างยิ่งเพื่อจำกัดขอบเขตความเสียหาย โดยแต่ละสภาพแวดล้อมมีวัตถุประสงค์ที่แตกต่างกันดังนี้

- **การพัฒนา** เป็นสภาพแวดล้อมที่เปลี่ยนแปลงบ่อยที่สุดสำหรับนักพัฒนาในการเขียนโคดและทดสอบเบื้องต้น ไม่มีความเสถียรและไม่เหมือนกับสภาพแวดล้อมจริง
- **การทดสอบ** เป็นสภาพแวดล้อมที่จำลองให้ใกล้เคียงกับการใช้งานจริงมากที่สุด มีไว้เพื่อเป็น "ด่านสุดท้าย" สำหรับการทดสอบที่ครอบคลุม เช่น การทดสอบโดยผู้ใช้ (UAT) การทดสอบประสิทธิภาพ และการทดสอบเจาะระบบ เพื่อให้มั่นใจว่าระบบทำงานได้ถูกต้องและปลอดภัยก่อนนำขึ้นใช้งานจริง การมีสภาพแวดล้อมนี้ช่วยป้องกันไม่ให้ข้อผิดพลาดหลุดรอดไปสร้างความเสียหายในการใช้งานจริงได้
- **การใช้งานจริง** เป็นสภาพแวดล้อมสำหรับให้บริการผู้ใช้จริง มีความเสถียรและต้องมีมาตรการควบคุมที่เข้มงวดที่สุด

๓.๒.๒) มาตรการควบคุม สภาพแวดล้อมการใช้งานจริงต้องถูกแยกขาดและมีมาตรการควบคุมที่เข้มงวดที่สุด สภาพแวดล้อมการทดสอบควรใช้ชุดข้อมูลที่ไม่ละเอียดอ่อน เพื่อทำการทดสอบและการเข้าถึงระหว่างสภาพแวดล้อมต่าง ๆ ต้องถูกควบคุมและตรวจสอบอย่างรัดกุม

๔) ข้อพิจารณาด้านความมั่นคงปลอดภัยในการใช้ API ของบุคคลที่สาม

การเรียกใช้โมเดลปัญญาประดิษฐ์ผ่าน API ของผู้ให้บริการภายนอก ก่อให้เกิดความเสี่ยงด้านการรั่วไหลของข้อมูลและความเป็นส่วนตัว องค์กรต้องมีมาตรการควบคุมดังนี้

๔.๑) การป้องกันข้อมูลรั่วไหล

๔.๑.๑) หลักการลดข้อมูลให้เหลือน้อยที่สุด ส่งข้อมูลไปยัง API เท่าที่จำเป็นต่อการประมวลผลเท่านั้น

๔.๑.๒) การกรองข้อมูลก่อนส่ง ใช้เทคนิคการปิดข้อมูล หรือการทำนามแฝงเพื่อลบหรือแทนที่ข้อมูลส่วนบุคคล หรือข้อมูลที่เป็นความลับขององค์กร ก่อนที่จะส่งข้อมูลออกจากเครือข่ายขององค์กร

๔.๒) ความมั่นคงปลอดภัยของช่องทางสื่อสาร

๔.๒.๑) ตรวจสอบให้แน่ใจว่าการเชื่อมต่อทั้งหมดใช้โปรโตคอลการเข้ารหัสลับที่แข็งแกร่ง เช่น TLS 1.2 หรือสูงกว่า

๔.๓) การจัดการข้อมูลลับ

๔.๓.๑) ห้าม เก็บ API Keys ในซอร์สโคดโดยเด็ดขาด แต่ให้จัดเก็บและเรียกใช้จากระบบจัดการข้อมูลลับที่ปลอดภัย

๕) ผลลัพธ์ที่คาดหวัง

เมื่อสิ้นสุดระยะที่ ๒ องค์กรควรมีผลลัพธ์ที่พร้อมสำหรับการทวนสอบในระยะต่อไป ได้แก่

๕.๑) ซอร์สโคดของแอปพลิเคชันปัญญาประดิษฐ์ที่ผ่านการตรวจสอบความปลอดภัย

๕.๒) โมเดลปัญญาประดิษฐ์ที่ผ่านการฝึกสอนและจัดเก็บอย่างปลอดภัยพร้อมการควบคุมเวอร์ชัน

๕.๓) ผลลัพธ์เชิงเอกสารที่ระบุส่วนประกอบซอฟต์แวร์ทั้งหมดของระบบ

๕.๔) หลักฐานการกำหนดค่าความปลอดภัยและการแบ่งแยกสภาพแวดล้อม

เอกสารอ้างอิง:

๑) ISO/IEC 27002:2022 security controls under “#Asset_management” “#Application_security” and “#Supplier_relationships_security” attribute

๒) ISO/IEC 42001:2023 A.4 Resources for AI systems, A.5 Assessing impacts of AI systems, A.6 AI system life cycle, A.10 Third-party and customer relationships

๓) ISO 28000:2022 Security and resilience — Security management systems — Requirements

๔) OWASP-AI-EXCHANGE 3. Development-time threats #DEVDATAPROTECT, #DEVSECURITY, #SUPPLYCHAINMANAGE

๓.๒.๔ ระยะที่ ๓ การทวนสอบด้านความมั่นคงปลอดภัย

ระยะการทวนสอบด้านความมั่นคงปลอดภัยเป็นขั้นตอนของการประกันคุณภาพที่สำคัญอย่างยิ่งก่อนการนำระบบปัญญาประดิษฐ์ ไปใช้งานจริง

วัตถุประสงค์หลักของระยะนี้คือการทวนสอบและทดสอบเพื่อให้มั่นใจว่ามาตรการควบคุมความปลอดภัยที่ถูกออกแบบในระยะที่ ๑ และพัฒนาขึ้นในระยะที่ ๒ นั้น สามารถทำงานได้อย่างมีประสิทธิผล สอดคล้องกับข้อกำหนด และทนทานต่อภัยคุกคามที่คาดการณ์ไว้ได้จริง

กระบวนการในระยะนี้เป็นการเปลี่ยนมุมมองจากการ "สร้าง" ไปสู่การ "ตรวจสอบ" อย่างเป็นระบบ เพื่อค้นหาข้อบกพร่องและช่องโหว่ที่อาจหลงเหลืออยู่ก่อนที่จะส่งผลกระทบในสภาพแวดล้อมการใช้งานจริง

๑) การทวนสอบการปฏิบัติตามข้อกำหนด

ขั้นตอนนี้เป็นการตรวจสอบเชิงเอกสารและเชิงเทคนิคว่าระบบที่พัฒนาขึ้นนั้น สอดคล้องกับนโยบาย ข้อกำหนด และมาตรฐานที่กำหนดไว้ในระยะการออกแบบหรือไม่ เป็นการตอบคำถามที่ว่า "เราได้สร้างสิ่งที่ตั้งใจจะสร้างหรือไม่?" (อ้างอิง ISO/IEC 27002:2022 A5.36)

๑.๑) บริบทของปัญญาประดิษฐ์ การทวนสอบไม่ได้จำกัดอยู่แค่โคด แต่ครอบคลุมถึงกระบวนการที่เกี่ยวข้องกับข้อมูลและโมเดลทั้งหมด

๑.๒) ตัวอย่างการทวนสอบ

๑.๒.๑) การทวนสอบการกำกับดูแลข้อมูล ตรวจสอบสคริปต์ (Script) ที่ใช้ในการประมวลผลข้อมูล เพื่อยืนยันว่าได้มีการใช้เทคนิคการทำข้อมูลนิรนามกับข้อมูลส่วนบุคคลจริงตามที่ออกแบบไว้ และไม่มีข้อมูลที่ละเอียดอ่อนรั่วไหลไปยังชุดข้อมูลที่ใช้ในการฝึกสอน

๑.๒.๒) การทวนสอบสถาปัตยกรรม ทบทวนการตั้งค่าบนระบบคลาวด์ เพื่อยืนยันว่าการแบ่งแยกสภาพแวดล้อม ระหว่าง Training และ Inference ได้ถูกนำไปใช้อย่างถูกต้อง และมี Network Policy ที่จำกัดการสื่อสารระหว่างกันจริง

๑.๒.๓) การทวนสอบข้อกำหนดโคด ตรวจสอบซอร์สโคด (Code Review) เพื่อยืนยันว่ามีการเรียกใช้ไลบรารีการเข้ารหัสลับ (Encryption Library) สำหรับการจัดเก็บโมเดลตามที่ระบุไว้ในข้อกำหนดด้านความปลอดภัย

๒) การทวนสอบประสิทธิผลของกลไกความปลอดภัย

ขั้นตอนนี้มุ่งเน้นการทดสอบเชิงปฏิบัติเพื่อยืนยันว่ากลไกความปลอดภัยที่ติดตั้งไว้นั้น "ทำงานได้จริงตามที่คาดหวัง" เมื่อต้องเผชิญกับสถานการณ์ต่าง ๆ ไม่ใช่แค่มีอยู่ตามข้อกำหนดเท่านั้น

๒.๑) บริบทของปัญญาประดิษฐ์ กลไกความปลอดภัยของปัญญาประดิษฐ์มีความซับซ้อนและต้องการการทดสอบที่จำเพาะเจาะจง

๒.๒) ตัวอย่างการทวนสอบ

๒.๒.๑) การทดสอบ Guardrails ของ LLM โดยการสร้างชุดทดสอบที่ประกอบด้วย Prompt ที่มีเนื้อหาไม่เหมาะสม เป็นอันตราย หรือพยายามชี้นำให้เกิดอคติ แล้วส่งไปยังโมเดล LLM เพื่อทวนสอบว่ากลไก Guardrails และ Content Filtering สามารถตรวจจับและปฏิเสธการตอบสนองได้อย่างถูกต้อง

๒.๒.๒) การทดสอบการควบคุมการเข้าถึง โดยการพยายามเข้าถึง API สำหรับการจัดการโมเดลด้วยสิทธิ์ของผู้ใช้งานทั่วไป เพื่อยืนยันว่าระบบจะปฏิเสธการเข้าถึงตามนโยบาย Role-Based Access Control (RBAC) ที่ออกแบบไว้

๒.๒.๓) การทดสอบการแจ้งเตือน โดยการจำลองการส่งคำร้องที่มีลักษณะผิดปกติ เช่น การส่งคำร้องจำนวนมากจาก IP เดียวไปยัง API ของโมเดล เพื่อทวนสอบว่าระบบเฝ้าระวัง สามารถตรวจจับและสร้างการแจ้งเตือนไปยังทีมความมั่นคงปลอดภัยได้จริง

๓) การทวนสอบช่องโหว่และความเสี่ยงผ่านการทดสอบเชิงรุก

ขั้นตอนนี้เป็นการทดสอบเชิงรุก หรือ "การทดสอบเชิงเจาะระบบ" โดยสมมติตัวเป็นผู้ไม่หวังดี เพื่อค้นหาช่องโหว่ที่ไม่เคยถูกพบมาก่อน เป็นการตอบคำถามที่ว่า "เรามองข้ามอะไรไปหรือไม่?" (อ้างอิง ISO/IEC 27002:2022 A8.29, A8.8 และ OWASP-AI-EXCHANGE 5. AI security testing)

๓.๑) บริบทของปัญญาประดิษฐ์ การทดสอบเชิงรุกสำหรับปัญญาประดิษฐ์ หรือที่เรียกว่า "AI Red Teaming" จะใช้เทคนิคการโจมตีที่จำเพาะเจาะจงกับปัญญาประดิษฐ์โดยเฉพาะ

๓.๒) ตัวอย่างการทวนสอบ

๓.๒.๑) การทดสอบด้วยตัวอย่างที่เป็นปฏิปักษ์ ทีม Red Team จะสร้างตัวอย่างข้อมูลที่ผ่านการดัดแปลงอย่างแยบยล เช่น ภาพที่มี Noise เล็กน้อย ข้อความที่เปลี่ยนตัวอักษรบางตัว ซึ่งไม่เคยอยู่ในชุดข้อมูลฝึกสอน มาใช้ทดสอบโมเดลเพื่อวัดความทนทานต่อการโจมตีแบบ Evasion Attacks ในสถานการณ์จริง

๓.๒.๒) การทดสอบ Prompt Injection และ Jailbreaking สำหรับ LLM ทีม Red Team จะใช้เทคนิคขั้นสูง เช่น การสวมบทบาท หรือการโจมตีทางอ้อม (Indirect Injection) เพื่อพยายามหลบเลี่ยงกลไก Guardrails ที่ป้องกันอยู่ และบังคับให้โมเดลทำงานนอกเหนือขอบเขตที่ได้รับอนุญาต

๓.๒.๓) การทดสอบการอนุมานข้อมูลส่วนบุคคล ทีมทดสอบพยายามใช้ผลลัพธ์จาก API สาธารณะของโมเดล เพื่ออนุมานข้อมูลที่ละเอียดอ่อนจากชุดข้อมูลฝึกสอน เช่น การโจมตีแบบ Membership Inference เพื่อตรวจสอบว่าข้อมูลของบุคคลใดบุคคลหนึ่งถูกใช้ในการฝึกสอนหรือไม่ ซึ่งเป็นการทวนสอบความเสี่ยงด้านการรั่วไหลของข้อมูล

๓.๒.๔) การทดสอบการใช้งานในทางที่ผิด ทีม Red Team จะพยายามใช้ระบบเพื่อสร้างผลลัพธ์ที่เป็นอันตราย เช่น การสร้าง Prompt เพื่อให้ปัญญาประดิษฐ์เขียนสคริปต์สำหรับมัลแวร์ หรือสร้างอีเมลฟิชชิ่งที่แนบเนียน เพื่อทวนสอบว่ากลไกป้องกันที่ออกแบบไว้สามารถทำงานได้จริง

๔) ผลลัพธ์ที่คาดหวัง

เมื่อสิ้นสุดระยะที่ ๓ องค์กรควรมีเอกสารและหลักฐานที่ชัดเจนเพื่อใช้ในการตัดสินใจก่อนนำระบบขึ้นใช้งานจริง

๔.๑) รายงานผลการทดสอบความมั่นคงปลอดภัย สรุปผลการทวนสอบการปฏิบัติตามข้อกำหนดและผลการทดสอบประสิทธิผลของกลไกต่าง ๆ

๔.๒) รายงานผลการทดสอบเจาะระบบ ระบุช่องโหว่ที่ค้นพบจากการทดสอบเชิงรุก พร้อมระดับความรุนแรงและข้อเสนอแนะในการแก้ไข

๔.๓) ทะเบียนความเสี่ยงฉบับทบทวนที่ได้รับการปรับปรุงข้อมูลความเสี่ยงใหม่ ๆ และยืนยันว่ามาตรการควบคุมที่มีอยู่มีประสิทธิผล

๔.๔) การอนุมัติเพื่อนำไปใช้งาน การตัดสินใจโดยผู้มีอำนาจว่าจะอนุมัติให้นำระบบไปใช้งานจริง แก้ไขข้อบกพร่องก่อน หรือระงับโครงการ ขึ้นอยู่กับระดับความเสี่ยงที่คงเหลืออยู่

เอกสารอ้างอิง:

๑) ISO/IEC 27002:2022 8.29 Security testing in development and acceptance, 8.8 Management of technical vulnerabilities, 5.36 Compliance with policies, rules and standards for information security

๒) ISO/IEC 42001:2023 A.5 Assessing impacts of AI systems, A.6 AI system life cycle

๓) OWASP-AI-EXCHANGE 5. AI security testing

๓.๒.๕ ระยะที่ ๔ การนำไปใช้งานอย่างมั่นคงปลอดภัย

ระยะการนำไปใช้งานอย่างมั่นคงปลอดภัยเป็นขั้นตอนการส่งมอบระบบปัญญาประดิษฐ์ที่ผ่านการทวนสอบและรับรองแล้ว จากสภาพแวดล้อมการทดสอบไปยังสภาพแวดล้อมการใช้งานจริง

วัตถุประสงค์หลักของระยะนี้คือ การรับประกันว่ากระบวนการติดตั้งและการกำหนดค่าทั้งหมดเป็นไปอย่างมั่นคงปลอดภัย และสภาพแวดล้อมการใช้งานจริงได้รับการเสริมความแข็งแกร่งเพื่อป้องกันภัยคุกคามในโลกแห่งความเป็นจริง

กิจกรรมในระยะนี้มุ่งเน้นการสร้างและรักษา “สถานะพื้นฐานด้านความมั่นคงปลอดภัย (Secure Baseline)” สำหรับการดำเนินงานของระบบปัญญาประดิษฐ์

๑) การเสริมความแข็งแกร่งให้โครงสร้างพื้นฐาน

เป็นกระบวนการกำหนดค่าโครงสร้างพื้นฐานที่รองรับการทำงานของปัญญาประดิษฐ์ ให้มีระดับความมั่นคงปลอดภัยสูงสุด เพื่อลดพื้นผิวการโจมตี (อ้างอิง ISO/IEC 27002:2022 A8.9 Configuration management)

๑.๑) บริบทของปัญญาประดิษฐ์ โครงสร้างพื้นฐานของปัญญาประดิษฐ์ไม่ได้มีแค่เซิร์ฟเวอร์ แต่ยังรวมถึงบริการจัดเก็บข้อมูล Container Orchestration และ API Gateway ซึ่งแต่ละส่วนต้องได้รับการกำหนดค่าอย่างรัดกุม

๑.๒) ตัวอย่างมาตรการควบคุม

๑.๒.๑) การบังคับใช้หลักการสิทธิ์น้อยที่สุด บัญชีที่ใช้ในการรันโมเดลปัญญาประดิษฐ์จะต้องได้รับสิทธิ์เฉพาะที่จำเป็นต่อการทำงานเท่านั้น เช่น สิทธิ์ในการอ่านไฟล์โมเดล สิทธิ์ในการเขียนไฟล์บันทึก (Log) และสิทธิ์ในการประมวลผล แต่ต้องไม่มีสิทธิ์ในการเข้าถึงชุดข้อมูลฝึกสอน หรือแก้ไขตัวโมเดลโดยเด็ดขาด เพื่อจำกัดขอบเขตความเสียหาย หาก Inference Service ถูกบุกรุก

๑.๒.๒) การใช้อิมเมจที่ผ่านการเสริมความแข็งแกร่ง ใช้ Container Image ที่ผ่านการปรับแต่งให้มีความปลอดภัย โดยการลบไลบรารี เครื่องมือ และ Shell ที่ไม่จำเป็นออกทั้งหมด เพื่อลดช่องทางที่ผู้บุกรุกจะสามารถใช้ประโยชน์ได้หลังเจาะเข้ามาในระบบ

๑.๒.๓) การแบ่งส่วนเครือข่าย กำหนดค่า Network Policy ให้ API Endpoint ของโมเดลปัญญาประดิษฐ์ สามารถสื่อสารได้เฉพาะกับส่วนประกอบที่ได้รับอนุญาตเท่านั้น และป้องกันด้วย Web Application Firewall (WAF) ที่มีการตั้งค่ากฎ สำหรับป้องกันการโจมตี API โดยเฉพาะ

๒) การคุ้มครองโมเดลและข้อมูลในสภาวะแวดล้อมการใช้งานจริง

เป็นการใช้เทคนิคการเข้ารหัสลับและเทคโนโลยีขั้นสูงเพื่อคุ้มครองสินทรัพย์ปัญญาประดิษฐ์ที่มีความสำคัญสูงสุด (โมเดลและข้อมูล) ในทุกสถานะ ทั้งขณะจัดเก็บ (At-Rest) ขณะส่งผ่าน (In-Transit) และที่สำคัญคือ ขณะประมวลผล (In-Use)

๒.๑) บริบทของปัญญาประดิษฐ์ โมเดลปัญญาประดิษฐ์คือทรัพย์สินทางปัญญา และข้อมูลที่ผู้ใช้ส่งเข้ามาประมวลผลอาจเป็นข้อมูลที่ละเอียดอ่อนอย่างยิ่ง

๒.๒) ตัวอย่างมาตรการควบคุม

๒.๒.๑) การเข้ารหัสลับโมเดลและการจัดการคีย์ ไฟล์โมเดลที่จัดเก็บไว้ใน Object Storage หรือ File Server จะต้องถูกเข้ารหัสลับไว้เสมอ เมื่อระบบเริ่มต้นทำงาน จะต้องมีการเรียกใช้คีย์สำหรับถอดรหัสจากบริการจัดการคีย์โดยเฉพาะ เช่น Hardware Security Module (HSM) หรือ Key Management Service (KMS) บนคลาวด์ ซึ่งการเข้าถึงคีย์ทุกครั้งจะต้องถูกบันทึกและตรวจสอบอย่างเข้มงวด

๒.๒.๒) การประมวลผลแบบเป็นความลับ เป็นเทคโนโลยีที่ใช้ฮาร์ดแวร์ในการสร้าง Trusted Execution Environment (TEE) หรือ "Secure Enclave" ซึ่งเป็นพื้นที่ที่ถูกแยกขาดและเข้ารหัสลับในหน่วยความจำ (RAM)

- ประโยชน์ต่อปัญญาประดิษฐ์ เทคโนโลยีนี้ช่วยให้สามารถประมวลผลข้อมูลที่ละเอียดอ่อน เช่น ข้อมูลทางการแพทย์ บนโมเดลปัญญาประดิษฐ์ได้ โดยที่แม้แต่ผู้ให้บริการคลาวด์หรือผู้ดูแลระบบก็ไม่สามารถเข้าถึงหรือมองเห็นข้อมูลและโมเดลขณะกำลังประมวลผลได้ เป็นการยกระดับ

การคุ้มครองข้อมูลในขณะประมวลผล ซึ่งเป็นสิ่งจำเป็นสำหรับระบบปัญญาประดิษฐ์ ที่จัดการกับข้อมูลที่มีความอ่อนไหวสูง

๓) การทดสอบและทวนสอบขั้นสุดท้ายก่อนการใช้งานจริง

เป็นด่านสุดท้ายในการประกันคุณภาพความมั่นคงปลอดภัย เพื่อยืนยันว่าระบบปัญญาประดิษฐ์ที่ติดตั้งบนสภาพแวดล้อมการใช้งานจริงนั้นมีความปลอดภัยตามที่คาดหวัง การทดสอบในระยะนี้มีความสำคัญเนื่องจากสภาพแวดล้อมการใช้งานจริงอาจมีการกำหนดค่าหรือนโยบายเครือข่ายที่แตกต่างจากสภาพแวดล้อม Testing (อ้างอิง ISO/IEC 42001:2023 A.6.2.5)

๓.๑) บริบทของปัญญาประดิษฐ์ เป็นการทวนสอบการทำงานร่วมกันของทุกองค์ประกอบในสภาพแวดล้อมจริง

๓.๒) ตัวอย่างการทวนสอบ

๓.๒.๑) การตรวจสอบความถูกต้องของการตั้งค่า ใช้เครื่องมืออัตโนมัติในการสแกนและเปรียบเทียบการตั้งค่าของสภาพแวดล้อมการใช้งานจริงกับเอกสารสถานะพื้นฐานด้านความมั่นคงปลอดภัย เพื่อให้แน่ใจว่าไม่มีการตั้งค่าที่ไม่ปลอดภัยหลงเหลืออยู่

๓.๒.๒) การทดสอบโดยทีม Red Team ทีม Red Team จะทำการทดสอบเจาะระบบในขอบเขตที่จำกัดบนสภาพแวดล้อม Production-Ready เพื่อทวนสอบว่ามาตรการป้องกันต่าง ๆ เช่น WAF, IAM Roles, และ Network Segmentation สามารถป้องกันการโจมตีในสถานการณ์จริงได้หรือไม่ การทดสอบในระยะนี้มีความสำคัญและควรแบ่งประเภทให้ชัดเจนตามสภาพแวดล้อม

๓.๓) การทดสอบในสภาพแวดล้อมเสมือนจริง

๓.๓.๑) วัตถุประสงค์ เพื่อค้นหาช่องโหว่และความผิดพลาดอย่างเต็มรูปแบบในสภาพแวดล้อมที่ใกล้เคียงของจริงที่สุด แต่ไม่ส่งผลกระทบต่อผู้ใช้

๓.๓.๒) ตัวอย่าง การทดสอบเจาะระบบ การทดสอบด้วยตัวอย่างที่เป็นปฏิปักษ์ การทดสอบภายใต้ภาระหนัก

๓.๔) การทวนสอบในสภาพแวดล้อมใช้งานจริง

๓.๔.๑) วัตถุประสงค์ เพื่อยืนยันว่าการตั้งค่าและมาตรการป้องกันบนสภาพแวดล้อมจริงทำงานได้ถูกต้องตามที่ออกแบบไว้ โดยต้องเป็นการทดสอบที่ไม่ส่งผลกระทบต่อบริการ

๓.๔.๒) ตัวอย่าง การตรวจสอบความถูกต้องของการตั้งค่า การทดสอบการเชื่อมต่อเบื้องต้นหลังการติดตั้ง การทดสอบโดยทีม Red Team ในขอบเขตจำกัดเพื่อทวนสอบกลไกป้องกัน เช่น WAF หรือ IAM Roles

๓.๕) การทวนสอบกระบวนการนำขึ้นระบบ ตรวจสอบว่ากระบวนการนำโคดและโมเดลขึ้นสู่การใช้งานจริง เป็นไปตามกระบวนการบริหารจัดการการเปลี่ยนแปลงที่ได้รับอนุมัติ และมีกลไกป้องกันการนำเวอร์ชันที่ไม่ผ่านการทดสอบขึ้นใช้งาน (อ้างอิง ISO/IEC 27002:2022 A8.19)

๔) ผลลัพธ์ที่คาดหวัง

เมื่อสิ้นสุดระยะที่ ๔ ระบบปัญญาประดิษฐ์ จะพร้อมสำหรับการให้บริการอย่างเป็นทางการ โดยมีผลลัพธ์ที่สำคัญคือ

๔.๑) ระบบปัญญาประดิษฐ์ ที่ทำงานบนสภาพแวดล้อมใช้งานจริงที่ผ่านการเสริมความแข็งแกร่ง

๔.๒) หลักฐานการติดตั้งและเปิดใช้งานมาตรการควบคุมความปลอดภัยทั้งหมด เช่น การเข้ารหัสลับ การจัดการคีย์ การตั้งค่าเครือข่าย

๔.๓) รายงานผลการทดสอบความปลอดภัยขั้นสุดท้ายบนสภาพแวดล้อมใช้งานจริง

๔.๔) การอนุมัติให้เริ่มใช้งานระบบอย่างเป็นทางการ

เอกสารอ้างอิง:

๑) ISO/IEC 27002:2022 security controls under “#Secure_configuration” and “#System_and_network_security” attribute, more specifically 8.9 Configuration management, 8.19 Installation of software on operational systems

๒) ISO/IEC 42001:2023 A.6.2.5 AI system deployment

๓.๒.๖ ระยะที่ ๕ การดำเนินงานและการบำรุงรักษาอย่างมั่นคงปลอดภัย

ระยะการดำเนินงานและการบำรุงรักษาเป็นขั้นตอนต่อเนื่องที่เริ่มต้นทันทีหลังจากระบบปัญญาประดิษฐ์ถูกนำขึ้นใช้งานจริง

วัตถุประสงค์หลักของระยะนี้คือ การรับประกันว่าคุณสมบัติด้านความมั่นคงปลอดภัยของระบบปัญญาประดิษฐ์จะถูกรักษาและพัฒนาให้ดีขึ้นอย่างต่อเนื่องตลอดวงจรชีวิตของระบบปัญญาประดิษฐ์ การใช้งานโดยมุ่งเน้นการเฝ้าระวังเชิงรุก การบริหารจัดการการเปลี่ยนแปลงที่รัดกุม และการเตรียมความพร้อมเพื่อตอบสนองต่อเหตุการณ์ที่ไม่คาดฝัน

ความมั่นคงปลอดภัยในระยะนี้ไม่ใช่สภาวะที่คงที่ แต่เป็นกระบวนการที่ต้องปรับตัวและพัฒนาอย่างไม่หยุดนิ่งเพื่อรับมือกับภัยคุกคาม ความต้องการทางธุรกิจ และสภาพแวดล้อมที่เปลี่ยนแปลงไป

๑) การเฝ้าระวังและประเมินสถานะ

เป็นกระบวนการเฝ้าระวังระบบอย่างต่อเนื่อง เพื่อให้ทราบถึงสถานะปัจจุบัน ตรวจจับความผิดปกติ และประเมินความเสี่ยงเชิงรุก

๑.๑) การบันทึกข้อมูลเชิงรุกและการเฝ้าระวังพฤติกรรม

๑.๑.๑) หลักการ วางรากฐานการเฝ้าระวังด้วยการบันทึกข้อมูลที่ละเอียดและครอบคลุมทั้งในระดับระบบข้อมูลนำเข้า ผลลัพธ์ และพฤติกรรมของโมเดล เพื่อใช้ในการวิเคราะห์เชิงรุก ตรวจจับการโจมตี และเฝ้าระวังการเสื่อมถอยของประสิทธิภาพโมเดล (อ้างอิง ISO/IEC 27002:2022 A.8.16)

๑.๑.๒) ตัวอย่าง

▪ **การเฝ้าระวังการโจมตี** บันทึกข้อมูลและวิเคราะห์ Prompt ทั้งหมดเพื่อตรวจจับความพยายามในการทำ Prompt Injection หรือ Jailbreaking และแจ้งเตือนทีม Security Operations Center (SOC) ทันที

▪ **การเฝ้าระวังการเสื่อมถอยของโมเดล** อาศัยข้อมูลที่บันทึกไว้เพื่อติดตามภาวะความเบี่ยงเบนของข้อมูล (Data Drift) และความสัมพันธ์ของข้อมูลที่เปลี่ยนไป (Concept Drift) เพื่อแจ้งเตือนให้เริ่มกระบวนการฝึกสอนโมเดลใหม่เมื่อจำเป็น ความเบี่ยงเบนของข้อมูลและความสัมพันธ์ของข้อมูลที่เปลี่ยนไปอย่างเป็นระบบควรใช้กระบวนการดังนี้

- **สร้างสถานะพื้นฐาน** หลังจากนำโมเดลขึ้นใช้งานจริงให้วัดและบันทึกค่าสถิติของข้อมูลนำเข้าและตัวชี้วัดประสิทธิภาพของโมเดล เช่น ความแม่นยำ ค่าความเชื่อมั่น เพื่อใช้เป็น "สถานะปกติ"
- **เฝ้าระวังอย่างต่อเนื่อง** ใช้เครื่องมือทางสถิติเพื่อเปรียบเทียบข้อมูลที่เข้ามาใหม่กับค่าพื้นฐาน และติดตามประสิทธิภาพของโมเดลอย่างสม่ำเสมอ
- **กำหนดเกณฑ์แจ้งเตือน** กำหนดระดับความเบี่ยงเบนที่ยอมรับได้ หากค่าที่เฝ้าระวังเบี่ยงเบนเกินเกณฑ์ที่กำหนด ให้ระบบสร้างการแจ้งเตือนอัตโนมัติ
- **วิเคราะห์และฝึกสอนใหม่** เมื่อได้รับการแจ้งเตือน ให้นักวิทยาศาสตร์ข้อมูลเข้าวิเคราะห์หาสาเหตุ หากยืนยันได้ว่าโมเดลเสื่อมประสิทธิภาพจริง ให้เริ่มกระบวนการฝึกสอนโมเดลใหม่ตามที่กำหนดไว้

๑.๒) การเฝ้าระวังและรับมือความเสี่ยงจากระบบปัญญาประดิษฐ์ของบุคคลที่สามเมื่อใช้บริการระบบปัญญาประดิษฐ์โดยผู้ให้บริการภายนอก องค์กรไม่สามารถควบคุมระบบภายในได้โดยตรง แต่สามารถบริหารความเสี่ยงได้โดย

๑.๒.๑) การเฝ้าระวังข้อมูลนำเข้าและผลลัพธ์ เฝ้าระวังข้อมูลที่นำเข้าระบบปัญญาประดิษฐ์และผลลัพธ์ที่ได้รับกลับมา เพื่อตรวจจับความผิดปกติ เนื้อหาที่ไม่เหมาะสม หรือสัญญาณการรั่วไหลของข้อมูล

๑.๒.๒) การเฝ้าระวังประสิทธิภาพบริการ ติดตามตัวชี้วัด เช่น อัตราความผิดพลาดหรือเวลาในการตอบสนอง หากค่าเหล่านี้เปลี่ยนแปลงอย่างมีนัยสำคัญอาจเป็นสัญญาณว่าเกิดปัญหาที่ฝั่งผู้ให้บริการ

๑.๒.๓) การเตรียมแผนเผชิญเหตุ จัดทำคู่มือปฏิบัติสำหรับกรณีที่บริการระบบปัญญาประดิษฐ์ขัดข้องหรือถูกโจมตี โดยต้องมีขั้นตอนการติดต่อผู้ให้บริการ การสลับไปใช้ระบบสำรอง (ถ้ามี) และการสื่อสารกับผู้ใช้งาน

๑.๓) การบริหารจัดการประสิทธิภาพและความพร้อมใช้งาน

๑.๓.๑) หลักการ วางแผนสถาปัตยกรรมให้พร้อมรองรับการใช้งานที่เพิ่มขึ้นเพื่อรักษาความพร้อมใช้งานของระบบและป้องกันปัญหาระบบไม่สามารถให้บริการได้

๑.๓.๒) ตัวอย่าง

๑.๓.๒.๑) การทดสอบภาระงาน จำลองการใช้งานจำนวนมากอย่างสม่ำเสมอเพื่อหาจุดคอขวด

๑.๓.๒.๒) การลดและขยายระบบอัตโนมัติ ออกแบบสถาปัตยกรรมให้สามารถเพิ่มและลดทรัพยากรได้อัตโนมัติตามปริมาณการใช้งานจริง

๑.๔) การบริหารจัดการช่องโหว่ทางเทคนิค

๑.๔.๑) หลักการ ค้นหา ประเมิน และแก้ไขช่องโหว่ในทุกองค์ประกอบของระบบอย่างต่อเนื่อง (อ้างอิง ISO/IEC 27002:2022 A8.8)

๑.๔.๒) ตัวอย่าง

- **การสแกนอัตโนมัติ** ใช้เครื่องมือ SCA สแกนช่องโหว่ในไลบรารีและใช้ DAST/SAST สแกนโคดที่พัฒนาขึ้นเอง
- **การบริหารจัดการแพตช์** จัดทำกระบวนการประเมินความรุนแรงและอัปเดตแพตช์ความปลอดภัยตามลำดับความสำคัญ โดยต้องผ่านการทดสอบก่อนเสมอ

๒) การบริหารจัดการการเปลี่ยนแปลง

เป็นกระบวนการปรับเปลี่ยนระบบอย่างมีแบบแผนและรัดกุม เพื่อให้แน่ใจว่าทุกการเปลี่ยนแปลงจะไม่นำมาซึ่งช่องโหว่ใหม่ (อ้างอิง ISO/IEC 27002:2022 8.32)

๒.๑) หลักการ การเปลี่ยนแปลงใด ๆ ต่อระบบ ไม่ว่าจะเป็นการอัปเดตโคด การฝึกสอนโมเดลใหม่ หรือการเพิ่มความสามารถใหม่ ๆ จะต้องผ่านกระบวนการจัดการที่มั่นคงปลอดภัยและครบวงจร

๒.๒) ตัวอย่างวงจรการอัปเดต

๒.๒.๑) การฝึกสอนโมเดลใหม่ เมื่อข้อมูลเปลี่ยนไปและต้องฝึกโมเดลใหม่ โมเดลเวอร์ชันใหม่จะต้องผ่านกระบวนการ ระยะที่ ๓ การทวนสอบด้านความมั่นคงปลอดภัยทั้งหมดอีกครั้งก่อนนำขึ้นใช้งาน

๒.๒.๒) การอัปเดตแพตช์ความปลอดภัย เมื่อมีการออกแพตช์สำหรับไลบรารี การอัปเดตจะต้องผ่านการทดสอบการถดถอย (Regression Testing) เพื่อให้แน่ใจว่าไม่ส่งผลกระทบต่อความแม่นยำของโมเดล

๒.๒.๓) การเพิ่มขีดความสามารถใหม่ เมื่อมีการร้องขอฟีเชอร์ใหม่ การพัฒนาจะต้องย้อนกลับไปสู่ ระยะที่ ๑ การออกแบบอย่างมั่นคงปลอดภัย เพื่อทำแบบจำลองภัยคุกคามสำหรับฟีเชอร์นั้น ๆ ก่อนดำเนินการพัฒนาและทดสอบ

๓) การตอบสนองต่อเหตุการณ์และธรรมาภิบาล

เป็นกระบวนการรับมือกับเหตุการณ์ที่ไม่คาดฝัน และกำกับดูแลให้ระบบดำเนินงานภายใต้กรอบของกฎหมายและข้อบังคับ

๓.๑) การจัดการและการตอบสนองต่อเหตุการณ์

๓.๑.๑) หลักการ เตรียมพร้อมรับมือกับเหตุการณ์ด้านความมั่นคงปลอดภัย โดยมีแผน ทีม คู่มือปฏิบัติที่ชัดเจน และต้องมีการซ้อมแผนอย่างสม่ำเสมอ เพื่อให้แน่ใจว่าทีมสามารถตอบสนองได้อย่างมีประสิทธิภาพเมื่อเกิดเหตุการณ์จริง (อ้างอิง ISO/IEC 27002:2022 A5.26)

๓.๑.๒) ตัวอย่างคู่มือปฏิบัติ หากตรวจพบ Evasion Attack ต้องมีขั้นตอนที่ชัดเจนในการ ๑) จำกัดความเสียหาย ๒) กำจัด และ ๓) กู้คืนระบบกลับสู่สภาวะปกติ

๓.๒) การกำกับดูแลและการปรับตัวตามกฎระเบียบ

๓.๒.๑) หลักการ กำกับดูแลให้ระบบปัญญาประดิษฐ์ทำงานสอดคล้องกับกฎหมาย ข้อบังคับ และนโยบายขององค์กร ซึ่งอาจเปลี่ยนแปลงได้ตลอดเวลา

๓.๒.๒) ตัวอย่าง:

- **การตรวจสอบการปฏิบัติตามกฎระเบียบ:** ตรวจสอบระบบเป็นประจำ เพื่อให้แน่ใจว่ายังคงสอดคล้องกับกฎระเบียบที่เกี่ยวข้อง เช่น กฎหมายคุ้มครองข้อมูลส่วนบุคคล หรือกฎระเบียบด้านปัญญาประดิษฐ์ใหม่ ๆ
- **กลไกการปรับตัว:** ออกแบบระบบให้ยืดหยุ่นพอที่จะปรับเปลี่ยนได้เมื่อมีกฎหมายใหม่ประกาศใช้ เช่น รองรับสิทธิ์ในการขอลบข้อมูลออกจากชุดข้อมูลฝึกสอน

๔) แหล่งข้อมูลสนับสนุนการรับมือเหตุการณ์

เพื่อให้เท่าทันภัยคุกคามรูปแบบใหม่ องค์กรควรติดตามข้อมูลจากแหล่งข่าวกรองที่เป็นที่ยอมรับในระดับสากล เช่น

๔.๑) หน่วยงานความมั่นคงปลอดภัยไซเบอร์ของสหภาพยุโรปซึ่งเผยแพร่รายงาน งานวิจัย และกรอบการดำเนินงานด้านความมั่นคงปลอดภัยระบบปัญญาประดิษฐ์อย่างสม่ำเสมอ

๔.๒) **OWASP AI Exchange** โครงการที่ขับเคลื่อนโดยชุมชนภายใต้ OWASP Foundation ทำหน้าที่รวบรวมเครื่องมือ ชุดข้อมูล และองค์ความรู้เกี่ยวกับภัยคุกคามและการทดสอบความปลอดภัยระบบปัญญาประดิษฐ์

๕) ผลลัพธ์ที่คาดหวัง

ระยะนี้เป็นกระบวนการต่อเนื่องไม่มีจุดสิ้นสุด ตราบเท่าที่ยังคงมีการใช้งานระบบปัญญาประดิษฐ์นั้นอยู่ โดยมีผลลัพธ์ที่สำคัญคือ

๕.๑) ระบบเฝ้าระวังและแจ้งเตือนที่ทำงานอย่างมีประสิทธิภาพ

๕.๒) รายงานสรุปประสิทธิภาพ ความมั่นคงปลอดภัย และความสามารถในการรองรับภาระงานของโมเดลเป็นประจำ

๕.๓) บันทึกการเปลี่ยนแปลง บันทึกการจัดการช่องโหว่ทั้งหมดที่เกิดขึ้นกับระบบ รายงานสรุปเหตุการณ์ และบทเรียนที่ได้รับเพื่อนำไปปรับปรุงต่อไป

๕.๔) เอกสารยืนยันการปฏิบัติตามกฎระเบียบที่เป็นปัจจุบัน

๕.๕) การคงไว้ซึ่งสถานะความมั่นคงปลอดภัยที่แข็งแกร่งของระบบปัญญาประดิษฐ์ตลอดอายุการใช้งาน

เอกสารอ้างอิง:

ISO/IEC 27002:2022 security controls under “#Secure_configuration” and “#System_and_network_security” attribute, more specifically 8.9 Configuration management, 8.19 Installation of software on operational systems

๓.๒.๗ ระยะที่ ๖ การกำจัดและทำลาย

ระยะการกำจัดและทำลายเป็นขั้นตอนสุดท้ายของวงจรชีวิตของระบบปัญญาประดิษฐ์ที่มั่นคงปลอดภัย ซึ่งมักถูกมองข้ามแต่มีความสำคัญอย่างยิ่งต่อการบริหารจัดการความเสี่ยง

วัตถุประสงค์หลักของระยะนี้คือ การปลดระวางระบบปัญญาประดิษฐ์ และสินทรัพย์ที่เกี่ยวข้องทั้งหมดอย่างเป็นระบบและมั่นคงปลอดภัย เพื่อป้องกันการรั่วไหลของข้อมูล การละเมิดทรัพย์สินทางปัญญา และการเกิดความเสี่ยงคงค้างในอนาคต

กระบวนการนี้ต้องพิจารณาข้อกำหนดจากหลายแหล่งทั้งมาตรฐานสากลด้านการจัดการความมั่นคงปลอดภัย เช่น ISO/IEC 27001 และ ISO/IEC 42001 กรอบการทำงานด้านความมั่นคงปลอดภัยของแอปพลิเคชัน เช่น กรอบการทำงานโดย OWASP และกฎหมายคุ้มครองข้อมูลส่วนบุคคลที่เข้มงวด เพื่อรับประกันว่าเมื่อระบบปัญญาประดิษฐ์สิ้นสุดอายุการใช้งาน จะไม่ทิ้งร่องรอยทางดิจิทัลที่อาจกลายเป็นช่องโหว่ให้แก่องค์กรได้

๑) การกำจัดและทำลายข้อมูลและโมเดลอย่างมั่นคงปลอดภัย

เป็นกระบวนการทำให้สินทรัพย์ดิจิทัลทั้งหมดที่เกี่ยวข้องกับระบบปัญญาประดิษฐ์ ไม่สามารถกู้คืนหรือเข้าถึงได้อีกต่อไปอย่างถาวร ซึ่งต้องเป็นไปตามนโยบายการจัดเก็บและทำลายข้อมูลขององค์กรและข้อกฎหมายที่เกี่ยวข้อง (อ้างอิง ISO/IEC 27002:2022 A8.10 Information deletion)

๑.๑) บริบทของปัญญาประดิษฐ์ สินทรัพย์ดิจิทัลของปัญญาประดิษฐ์มีความหลากหลายและละเอียดอ่อน ตั้งแต่ข้อมูลฝึกสอนไปจนถึงพารามิเตอร์ของโมเดล

๑.๒) ตัวอย่างมาตรการควบคุม

๑.๒.๑) ชุดข้อมูลฝึกสอน โดยเฉพาะข้อมูลที่มีข้อมูลส่วนบุคคล หรือความลับทางการค้า การลบไฟล์แบบปกตินั้นไม่เพียงพอ จะต้องใช้วิธีการทำลายข้อมูลอย่างปลอดภัย เช่น การเขียนทับข้อมูลซ้ำหลายครั้ง หรือการทำลายคีย์ที่ใช้เข้ารหัสลับข้อมูล ซึ่งจะทำให้ข้อมูลที่เข้ารหัสลับไว้ไม่สามารถอ่านได้อีกต่อไป

๑.๒.๒) โมเดลปัญญาประดิษฐ์และ Checkpoints โมเดลที่ฝึกสอนแล้วคือ ทรัพย์สินทางปัญญาที่มีมูลค่าสูง รวมถึงไฟล์ Checkpoints ที่ถูกบันทึกไว้ระหว่างการฝึกสอนก็อาจสามารถใช้ทำวิศวกรรมย้อนกลับเพื่ออนุมานข้อมูลบางส่วนได้ สินทรัพย์เหล่านี้ทั้งหมดจะต้องถูกกำจัดด้วยวิธีการทำลายข้อมูลที่ปลอดภัยเช่นเดียวกับชุดข้อมูลฝึกสอน

๑.๒.๓) ไฟล์บันทึก ไฟล์บันทึกการทำงานของระบบอาจมีข้อมูลที่ละเอียดอ่อน เช่น Prompt ที่ผู้ใช้ป้อนเข้ามา หรือผลลัพธ์ที่โมเดลสร้างขึ้น จะต้องถูกทำลายอย่างปลอดภัยตามนโยบายการเก็บรักษาข้อมูล

๑.๓) เทคนิคการทำลายข้อมูลที่นำไปปฏิบัติได้

๑.๓.๑) การลบโดยใช้การเข้ารหัสลับ (Cryptographic Erasure หรือ Crypto Shredding)

หลักการ เป็นวิธีที่ทันสมัยและมีประสิทธิภาพสูงสุดสำหรับข้อมูลขนาดใหญ่และระบบคลาวด์ โดยข้อมูลทั้งหมดจะถูกเข้ารหัสลับไว้ตั้งแต่แรก เมื่อต้องการทำลายข้อมูล เราเพียงแค่ทำลายกุญแจเข้ารหัสที่ใช้ในการถอดรหัสข้อมูลนั้นทิ้งไปอย่างถาวร

ข้อดี เมื่อไม่มีกุญแจ ข้อมูลที่เข้ารหัสลับไว้จะกลายเป็นข้อมูลขยะที่ไม่สามารถอ่านได้อีกต่อไป เป็นการทำลายข้อมูลจำนวนมหาศาลได้อย่างรวดเร็วและสมบูรณ์ เหมาะสำหรับ Cloud Storage หรือ Big Data Lake

๑.๓.๒) การเขียนทับข้อมูล

หลักการ เขียนข้อมูลแบบสุ่มทับลงบนพื้นที่จัดเก็บข้อมูลเดิมซ้ำ ๆ หลายรอบ เพื่อให้ข้อมูลดั้งเดิมไม่สามารถกู้คืนได้

ข้อควรระวัง เหมาะสำหรับฮาร์ดดิสก์แบบจานแม่เหล็ก (HDD) แต่อาจมีประสิทธิภาพลดลงหรือไม่เหมาะสมกับ Solid-State Drives (SSD) เนื่องจากเทคโนโลยีการจัดเก็บข้อมูลที่ซับซ้อนกว่า

๑.๓.๓) การทำลายทางกายภาพ

หลักการ การทำลายตัวสื่อบันทึกข้อมูลโดยตรง เป็นวิธีที่ปลอดภัยที่สุดสำหรับอุปกรณ์ที่หมดอายุการใช้งาน

ตัวอย่าง การบดย่อย (Shredding) การทำลายด้วยสนามแม่เหล็กแรงสูง (Degaussing) หรือการหลอมละลาย (Pulverizing)

เอกสารอ้างอิง:

ISO/IEC 27002:2022 (A8.10 Information deletion), ISO/IEC 27040 (Storage Security), EU GDPR (Article 17), OWASP Top 10 for LLMs (LLM06)

๒) การเรียกคืนทรัพยากรและยกเลิกสิทธิ์การเข้าถึง

เป็นกระบวนการปลดระวางโครงสร้างพื้นฐานทั้งฮาร์ดแวร์และซอฟต์แวร์ที่เคยรองรับการทำงานของระบบปัญญาประดิษฐ์ พร้อมทั้งเพิกถอนสิทธิ์การเข้าถึงและข้อมูลยืนยันตัวตนทั้งหมด เพื่อป้องกันการเกิดบัญชีผู้ใช้ร้างหรือทรัพยากรที่ไม่มีผู้ดูแลซึ่งอาจกลายเป็นช่องโหว่ได้

๒.๑) บริบทของปัญญาประดิษฐ์ ระบบปัญญาประดิษฐ์มักใช้ทรัพยากรประมวลผลและจัดเก็บข้อมูลขนาดใหญ่ ซึ่งต้องจัดการอย่างระมัดระวังเมื่อเลิกใช้งาน

๒.๒) ตัวอย่างมาตรการควบคุม

๒.๒.๑) ทรัพยากรฮาร์ดแวร์และสื่อบันทึกข้อมูล เซิร์ฟเวอร์และอุปกรณ์จัดเก็บข้อมูลที่ใช้ในการฝึกสอนโมเดลปัญญาประดิษฐ์จะต้องผ่านกระบวนการล้างข้อมูลอย่างปลอดภัยก่อนนำไปใช้ใหม่ หรือหากต้องการทิ้งจะต้องทำลายทางกายภาพ เช่น การทำลายด้วยสนามแม่เหล็กหรือการบดย่อย เพื่อป้องกันการกู้คืนข้อมูล (อ้างอิง ISO/IEC 27002:2022 7.10 Storage media) ตามเทคนิคการทำลายข้อมูลที่เหมาะสม

เอกสารอ้างอิง:

ISO/IEC 27002:2022 (A7.10 Storage Media)

๒.๒.๒) ทรัพยากรบนคลาวด์และ API Keys ดำเนินการตามแผนการปลดระวางที่กำหนดไว้ เช่น การลบ Cloud Storage Buckets ทั้งหมด การยุติการทำงานของ Virtual Machines และ Containers การลบ IAM Roles และ Service Accounts ที่สร้างขึ้นสำหรับระบบปัญญาประดิษฐ์โดยเฉพาะ และการลบ API Keys ทั้งหมดออกจากระบบจัดการข้อมูลลับ เพื่อป้องกันปัญหาการควบคุมสิทธิ์การเข้าถึงที่บกพร่อง (Broken Access Control)

เอกสารอ้างอิง:

ISO/IEC 27002:2022 (A5.15 Access control, A5.18 Access rights), OWASP Top 10 (A01:2021)

๒.๒.๓) บริการและ API ของบุคคลที่สาม ยกเลิกสัญญาการใช้บริการกับผู้ให้บริการข้อมูลภายนอก และเพิกถอน API keys ทั้งหมดที่ระบบปัญญาประดิษฐ์เคยใช้เชื่อมต่อ เพื่อป้องกันการใช้งานที่ไม่ได้รับอนุญาตในอนาคต

๓) การปฏิบัติตามข้อกำหนดและการจัดเก็บเอกสาร

เป็นขั้นตอนสุดท้ายในเชิงบริหารจัดการ เพื่อรับประกันการปฏิบัติตามข้อกำหนดทางกฎหมายและกฎระเบียบ พร้อมทั้งเก็บรักษาหลักฐานการดำเนินงานตลอดวงจรชีวิตของระบบปัญญาประดิษฐ์ไว้เพื่อการตรวจสอบในอนาคต

๓.๑) บริบทของปัญญาประดิษฐ์ การมีหลักฐานที่ตรวจสอบได้เป็นสิ่งสำคัญอย่างยิ่ง โดยเฉพาะระบบปัญญาประดิษฐ์ที่ส่งผลกระทบต่อบุคคลในวงกว้าง

๓.๒) ตัวอย่างมาตรการควบคุม

๓.๒.๑) การจัดทำใบรับรองการทำลาย สร้างและจัดเก็บเอกสารที่เป็นทางการเพื่อรับรองการทำลายข้อมูลและโมเดล เอกสารนี้ควรระบุรายละเอียดว่าสินทรัพย์ใดถูกทำลาย ใช้วิธีการใด วันที่และเวลา และผู้ที่รับผิดชอบ ซึ่งเป็นหลักฐานสำคัญสำหรับการตรวจสอบด้านการคุ้มครองข้อมูลส่วนบุคคล

๓.๒.๒) การจัดเก็บเอกสารตลอดวงจรชีวิตของระบบปัญญาประดิษฐ์ เอกสารสำคัญทั้งหมดที่สร้างขึ้นตั้งแต่ระยะที่ ๐ ถึง ๕ เช่น รายงานการประเมินความเสี่ยง แบบจำลองภัยคุกคาม รายงานผลการทดสอบเจาะระบบ และรายงานการรับมือเหตุการณ์จะต้องถูกจัดเก็บในที่ที่ปลอดภัยตามนโยบายขององค์กร เพื่อใช้เป็นหลักฐานแสดงความรับผิดชอบ หากมีการฟ้องร้องหรือการตรวจสอบเกิดขึ้นในภายหลัง

เอกสารอ้างอิง:

ISO/IEC 27002:2022 (A5.33 Protection of records), ISO/IEC 42001 (Clause 7.5 Documented information), EU AI Act (Record-keeping obligations)

๓.๒.๓) การส่งคืนสินทรัพย์ จัดทำเอกสารการส่งคืนหรือการกำจัดสินทรัพย์ทั้งหมดที่เกี่ยวข้องกับโครงการอย่างเป็นทางการตามนโยบายขององค์กร

เอกสารอ้างอิง:

ISO/IEC 27002:2022 (A5.11 Return of assets)

๔) ผลลัพธ์ที่คาดหวัง

เมื่อสิ้นสุดระยะที่ ๖ ระบบปัญญาประดิษฐ์จะถูกปลดระวางอย่างสมบูรณ์และปลอดภัย โดยมีผลลัพธ์ที่สำคัญคือ

๔.๑) ใบรับรองการทำลายข้อมูลและโมเดล

๔.๒) หลักฐานการเคลียร์ทรัพยากรฮาร์ดแวร์และคลาวด์อย่างมั่นคงปลอดภัย

๔.๓) คลังเอกสารฉบับสมบูรณ์ของโครงการที่จัดเก็บอย่างปลอดภัย

๔.๔) การปิดโครงการระบบปัญญาประดิษฐ์อย่างเป็นทางการ พร้อมการยืนยันว่าความเสี่ยงคงค้างทั้งหมดได้รับการจัดการเรียบร้อยแล้ว

เอกสารอ้างอิง:

ISO/IEC 27002:2022 5.11 Return of assets, 5.33 Protection of records, 7.10 Storage media, 8.10 Information deletion

บทที่ ๔
การกำกับดูแลและการบริหารความเสี่ยงสำหรับระบบปัญญาประดิษฐ์

หัวใจสำคัญของบทนี้ตั้งอยู่บน**หลักการกำกับดูแล การบริหารความเสี่ยง และการปฏิบัติตามข้อกำหนด (Governance, Risk Management, and Compliance: GRC)** ซึ่งเป็นเสาหลักของการประยุกต์ใช้ปัญญาประดิษฐ์อย่างมีความรับผิดชอบและยั่งยืน การจัดตั้งโครงสร้างธรรมาภิบาลที่ชัดเจนและการกำหนดบทบาทหน้าที่ที่รัดกุมจะช่วยให้องค์กรปฏิบัติตามข้อบังคับทางกฎหมาย สร้างความไว้วางใจให้แก่ผู้มีส่วนได้ส่วนเสียและส่งเสริมวัฒนธรรมความมั่นคงปลอดภัยทั่วทั้งองค์กร

โดยแนวปฏิบัตินี้จะทำหน้าที่มอบกรอบการดำเนินงานและเครื่องมือที่จำเป็น แต่สิ่งสำคัญที่สุดคือ**องค์กรและผู้ใช้งานจะต้องเป็นผู้ประเมินความเสี่ยงในบริบทการใช้งานของตนเอง** เนื่องจากแต่ละระบบปัญญาประดิษฐ์มีความจำเพาะทั้งในด้านวัตถุประสงค์ สถาปัตยกรรม และข้อมูล ความรับผิดชอบในการระบุ วิเคราะห์ และจัดการความเสี่ยงจึงเป็นขององค์กรผู้ใช้งานโดยตรง

๔.๑ การกำหนดบทบาทหน้าที่และความรับผิดชอบ

รากฐานที่สำคัญที่สุดของการกำกับดูแลปัญญาประดิษฐ์ที่มีประสิทธิภาพ คือ การกำหนดบทบาท ความรับผิดชอบ และอำนาจหน้าที่ที่เกี่ยวข้องกับความมั่นคงปลอดภัยของระบบปัญญาประดิษฐ์อย่างชัดเจนและเป็นลายลักษณ์อักษร การขาดความชัดเจนในเรื่องนี้มักนำไปสู่ช่องว่างด้านความรับผิดชอบ ซึ่งเป็นบ่อเกิดของความเสี่ยงที่ร้ายแรง

การกำกับดูแลและการบริหารความเสี่ยงของระบบปัญญาประดิษฐ์ เป็นองค์ประกอบสำคัญที่ช่วยให้การใช้งานปัญญาประดิษฐ์ในองค์กรมีความมั่นคงปลอดภัย โปร่งใส และตรวจสอบได้ โดยต้องบูรณาการเข้ากับ Enterprise Risk Management (ERM) และสอดคล้องกับกฎหมายและมาตรฐานที่เกี่ยวข้อง เช่น ISO/IEC 42001:2023 ISO/IEC 23894:2023 ISO/IEC 27001:2022 และกฎหมายไทย เช่น พระราชบัญญัติคุ้มครองข้อมูลส่วนบุคคล และพระราชบัญญัติการรักษาความมั่นคงปลอดภัยไซเบอร์

องค์กรควรจัดตั้งโครงสร้างบุคลากรที่ครอบคลุมตั้งแต่ระดับยุทธศาสตร์ไปจนถึงระดับปฏิบัติการ โดยแต่ละบทบาทมีหน้าที่ที่จำเพาะเจาะจงต่อวงจรชีวิตของระบบปัญญาประดิษฐ์ ดังตัวอย่างต่อไปนี้

๑) คณะกรรมการหรือผู้บริหารระดับสูง

บทบาท กำหนดทิศทางเชิงกลยุทธ์และนโยบายระดับองค์กร และเป็นผู้รับผิดชอบสูงสุดต่อความเสี่ยงทั้งหมดที่เกี่ยวข้องกับปัญญาประดิษฐ์

ความรับผิดชอบด้านความมั่นคงปลอดภัยปัญญาประดิษฐ์

๑.๑) กำหนดนโยบาย มาตรฐาน และทิศทาง

๑.๒) รับผิดชอบตรวจสอบความสอดคล้องกับกฎหมายและข้อบังคับ

๑.๓) กำหนดระดับความเสี่ยงที่องค์กรยอมรับได้ สำหรับโครงการปัญญาประดิษฐ์ต่าง ๆ

๑.๓.๑) อนุมัตินโยบายธรรมาภิบาลและการใช้งานปัญญาประดิษฐ์อย่างมีจริยธรรมขององค์กร

๑.๓.๒) จัดสรรทรัพยากร (งบประมาณ บุคลากร เทคโนโลยี) ให้เพียงพอต่อการดำเนินงานด้านความมั่นคงปลอดภัยของปัญญาประดิษฐ์

๑.๔) ทบทวนรายงานความเสี่ยงด้านปัญญาประดิษฐ์ที่สำคัญเป็นประจำ

๒) คณะกรรมการกำกับดูแลปัญญาประดิษฐ์

บทบาท คณะกรรมการชุดนี้ทำหน้าที่เป็นองค์กรกำกับดูแลระดับยุทธศาสตร์ กำหนดกรอบธรรมาภิบาล นโยบาย และมาตรฐานการใช้งานปัญญาประดิษฐ์ทั้งหมดในองค์กร เพื่อให้มั่นใจว่าการพัฒนาและใช้งานปัญญาประดิษฐ์ สอดคล้องกับทิศทางธุรกิจ ค่านิยมองค์กร และหลักจริยธรรม

๒.๑) ความรับผิดชอบด้านความมั่นคงปลอดภัยปัญญาประดิษฐ์

๒.๒) พิจารณาและอนุมัติกรอบธรรมาภิบาลปัญญาประดิษฐ์ขององค์กร

๒.๓) ทบทวนและให้ความเห็นชอบโครงการปัญญาประดิษฐ์ที่มีความเสี่ยงสูงหรือส่งผลกระทบในวงกว้าง

๒.๔) ส่งเสริมวัฒนธรรมการใช้งานปัญญาประดิษฐ์อย่างมีความรับผิดชอบทั่วทั้งองค์กร

๓) คณะทำงานด้านความเสี่ยงปัญญาประดิษฐ์

บทบาท คณะทำงานที่ประกอบด้วยผู้แทนจากหน่วยงานหลัก เช่น ความเสี่ยง ความมั่นคงปลอดภัย กฎหมาย และวิทยาศาสตร์ข้อมูล ทำหน้าที่ในระดับยุทธวิธีเพื่อบริหารจัดการความเสี่ยงที่เกี่ยวข้องกับปัญญาประดิษฐ์โดยเฉพาะ

ความรับผิดชอบด้านความมั่นคงปลอดภัยปัญญาประดิษฐ์

๓.๑) ดำเนินการประเมินและติดตามความเสี่ยงที่จำเพาะเจาะจงกับระบบปัญญาประดิษฐ์อย่างสม่ำเสมอ

๓.๒) รับผิดชอบในการบูรณาการความเสี่ยงด้านปัญญาประดิษฐ์เข้ากับ ทะเบียนความเสี่ยงรวมขององค์กร

๓.๓) รายงานผลการประเมินและสถานะความเสี่ยงต่อคณะกรรมการกำกับดูแลฯ และผู้บริหารระดับสูง

๔) เจ้าของผลิตภัณฑ์และเจ้าของระบบปัญญาประดิษฐ์

บทบาท ผู้รับผิดชอบหลักสำหรับระบบปัญญาประดิษฐ์ระบบใดระบบหนึ่งตลอดวงจรชีวิตของระบบปัญญาประดิษฐ์ โดยทั่วไปบทบาทนี้จะมาจากฝั่งธุรกิจเพื่อให้การกำกับดูแลเป็นไปในลักษณะจากบนลงล่าง และเป็นสะพานเชื่อมที่สำคัญระหว่างทีมธุรกิจและทีมเทคนิค

ความรับผิดชอบด้านความมั่นคงปลอดภัยปัญญาประดิษฐ์

๔.๑) รับผิดชอบต่อผลกระทบและความเสี่ยงทั้งหมดที่เกิดจากระบบปัญญาประดิษฐ์ที่ตนดูแล

๔.๒) ตัดสินใจอนุมัติการนำระบบขึ้นใช้งานจริง โดยพิจารณาจากความเสี่ยงทางธุรกิจ ชื่อเสียง และความสอดคล้องกับกฎหมายเป็นสำคัญ ควบคู่ไปกับรายงานผลการประเมินความเสี่ยงและผลการทดสอบความปลอดภัย

๔.๓) กำหนดและบังคับใช้นโยบายการใช้งานที่ยอมรับได้สำหรับระบบปัญญาประดิษฐ์นั้น ๆ

๔.๔) ทำงานร่วมกับทีมกฎหมายและกำกับการปฏิบัติตาม เพื่อประเมินผลกระทบด้านกฎระเบียบ

๔.๕) กำกับดูแลและติดตาม ให้มั่นใจว่ามีการนำมาตรการด้านความมั่นคงปลอดภัยที่จำเป็นไปปฏิบัติจริง

๔.๖) รายงานสถานะความเสี่ยงเหตุการณ์ด้านความมั่นคงปลอดภัย หรือประเด็นปัญหาที่สำคัญต่อคณะทำงานด้านความเสี่ยงหรือผู้บริหารที่เกี่ยวข้อง

๕) ทีมวิทยาศาสตร์ข้อมูลและวิศวกรการเรียนรู้ของเครื่อง

บทบาท ผู้ออกแบบ พัฒนา และฝึกสอนโมเดลปัญญาประดิษฐ์ หรือเทียบเท่ากับผู้สร้าง

ความรับผิดชอบด้านความมั่นคงปลอดภัยปัญญาประดิษฐ์

๕.๑) นำข้อกำหนดด้านความปลอดภัยจาก ระยะที่ ๑ มาปฏิบัติให้เกิดผลจริงใน ระยะที่ ๒

๕.๒) ใช้เทคนิคการสร้างโมเดลที่ทนทานต่อการโจมตี เช่น Adversarial Training

๕.๓) ประยุกต์ใช้เทคนิคการคุ้มครองความเป็นส่วนตัวในข้อมูล เช่น Data Anonymization และ Differential Privacy

๕.๔) จัดทำเอกสารที่เกี่ยวข้องกับโมเดล เพื่อระบุข้อจำกัด อคติที่อาจเกิดขึ้น และประสิทธิภาพของโมเดล

๖) ทีมวิศวกรรมความมั่นคงปลอดภัย

บทบาท ผู้เชี่ยวชาญด้านความมั่นคงปลอดภัยที่ให้คำปรึกษา ตรวจสอบ และเฝ้าระวัง "ผู้แนะนำและผู้ทวนสอบ"

ความรับผิดชอบด้านความมั่นคงปลอดภัยปัญญาประดิษฐ์

๖.๑) นำกระบวนการการสร้างแบบจำลองภัยคุกคามสำหรับระบบปัญญาประดิษฐ์ในระยะการออกแบบ

๖.๒) ดำเนินการทดสอบเจาะระบบสำหรับปัญญาประดิษฐ์เพื่อค้นหาช่องโหว่ในระยะการทวนสอบ

๖.๓) ออกแบบและติดตั้งระบบบันทึกข้อมูลและเฝ้าระวังสำหรับตรวจจับการโจมตีที่จำเพาะเจาะจงกับปัญญาประดิษฐ์

๖.๔) พัฒนาคู่มือการรับมือเหตุการณ์สำหรับสถานการณ์ที่เกี่ยวข้องกับปัญญาประดิษฐ์โดยเฉพาะ

๗) ทีมกฎหมายและกำกับการปฏิบัติตาม

บทบาท ผู้ดูแลให้การดำเนินงานของระบบปัญญาประดิษฐ์ สอดคล้องกับกฎหมาย กฎระเบียบ และมาตรฐานที่เกี่ยวข้อง

ความรับผิดชอบด้านความมั่นคงปลอดภัยปัญญาประดิษฐ์

๗.๑) ตีความและให้คำแนะนำเกี่ยวกับข้อกฎหมายคุ้มครองข้อมูลส่วนบุคคล เช่น PDPA ที่มีผลต่อการรวบรวมและใช้ข้อมูลฝึกสอน

๗.๒) ทบทวนและอนุมัตินโยบายการจัดเก็บและทำลายข้อมูล

๗.๓) ประเมินความรับผิดทางกฎหมายที่อาจเกิดขึ้นจากการตัดสินใจที่ผิดพลาดของปัญญาประดิษฐ์

๘) เจ้าหน้าที่คุ้มครองข้อมูลส่วนบุคคล

บทบาท ผู้กำกับดูแลและให้คำปรึกษาด้านการคุ้มครองข้อมูลส่วนบุคคลโดยเฉพาะ

ความรับผิดชอบด้านความมั่นคงปลอดภัยปัญญาประดิษฐ์

๘.๑) ดำเนินการประเมินผลกระทบด้านการคุ้มครองข้อมูลส่วนบุคคลสำหรับโครงการปัญญาประดิษฐ์

๘.๒) ให้คำปรึกษาเกี่ยวกับความชอบด้วยกฎหมายในการประมวลผลข้อมูลเพื่อใช้ฝึกสอนปัญญาประดิษฐ์

๘.๓) เป็นจุดประสานงานสำหรับเจ้าของข้อมูลในการใช้สิทธิ์ของตนที่เกี่ยวข้องกับระบบปัญญาประดิษฐ์

๙) ผู้ใช้งานระบบปัญญาประดิษฐ์

บทบาท บุคลากรหรือลูกค้าที่ใช้งานหรือมีปฏิสัมพันธ์กับระบบปัญญาประดิษฐ์โดยตรง

ความรับผิดชอบด้านความมั่นคงปลอดภัยปัญญาประดิษฐ์

๙.๑) ตระหนักรู้ว่ากำลังใช้งานหรือได้รับผลจากการทำงานของระบบปัญญาประดิษฐ์

๙.๒) ปฏิบัติตามนโยบายการใช้งานที่ยอมรับได้ของระบบนั้น ๆ

๙.๓) รายงานพฤติกรรมของระบบที่ผิดปกติ น่าสงสัย หรืออาจก่อให้เกิดความเสียหายแก่เจ้าของระบบ

การกำหนดบทบาทและความรับผิดชอบที่ชัดเจนดังตารางสรุปด้านล่างนี้ จะช่วยสร้างวัฒนธรรมความรับผิดชอบร่วมกันและทำให้มั่นใจได้ว่ามิติด้านความมั่นคงปลอดภัยจะถูกบูรณาการเข้าไปในทุกกระบวนการอย่างแท้จริง

เพื่อให้สามารถนำแนวทางนี้ไปปฏิบัติได้ชัดเจนขึ้นจึงได้ใช้แผนภูมิแสดงความรับผิดชอบตามหลักการ RACI ซึ่งเป็นเครื่องมือที่ใช้ในการบริหารจัดการโครงการหรือการดำเนินงาน เพื่อกำหนดบทบาทและความรับผิดชอบของผู้ที่เกี่ยวข้องในแต่ละกิจกรรมหรืองานให้ชัดเจน ช่วยลดความสับสน ป้องกันงานซ้ำซ้อน และทำให้มั่นใจว่ามีผู้รับผิดชอบในทุกขั้นตอน ซึ่งเป็นกรอบแนวคิดและแนวปฏิบัติที่ดีที่ได้รับการยอมรับและนำไปประยุกต์ใช้อย่างแพร่หลายในมาตรฐานและกรอบการทำงานชั้นนำต่าง ๆ ดังนี้

๑) Project Management Institute (PMI)

๑.๑) ในคู่มือ A Guide to the Project Management Body of Knowledge (PMBOK Guide) ซึ่งเป็นมาตรฐานหลักสำหรับผู้ประกอบวิชาชีพจัดการโครงการ (PMP) ได้กล่าวถึงเครื่องมือที่เรียกว่า Responsibility Assignment Matrix (RAM) หรือเมทริกซ์การกำหนดความรับผิดชอบ โดยระบุว่า RACI Chart เป็นรูปแบบหนึ่งของ RAM ที่ได้รับความนิยมมากที่สุด

๑.๒) การอ้างอิง: สามารถอ้างอิงได้ว่า RACI เป็นเครื่องมือที่ได้รับการยอมรับและเป็นส่วนหนึ่งขององค์ความรู้ด้านการบริหารโครงการตามแนวทางของ PMI

๒) ISACA (Information Systems Audit and Control Association)

๒.๑) ในกรอบการกำกับดูแลและบริหารจัดการเทคโนโลยีสารสนเทศ COBIT (Control Objectives for Information and Related Technologies) ได้มีการใช้ RACI Chart อย่างชัดเจน เพื่อช่วยในการกำหนดบทบาทและความรับผิดชอบสำหรับกระบวนการต่าง ๆ ที่เกี่ยวข้องกับการกำกับดูแลและการบริหารจัดการไอที

๒.๒) การอ้างอิง: สามารถอ้างอิงได้ว่า RACI เป็นเครื่องมือสำคัญที่ใช้ในกรอบการทำงาน COBIT ของ ISACA เพื่อสร้างธรรมาภิบาลด้านไอทีที่ดี

๓) ITIL (Information Technology Infrastructure Library)

๓.๑) ITIL ซึ่งเป็นกรอบแนวปฏิบัติที่ดีที่สุดสำหรับการบริหารจัดการบริการสารสนเทศ ได้นำ RACI มาใช้อย่างกว้างขวางเพื่อกำหนดว่าใครมีบทบาทอะไรในกระบวนการต่าง ๆ เช่น การจัดการเหตุการณ์ และการจัดการการเปลี่ยนแปลง

๓.๒) สามารถอ้างอิงได้ว่า RACI เป็นเครื่องมือมาตรฐานในการกำหนดบทบาทและความรับผิดชอบในกระบวนการบริหารจัดการบริการสารสนเทศ ตามแนวทางของ ITIL

ตัวอักษรแต่ละตัวใน RACI มีความหมายที่แตกต่างกันอย่างชัดเจน ดังนี้

R - Responsible (ผู้ปฏิบัติ)

- หมายถึง **ผู้ที่ลงมือทำงานหรือกิจกรรมนั้น ๆ ให้สำเร็จ**
- ในหนึ่งกิจกรรมอาจมีผู้ปฏิบัติ (R) มากกว่าหนึ่งคนได้

A - Accountable (ผู้รับผิดชอบสูงสุด)

- หมายถึง **บุคคลเพียงคนเดียวที่มีอำนาจตัดสินใจขั้นสุดท้ายและต้องรับผิดชอบต่อผลลัพธ์ของงานทั้งหมด**
- เป็นผู้มอบหมายงานให้ R และเป็นผู้ตรวจสอบอนุมัติงานที่ R ทำเสร็จแล้ว
- **ข้อสำคัญ** ในแต่ละกิจกรรม **จะต้องมีผู้รับผิดชอบสูงสุด (A) เพียงคนเดียวเท่านั้น** เพื่อป้องกันการเกี่ยงกันรับผิดชอบ

C - Consulted (ผู้ให้คำปรึกษา)

- หมายถึง ผู้เชี่ยวชาญที่ควรได้รับการปรึกษาหารือ **ก่อน**ที่จะลงมือทำหรือตัดสินใจในกิจกรรมนั้น ๆ
- เป็นการสื่อสารแบบสองทาง (Two-Way Communication) เพื่อขอความเห็นหรือข้อมูลเพิ่มเติม

I - Informed (ผู้รับทราบข้อมูล)

- หมายถึง ผู้ที่ต้องได้รับแจ้งข้อมูลความคืบหน้าหรือผลลัพธ์ของกิจกรรม **หลังจาก**ที่งานนั้นสำเร็จหรือมีการตัดสินใจไปแล้ว
- เป็นการสื่อสารแบบทางเดียว (One-Way Communication) โดยไม่จำเป็นต้องขอความเห็นกลับ

ตารางที่ ๖ แสดงแผนภูมิแสดงความรับผิดชอบตามหลักการ RACI

กิจกรรม	๑. ผู้บริหารระดับสูง	๒. AI Governance Board	๓. AI Risk Committee	๔. เจ้าของระบบปัญญาประดิษฐ์ (ธุรกิจ)	๕. ทีม Data Science/ML	๖. ทีม Security	๗. ทีม Legal/ Compliance	๘. DPO	๙. ผู้ใช้งาน
๑) ด้านธรรมาภิบาลและนโยบาย									
กำหนดและอนุมัตินโยบาย/ธรรมาภิบาล AI	A	R	C	I	I	C	C	C	I
กำหนดระดับความเสี่ยงที่ยอมรับได้ (Risk Appetite)	A	C	R	C	I	C	I	I	-
จัดสรรทรัพยากร (งบประมาณ บุคลากร)	A	C	I	R	C	C	I	I	-
๒) ด้านการบริหารความเสี่ยงและกฎหมาย									
ประเมินความเสี่ยงและผลกระทบ (DPIA)	I	C	R	A	C	C	C	R	-
บูรณาการความเสี่ยง AI เข้ากับ ERM	I	I	A/R	C	I	I	I	I	-

กิจกรรม	๑. ผู้บริหารระดับสูง	๒. AI Governance Board	๓. AI Risk Committee	๔. เจ้าของระบบปัญญาประดิษฐ์ (ธุรกิจ)	๕. ทีม Data Science/ML	๖. ทีม Security	๗. ทีม Legal/ Compliance	๘. DPO	๙. ผู้ใช้งาน
ตรวจสอบการปฏิบัติตามกฎหมาย/ข้อบังคับ	I	C	C	R	I	I	A	R	-
๓) ด้านการพัฒนาและปฏิบัติการ									
ออกแบบและพัฒนาโมเดลที่ปลอดภัย	-	I	I	A	R	C	C	C	-
สร้างแบบจำลองภัยคุกคาม (Threat Modeling)	-	I	I	A	C	R	I	I	-
ทดสอบเจาะระบบ (AI Red Teaming)	-	I	I	A	C	R	I	I	-
อนุมัติการนำระบบขึ้นใช้งาน (Go/No-Go)	I	C	C	A	C	C	C	C	-
เฝ้าระวังและรับมือเหตุการณ์ความปลอดภัย	I	I	I	A	C	R	I	I	-

กิจกรรม	๑. ผู้บริหารระดับสูง	๒. AI Governance Board	๓. AI Risk Committee	๔. เจ้าของระบบปัญญาประดิษฐ์ (ธุรกิจ)	๕. ทีม Data Science/ML	๖. ทีม Security	๗. ทีม Legal/ Compliance	๘. DPO	๙. ผู้ใช้งาน
๔) ด้านการใช้งาน									
ใช้งานระบบตามนโยบาย	-	-	-	A	-	-	-	-	R
รายงานพฤติกรรมของระบบที่ผิดปกติ	-	-	I	A	C	C	-	-	R

๔.๒ การบูรณาการความเสี่ยงปัญญาประดิษฐ์เข้ากับกรอบการบริหารความเสี่ยงองค์กร (Integrating AI Risks into ERM)

ความเสี่ยงที่เกิดจากระบบปัญญาประดิษฐ์ไม่ใช่เป็นเพียงความเสี่ยงทางเทคโนโลยีที่จำกัดอยู่แค่ในฝ่ายเทคโนโลยีสารสนเทศเท่านั้น แต่เป็นความเสี่ยงทางธุรกิจที่สามารถส่งผลกระทบในระดับองค์กรได้ ดังนั้น การบริหารจัดการความเสี่ยงปัญญาประดิษฐ์ในลักษณะแยกส่วน จึงไม่มีประสิทธิภาพเพียงพอ องค์กรจึงจำเป็นต้องบูรณาการความเสี่ยงที่จำเพาะเจาะจงกับปัญญาประดิษฐ์ เข้าไปในกรอบการบริหารความเสี่ยงระดับองค์กร (Enterprise Risk Management - ERM) ที่มีอยู่เดิม

กระบวนการนี้ทำให้มั่นใจได้ว่าผู้บริหารระดับสูงและคณะกรรมการบริษัทจะสามารถมองเห็น ประเมิน และตัดสินใจเกี่ยวกับความเสี่ยงปัญญาประดิษฐ์ได้ในบริบทเดียวกับความเสี่ยงอื่น ๆ ขององค์กร เช่น ความเสี่ยงทางการเงิน การตลาด หรือการปฏิบัติการ ซึ่งนำไปสู่การจัดสรรทรัพยากรและการกำหนดกลยุทธ์ที่สอดคล้องกันทั่วทั้งองค์กร

กรอบการดำเนินงานและมาตรฐานที่เกี่ยวข้อง

เพื่อดำเนินการบูรณาการอย่างเป็นระบบ องค์กรควรใช้ประโยชน์จากมาตรฐานและกรอบการทำงานที่เป็นที่ยอมรับในระดับสากล

๑) COSO ERM Framework (Integrating with Strategy and Performance) เป็นกรอบการทำงานชั้นนำสำหรับ ERM ที่ช่วยให้องค์กรสามารถเชื่อมโยงการบริหารความเสี่ยงเข้ากับกลยุทธ์และผลการดำเนินงาน การนำความเสี่ยงปัญญาประดิษฐ์มาพิจารณาภายใต้กรอบของ COSO จะช่วยให้องค์กรเข้าใจว่าความเสี่ยงเหล่านี้ส่งผลต่อการบรรลุวัตถุประสงค์เชิงกลยุทธ์ได้อย่างไร

๒) ISO 31000:2018 (Risk Management — Guidelines) เป็นมาตรฐานสากลที่ให้แนวทางและหลักการพื้นฐานสำหรับกระบวนการบริหารความเสี่ยงทั้งหมด ตั้งแต่การกำหนดบริบท การประเมินความเสี่ยงไปจนถึงการจัดการความเสี่ยง ซึ่งสามารถใช้เป็นแนวทางหลักในการประเมินความเสี่ยงของระบบปัญญาประดิษฐ์ได้

กระบวนการบูรณาการความเสี่ยงปัญญาประดิษฐ์เข้ากับประเภทความเสี่ยงใน ERM

องค์กรสามารถจับคู่ความเสี่ยงเฉพาะทางของปัญญาประดิษฐ์เข้ากับหมวดหมู่ความเสี่ยงของ ERM ที่มีอยู่เดิมได้ ดังตัวอย่างต่อไปนี้

๑) ความเสี่ยงด้านกลยุทธ์

๑.๑) คำจำกัดความ ความเสี่ยงที่ส่งผลกระทบต่อความสามารถในการบรรลุเป้าหมายทางธุรกิจในระยะยาว

๑.๒) ตัวอย่างความเสี่ยงปัญญาประดิษฐ์ การเสื่อมถอยของโมเดล (Model Degradation / Concept Drift)

สถานการณ์ โมเดลปัญญาประดิษฐ์วิเคราะห์แนวโน้มตลาดขององค์กรทำงานได้ไม่แม่นยำเหมือนเดิม เนื่องจากพฤติกรรมผู้บริโภคเปลี่ยนแปลงไปหลังเกิดวิกฤตเศรษฐกิจ (Concept Drift) หากองค์กรไม่สามารถตรวจจับและฝึกสอนโมเดลใหม่ได้ทันท่วงที อาจนำไปสู่การตัดสินใจทางกลยุทธ์ที่ผิดพลาดและสูญเสียความสามารถในการแข่งขัน

๒) ความเสี่ยงด้านปฏิบัติการ

๒.๑) คำจำกัดความ ความเสี่ยงที่จะเกิดความเสียหายจากความล้มเหลวของกระบวนการภายใน บุคลากร หรือระบบ

๒.๒) ตัวอย่างความเสี่ยงปัญญาประดิษฐ์ การโจมตีห่วงโซ่อุปทาน

สถานการณ์ ทีมพัฒนาได้ดาวน์โหลดโมเดลภาษาจากแหล่งที่ไม่น่าเชื่อถือซึ่งถูกฝัง Backdoor จากผู้ไม่หวังดี เมื่อนำมาใช้งานจริงทำให้ผู้โจมตีสามารถควบคุมผลลัพธ์ของระบบ Chatbot ที่ให้บริการลูกค้าได้ ซึ่งถือเป็นความล้มเหลวของกระบวนการพัฒนาและจัดหาที่มั่นคงปลอดภัย

๒.๓) ตัวอย่างความเสี่ยงปัญญาประดิษฐ์ การใช้ Prompt Injection เพื่อก่อให้เกิดความเสียหายต่อระบบ

สถานการณ์ ผู้ไม่หวังดีใช้เทคนิค Prompt Injection กับระบบปัญญาประดิษฐ์ภายในที่เชื่อมต่อกับระบบอัตโนมัติของฝ่ายปฏิบัติการ โดยหลอกล่อให้ปัญญาประดิษฐ์แปลความหมายและส่งชุดคำสั่งที่เป็นอันตราย ส่งผลให้ระบบเซิร์ฟเวอร์หลักหยุดทำงาน ถือเป็นความล้มเหลวของระบบควบคุมข้อมูลนำเข้า (Input Control) ที่ก่อให้เกิดความเสียหายต่อการดำเนินงานโดยตรง

๓) ความเสี่ยงด้านการเงิน

๓.๑) คำจำกัดความ ความเสี่ยงที่จะเกิดความสูญเสียทางการเงินทั้งทางตรงและทางอ้อม

๓.๒) ตัวอย่างความเสี่ยงปัญญาประดิษฐ์ การรั่วไหลของทรัพย์สินทางปัญญา

สถานการณ์ คู่แข่งทางธุรกิจใช้เทคนิค Model Extraction โดยการส่งคำร้องจำนวนมหาศาลมายัง API ของระบบปัญญาประดิษฐ์ แนะนำราคาผลิตภัณฑ์ขององค์กร เพื่อสร้างโมเดลลอกเลียนแบบที่มีความสามารถใกล้เคียงกัน ส่งผลให้องค์กรสูญเสียความได้เปรียบทางการแข่งขันและรายได้ในอนาคต ซึ่งถือเป็นความเสียหายทางการเงินทางอ้อม

๔) ความเสี่ยงด้านการปฏิบัติตามกฎหมายและกฎระเบียบ

๔.๑) คำจำกัดความ ความเสี่ยงที่จะถูกลงโทษหรือคว่ำบาตรจากการไม่ปฏิบัติตามกฎหมายกฎระเบียบหรือมาตรฐาน

๔.๒) ตัวอย่างความเสี่ยงปัญญาประดิษฐ์ การละเมิดกฎหมายคุ้มครองข้อมูลส่วนบุคคล

สถานการณ์ โมเดลปัญญาประดิษฐ์ที่ให้บริการสรุปเอกสารเกิดความผิดพลาดและเปิดเผยข้อมูลส่วนบุคคลที่ละเอียดอ่อนของลูกค้าซึ่งอยู่ในชุดข้อมูลฝึกสอน ถือเป็นการละเมิดพระราชบัญญัติคุ้มครองข้อมูลส่วนบุคคลและอาจนำไปสู่โทษปรับทางปกครองจำนวนมหาศาล

๕) ความเสี่ยงด้านชื่อเสียง

๕.๑) คำจำกัดความ ความเสี่ยงที่จะเกิดความเสียหายต่อภาพลักษณ์และความไว้วางใจขององค์กร

๕.๒) ตัวอย่างความเสี่ยงปัญญาประดิษฐ์ การสร้างผลลัพธ์ที่มีอคติและไม่เป็นธรรม

สถานการณ์ ระบบปัญญาประดิษฐ์ที่ใช้คัดกรองใบสมัครงานถูกค้นพบว่า มีอคติต่อผู้สมัครเพศหญิงอย่างมีนัยสำคัญ เนื่องจากถูกฝึกสอนด้วยข้อมูลในอดีตที่มีอคติแฝงอยู่ เมื่อเรื่องนี้ถูกเปิดเผยสู่สาธารณะ ย่อมสร้างความเสียหายอย่างรุนแรงต่อชื่อเสียงขององค์กรในฐานะนายจ้างที่ไม่เท่าเทียม

๕.๓) ตัวอย่างความเสี่ยงปัญญาประดิษฐ์ การถูกใช้เป็นเครื่องมือสร้าง Deepfake

สถานการณ์ ระบบปัญญาประดิษฐ์สร้างภาพขององค์กรถูกนำไปใช้สร้างวิดีโอปลอมของผู้บริหารกล่าวข้อมูลที่สร้างความเสียหายต่อความเชื่อมั่นของนักลงทุน ส่งผลกระทบต่อราคาหุ้นและชื่อเสียงขององค์กรอย่างรุนแรง

โดยสรุปการจับคู่ความเสี่ยง AI เข้ากับประเภทความเสี่ยงใน ERM เพื่อให้ผู้บริหารมองเห็นภาพรวมในระดับองค์กร ดังตารางนี้

ตารางที่ ๗ การจำแนกประเภทและความเสี่ยงด้านความมั่นคงปลอดภัยของปัญญาประดิษฐ์ ภายใต้กรอบ ERM

ประเภทความเสี่ยง ERM	คำจำกัดความ	ตัวอย่างความเสี่ยง AI ที่เกี่ยวข้อง
๑) ด้านกลยุทธ์	ความเสี่ยงที่กระทบต่อการบรรลุเป้าหมายทางธุรกิจ	**การเสื่อมถอยของโมเดล (Model Degradation / Concept Drift)** โมเดลวิเคราะห์แนวโน้มตลาดไม่แม่นยำ เนื่องจากพฤติกรรมผู้บริโภคเปลี่ยนไป นำไปสู่การตัดสินใจเชิงกลยุทธ์ที่ผิดพลาด
๒) ด้านปฏิบัติการ	ความเสี่ยงจากความล้มเหลวของกระบวนการภายในคนหรือระบบ	**การโจมตีห่วงโซ่อุปทาน** ทีมพัฒนานำ Pre-trained Model ที่ถูกฝัง Backdoor มาใช้งาน ทำให้ผู้โจมตีสามารถควบคุมผลลัพธ์ของระบบได้ ผู้ไม่หวังดีใช้เทคนิค Prompt Injection กับระบบปัญญาประดิษฐ์ภายในที่เชื่อมต่อกับระบบอัตโนมัติของฝ่ายปฏิบัติการ โดยหลอกล่อให้ปัญญาประดิษฐ์

ประเภทความเสี่ยง ERM	คำจำกัดความ	ตัวอย่างความเสี่ยง AI ที่เกี่ยวข้อง
		แปลความหมายและส่งชุดคำสั่งที่เป็นอันตราย ส่งผลให้ระบบเซิร์ฟเวอร์หลักหยุดทำงาน
๓) ด้านการเงิน	ความเสี่ยงที่จะเกิดความสูญเสียทางการเงิน	**การรั่วไหลของทรัพย์สินทาง** คู่แข่งใช้เทคนิคคัดลอกโมเดล ทำให้องค์กรสูญเสียความได้เปรียบทางการแข่งขันและรายได้
๔) ด้านการปฏิบัติตามกฎหมาย	ความเสี่ยงจากการไม่ปฏิบัติตามกฎหมายหรือกฎระเบียบ	**การละเมิดกฎหมายคุ้มครองข้อมูล** โมเดลปัญญาประดิษฐ์เปิดเผยข้อมูลส่วนบุคคล ที่ละเอียดอ่อนของลูกค้าโดยไม่ได้ตั้งใจ ซึ่งเป็นการละเมิด PDPA และอาจนำไปสู่โทษปรับมหาศาล
๕) ด้านชื่อเสียง	ความเสี่ยงที่จะเกิดความเสียหายต่อภาพลักษณ์และความไว้วางใจ	**การสร้างผลลัพธ์ที่มีอคติ** ระบบปัญญาประดิษฐ์คัดกรองใบสมัครงานมีอคติต่อเพศใดเพศหนึ่ง ทำให้เกิดภาพลักษณ์เชิงลบต่อองค์กร ระบบปัญญาประดิษฐ์สร้างภาพขององค์กรถูกนำไปใช้สร้างวิดีโอปลอมของผู้บริหาร กล่าวข้อมูลที่สร้างความเสียหายต่อความเชื่อมั่นของนักลงทุน ส่งผลกระทบต่อราคาหุ้นและชื่อเสียงขององค์กรอย่างรุนแรง

การบูรณาการความเสี่ยงปัญญาประดิษฐ์เข้ากับ ERM ไม่ใช่เพียงกิจกรรมทางเทคนิค แต่เป็นหน้าที่สำคัญของการกำกับดูแลที่จะช่วยให้องค์กรสามารถขับเคลื่อนนวัตกรรมไปพร้อมกับการบริหารจัดการความเสี่ยงได้อย่างสมดุลและยั่งยืน

กระบวนการบริหารความเสี่ยงปัญญาประดิษฐ์ตามแนวทาง ISO/IEC 23894

นอกเหนือจากกรอบการทำงานระดับองค์กรแล้ว การจัดการความเสี่ยงปัญญาประดิษฐ์ในทางปฏิบัติควรดำเนินตามกระบวนการที่เป็นวงจรและต่อเนื่องตามมาตรฐานเฉพาะทางอย่าง ISO/IEC 23894:2023 ซึ่งประกอบด้วย ๔ ขั้นตอนสำคัญ ดังนี้

๑) การระบุความเสี่ยง

เป็นการค้นหาและระบุภัยคุกคามและช่องโหว่ที่อาจเกิดขึ้นกับระบบปัญญาประดิษฐ์ตลอดวงจรชีวิตของระบบปัญญาประดิษฐ์ โดยครอบคลุมความเสี่ยงในมิติต่าง ๆ เช่น

๑.๑) ความเสี่ยงทางเทคนิค Data Poisoning, Model Evasion, Prompt Injection และ AI Supply Chain Attack

๑.๒) ความเสี่ยงเชิงองค์ประกอบ ความเสี่ยงด้านข้อมูล โมเดล แอปพลิเคชัน และโครงสร้างพื้นฐาน

๒) การวิเคราะห์และประเมินความเสี่ยง

หลังจากระบุความเสี่ยงแล้ว จะทำการวิเคราะห์ โอกาสที่จะเกิดและผลกระทบ หากความเสี่ยงนั้นเกิดขึ้นจริง โดยมักใช้ตารางเมทริกซ์ (Likelihood × Impact Matrix) เพื่อประเมินและจัดลำดับความสำคัญของความเสี่ยง เช่น สูง กลาง ต่ำ

๓) การตอบสนองต่อความเสี่ยง

กำหนดแนวทางการจัดการความเสี่ยงที่ประเมินไว้ โดยเลือกจาก ๔ แนวทางหลัก

๓.๑) หลีกเลี่ยง ยกเลิกกิจกรรมที่ก่อให้เกิดความเสี่ยง

๓.๒) บรรเทา กำหนดมาตรการควบคุมเพื่อลดโอกาสเกิดหรือผลกระทบ เช่น การควบคุมการเข้าถึงการเก็บข้อมูลจราจรคอมพิวเตอร์ การเข้ารหัสลับ

๓.๓) ถ่ายโอน โอนย้ายความรับผิดชอบไปยังบุคคลที่สาม เช่น การทำประกันภัย

๓.๔) ยอมรับ ยอมรับความเสี่ยงไว้หากอยู่ในระดับที่องค์กรรับได้

๔) การติดตามและทบทวน

เป็นกระบวนการต่อเนื่องเพื่อตรวจสอบประสิทธิภาพของมาตรการควบคุมที่กำหนดไว้ และทบทวนสถานะความเสี่ยงอย่างสม่ำเสมอ เช่น รายไตรมาส รายปี พร้อมทั้งรายงานผลต่อผู้บริหารและคณะกรรมการกำกับดูแล เพื่อให้มั่นใจว่าการจัดการความเสี่ยงยังคงเหมาะสมกับสถานการณ์ที่เปลี่ยนแปลงไป

หลังจากที่ดำเนินกระบวนการบริหารความเสี่ยงครบทั้ง ๔ ขั้นตอนแล้ว ผลลัพธ์ทั้งหมดจะถูกนำมาบันทึกและติดตามอย่างเป็นระบบในเอกสารที่เรียกว่า **ทะเบียนความเสี่ยงด้านความมั่นคงปลอดภัยปัญญาประดิษฐ์** ซึ่งถือเป็นเครื่องมือสำคัญในการกำกับดูแล

ทะเบียนความเสี่ยง คือ เอกสารกลางที่ใช้รวบรวม ติดตาม และจัดการความเสี่ยงทั้งหมดของระบบปัญญาประดิษฐ์อย่างต่อเนื่องโดยข้อมูลในแต่ละส่วนของทะเบียนความเสี่ยงจะมาจากผลลัพธ์ของกระบวนการบริหารความเสี่ยงในแต่ละขั้นตอนโดยตรง

ขั้นตอนที่ ๑ การระบุความเสี่ยง ⟶ *สร้างรายการความเสี่ยง*

ขั้นตอนนี้จะให้ข้อมูลพื้นฐานสำหรับคอลัมน์หลัก ๆ ในทะเบียนความเสี่ยง

- รหัสความเสี่ยง (Risk ID) รหัสอ้างอิงเฉพาะสำหรับแต่ละความเสี่ยง เช่น AI-SEC-001
- ชื่อความเสี่ยง (Risk Title) ชื่อที่กระชับและสื่อความหมาย เช่น Data Poisoning Attack
- รายละเอียดความเสี่ยง คำอธิบายสถานการณ์ของความเสี่ยง สาเหตุ และผลกระทบที่อาจเกิดขึ้น
- หมวดหมู่ความเสี่ยง ประเภทของความเสี่ยงที่จับคู่กับ ERM เช่น ด้านปฏิบัติการด้านชื่อเสียง

ขั้นตอนที่ ๒ การวิเคราะห์และประเมิน ⟶ *ประเมินระดับความรุนแรง*

ผลการวิเคราะห์จากขั้นตอนนี้จะถูกนำมาใส่ในคอลัมน์ที่ใช้ประเมินความรุนแรง

- โอกาสที่จะเกิด ระดับความเป็นไปได้ที่จะเกิดความเสี่ยง เช่น ๑ - ๕
- ผลกระทบ ระดับความรุนแรงของความเสียหายหากเกิดขึ้น เช่น ๑ - ๕
- ระดับความเสี่ยง ผลคูณของ โอกาส × ผลกระทบ เพื่อจัดลำดับความสำคัญ เช่น สูง กลาง ต่ำ

ขั้นตอนที่ ๓ การตอบสนอง ⟶ *กำหนดแผนจัดการ*

แผนการจัดการที่กำหนดขึ้นจะถูกบันทึกไว้ในคอลัมน์สำหรับการดำเนินการ

- แนวทางการตอบสนอง กลยุทธ์ที่เลือกใช้ หลีกเลี่ยง บรรเทา ถ่ายโอน ยอมรับ
- มาตรการควบคุม รายละเอียดของมาตรการที่นำมาใช้เพื่อลดความเสี่ยง
- ผู้รับผิดชอบ บุคคลหรือตำแหน่งที่รับผิดชอบในการจัดการความเสี่ยงนี้
- กำหนดแล้วเสร็จ วันที่คาดว่ามาตรการควบคุมจะดำเนินการแล้วเสร็จ

ขั้นตอนที่ ๔ การติดตามและทบทวน ⟶ *ปรับปรุงสถานะ*

ขั้นตอนนี้ทำให้ทะเบียนความเสี่ยงเป็นเอกสารที่ไม่หยุดนิ่ง โดยใช้คอลัมน์สำหรับติดตามความคืบหน้า

- สถานะ สถานะปัจจุบันของการจัดการความเสี่ยง เช่น เปิดใช้ กำลังดำเนินการ ปิดแล้ว
- วันที่ทบทวนล่าสุด วันที่ตรวจสอบและประเมินความเสี่ยงนี้ครั้งล่าสุด
- วันที่ทบทวนครั้งถัดไป กำหนดการทบทวนครั้งต่อไป

การจัดทำทะเบียนความเสี่ยงด้านความมั่นคงปลอดภัยปัญญาประดิษฐ์ ที่ดีจะช่วยให้องค์กรมีมุมมองที่ชัดเจนต่อความเสี่ยงทั้งหมด ทำให้สามารถจัดลำดับความสำคัญในการใช้ทรัพยากร และแสดงถึงการกำกับดูแลกิจการที่ดีต่อผู้มีส่วนได้ส่วนเสียทุกฝ่าย

ตัวอย่าง ตารางทะเบียนความเสี่ยงด้านความมั่นคงปลอดภัยปัญญาประดิษฐ์ จำนวน ๘ ตัวอย่าง โดยมีการตอบสนองครบทั้ง ๔ กลยุทธ์ บรรเทา ถ่ายโอน หลีกเลี่ยง และยอมรับ แสดงได้ดังนี้

ตารางที่ ๘ ตัวอย่างตารางทะเบียนความเสี่ยง

รหัส	ชื่อความเสี่ยง	รายละเอียด	หมวดหมู่	โอกาส (๑-๕)	ผลกระทบ (๑-๕)	ระดับความเสี่ยง	แนวทางการตอบสนอง	มาตรการควบคุม	ผู้รับผิดชอบ	กำหนดเสร็จ	สถานะ
AI-SEC-๐๐๑	Data Poisoning Attack	ผู้ไม่หวังดีแฝงข้อมูลที่เป็นพิษเข้ามาในชุดข้อมูลฝึกสอน ทำให้โมเดลเรียนรู้และตัดสินใจผิดพลาด	ปฏิบัติการ	๓ (กลาง)	๔ (สูง)	๑๒ (สูง)	บรรเทา	๑. ตรวจสอบความถูกต้องของแหล่งข้อมูล ๒. ใช้ Anomaly Detection ตรวจจับข้อมูลผิดปกติ ๓. จำกัดสิทธิ์การเข้าถึงข้อมูลฝึกสอน	Head of Data Science	๓๑/๑๒/๒๕๖๘	กำลังดำเนินการ
AI-SEC-๐๐๒	ผลลัพธ์ที่มีอคติ (Biased Outcome)	โมเดลคัดกรองผู้สมัครงานมีอคติต่อเพศ/เชื้อชาติ เนื่องจากข้อมูลที่ใช้สอนมีอคติแฝงอยู่	ชื่อเสียง / ปฏิบัติตามกฎหมาย	๔ (สูง)	๔ (สูง)	๑๖ (สูงมาก)	บรรเทา	๑. ใช้เครื่องมือตรวจสอบอคติ ๒. จัดตั้งทีมตรวจสอบด้านจริยธรรมปัญญาประดิษฐ์ ๓. ปรับปรุงชุดข้อมูลให้มีความหลากหลาย	Chief People Officer	๓๐/๐๙/๒๕๖๘	กำลังดำเนินการ
AI-SEC-๐๐๓	Prompt Injection	ผู้ใช้งานสร้าง Prompt หลอกล่อให้ LLM ที่เชื่อมต่อกับระบบ	ปฏิบัติการ / ข้อมูล	๓ (กลาง)	๕ (สูงมาก)	๑๕ (สูงมาก)	บรรเทา	๑. ใช้เทคนิค Input Sanitization และ Output Filtering	AI Product Manager	๓๑/๑๐/๒๕๖๘	เปิดใช้

รหัส	ชื่อความเสี่ยง	รายละเอียด	หมวดหมู่	โอกาส (๑-๕)	ผลกระทบ (๑-๕)	ระดับความเสี่ยง	แนวทางการตอบสนอง	มาตรการควบคุม	ผู้รับผิดชอบ	กำหนดเสร็จ	สถานะ
		ภายใน เปิดเผยข้อมูลที่ไม่ได้รับอนุญาต						๒. จำกัดสิทธิ์ของระบบที่ LLM สามารถเข้าถึงได้ ๓. มีมนุษย์คอยตรวจสอบผลลัพธ์ก่อนดำเนินการ			
AI-SEC-๐๐๔	DoS Attack บน API ของ AI	ผู้โจมตีส่งคำขอมหาศาลมายัง API ของระบบ ปัญญาประดิษฐ์ทำให้ระบบล่มและผู้ใช้ทั่วไปไม่สามารถใช้งานได้	ปฏิบัติการ / การเงิน	๒ (ต่ำ)	๕ (สูงมาก)	๑๐ (สูง)	ถ่ายโอน	ใช้บริการ Cloud Provider ที่มีระบบป้องกัน DDoS/DoS Attack ในตัว	Head of Infrastructure	N/A	ปิดแล้ว
AI-SEC-๐๐๕	ใช้ เปิดใช้-Source AI ที่ไม่น่าเชื่อถือ	ทีมพัฒนาต้องการใช้โมเดลปัญญาประดิษฐ์จากแหล่งที่ไม่น่าเชื่อถือ ซึ่งอาจมี Backdoor แฝงอยู่	กลยุทธ์ / การเงิน	๔ (สูง)	๕ (สูงมาก)	๒๐ (วิกฤต)	หลีกเลี่ยง	ไม่อนุมัติการใช้งานโมเดลดังกล่าว และกำหนดนโยบายให้พัฒนาโมเดลภายในหรือจัดหาจากแหล่งที่ผ่านการรับรองเท่านั้น	CTO / AI Governance Committee	N/A	ปิดแล้ว
AI-SEC-๐๐๖	Model Drift ในระบบไม่สำคัญ	โมเดลแนะนำข่าวสารภายในองค์กรมีความ	ปฏิบัติการ	๓ (กลาง)	๑ (ต่ำมาก)	๓ (ต่ำ)	ยอมรับ	ยอมรับความเสี่ยงและติดตามผลเป็นระยะ (ทุก ๖ เดือน) โดยยังไม่	Internal Comms Manager	N/A	ปิดแล้ว

รหัส	ชื่อความเสี่ยง	รายละเอียด	หมวดหมู่	โอกาส (๑-๕)	ผลกระทบ (๑-๕)	ระดับความเสี่ยง	แนวทางการตอบสนอง	มาตรการควบคุม	ผู้รับผิดชอบ	กำหนดเสร็จ	สถานะ
		แม่นยำลดลงเล็กน้อยเมื่อเวลาผ่านไป						มีมาตรการควบคุมเพิ่มเติมในขณะนี้เนื่องจากผลกระทบต่ำมาก			
AI-SEC-๐๐๗	การรั่วไหลของข้อมูลผ่าน Model Inversion	ผู้โจมตีสามารถสร้างข้อมูลบางส่วนที่ใช้ในการฝึกสอนโมเดลขึ้นมาใหม่ได้ เช่น ภาพใบหน้าจากโมเดลจดจำใบหน้า	ข้อมูล / ปฏิบัติตามกฎหมาย	๒ (ต่ำ)	๔ (สูง)	๘ (กลาง)	บรรเทา	๑. ใช้เทคนิค Differential Privacy ในระหว่างการฝึกสอนโมเดล ๒. จำกัดจำนวนการส่งคำร้องที่ส่งมายัง API เพื่อป้องกันการโจมตี	Chief Information Security Officer (CISO)	๓๑/๐๑/๒๕๖๙	เปิดใช้
AI-SEC-๐๐๘	การโจมตีแบบ Adversarial Attack	ผู้โจมตีสร้างข้อมูลป้อนเข้าที่ถูกปรับแต่งเล็กน้อย เช่น สติกเกอร์บนป้ายจราจร เพื่อหลอกล่อให้ปัญญาประดิษฐ์ ของรถยนต์ไร้คนขับตีความผิดพลาด	ปฏิบัติการ / ความปลอดภัย	๒ (ต่ำ)	๕ (สูงมาก)	๑๐ (สูง)	บรรเทา	๑. ใช้เทคนิค Adversarial Training เพื่อให้โมเดลทนทานต่อการโจมตี ๒. เพิ่มกระบวนการตรวจสอบจากเซ็นเซอร์อื่น ๆ เช่น LiDAR, Radar เพื่อยืนยันผล			

หลังจากที่มีทะเบียนความเสี่ยงแล้ว ขั้นตอนต่อไป คือ การกำหนดตัวชี้วัดความเสี่ยงหลัก (Key Risk Indicators - KRIs) เพื่อใช้เป็นระบบแจ้งเตือนล่วงหน้า ทำให้สามารถติดตามและจัดการความเสี่ยงได้อย่างมีประสิทธิภาพก่อนที่มันจะเกิดเป็นปัญหาใหญ่

ความสัมพันธ์ระหว่างทะเบียนความเสี่ยงและตัวชี้วัดความเสี่ยงหลัก

ความสัมพันธ์ของทั้งสองสิ่งนี้เปรียบเสมือน "การวินิจฉัยโรค" กับ "การวัดสัญญาณชีพ"

๑) ทะเบียนความเสี่ยง คือ "ผลการวินิจฉัย" เป็นเอกสารที่ระบุว่าองค์กรมีความเสี่ยงอะไรบ้าง เช่น "มีความเสี่ยงสูงที่จะเป็นโรคหัวใจ" จากการประเมินปัจจัยต่าง ๆ

๒) ตัวชี้วัดความเสี่ยงหลัก คือ "สัญญาณชีพ" เป็นค่าที่วัดผลได้จริงและต่อเนื่อง เช่น อัตราการเต้นของหัวใจ ความดันโลหิต เพื่อติดตาม "ความเสี่ยง" ที่วินิจฉัยไว้นั้น หากค่า KRI ผิดปกติ เช่น ความดันสูงขึ้น ก็เป็นสัญญาณเตือนว่าความเสี่ยงกำลังจะเกิดขึ้นจริง และต้องรีบจัดการ

ดังนั้น ตัวชี้วัดความเสี่ยงหลัก จึงเป็นตัวชี้วัดที่ถูกออกแบบมาเพื่อติดตามความเสี่ยง และวัดประสิทธิผลของมาตรการควบคุมที่กำหนดไว้ในทะเบียนความเสี่ยง

ตัวอย่างการเชื่อมโยงจากทะเบียนความเสี่ยงสู่ตัวชี้วัดความเสี่ยงหลัก

สามารถจับคู่ความเสี่ยงในทะเบียน มาสู่ตัวชี้วัดความเสี่ยงหลักที่วัดผลได้จริง ดังนี้

ตารางที่ ๙ การเชื่อมโยงความเสี่ยงด้านปัญญาประดิษฐ์สู่ตัวชี้วัดความเสี่ยงหลัก (KRI Mapping)

รหัสความเสี่ยง (จากทะเบียนความเสี่ยง)	ชื่อความเสี่ยง	มาตรการควบคุม (ตัวอย่าง)	ตัวชี้วัดความเสี่ยงหลักที่เกี่ยวข้อง	เกณฑ์แจ้งเตือน
AI-SEC-๐๐๑	Data Poisoning Attack	ใช้ Anomaly Detection ตรวจจับข้อมูลผิดปกติ	● จำนวนครั้งที่ระบบตรวจจับข้อมูลผิดปกติต่อสัปดาห์	**> ๑๐ ครั้ง/สัปดาห์ (Yellow 🟡) > ๒๐ ครั้ง/สัปดาห์ (Red 🔴)**
AI-SEC-๐๐๒	ผลลัพธ์ที่มีอคติ	ใช้เครื่องมือตรวจสอบอคติ	● ค่าคะแนนความเท่าเทียมของโมเดล เช่น Disparate Impact Ratio	**< ๐.๙ (Yellow 🟡) < ๐.๘ (Red 🔴)**
AI-SEC-๐๐๓	Prompt Injection	ใช้เทคนิค Input Sanitization และ Output Filtering	● ร้อยละ ของ Prompt ที่ถูกปฏิเสธ โดยระบบ Filter	**> ร้อยละ ๕ (Yellow 🟡)**
AI-SEC-๐๐๗	การรั่วไหลของข้อมูล	ใช้เทคนิค Differential Privacy จำกัดจำนวนการส่งคำร้อง	● จำนวนการส่งคำร้องเฉลี่ยต่อผู้ใช้ในหนึ่งชั่วโมง	**> ๑,๐๐๐ queries/hr (Red 🔴)**

ตัวอย่างตัวชี้วัดความเสี่ยงหลักเพิ่มเติมในมิติต่าง ๆ

ตัวอย่างตัวชี้วัดความเสี่ยงหลักอื่น ๆ ที่นิยมใช้ในการกำกับดูแลปัญญาประดิษฐ์ มีดังต่อไปนี้

๑) ด้านข้อมูลและความเป็นส่วนตัว

- อัตราผลบวกลวงจากระบบตรวจจับข้อมูลส่วนบุคคลในผลลัพธ์ของปัญญาประดิษฐ์ หากสูงเกินไปอาจแปลว่าโมเดลมีความเสี่ยงที่จะเปิดเผยข้อมูลอ่อนไหว
- ร้อยละของชุดข้อมูลฝึกสอนที่ผ่านการตรวจสอบแหล่งที่มา ตัวเลขที่ต่ำบ่งชี้ความเสี่ยงด้าน Supply Chain Attack

๒) ด้านโมเดลและการปฏิบัติการ

- อัตราการเบี่ยงเบนของความแม่นยำโมเดล หากความแม่นยำลดลงอย่างต่อเนื่อง แสดงว่าโมเดลเริ่มล้าสมัยและเสี่ยงต่อการตัดสินใจผิดพลาด
- จำนวนครั้งที่ตรวจพบและบล็อกการโจมตีแบบ Adversarial Attack บอกถึงภัยคุกคามที่องค์กรกำลังเผชิญ
- ร้อยละขององค์ประกอบของปัญญาประดิษฐ์มีช่องโหว่ความรุนแรงระดับสูง วัดความเสี่ยงจากโครงสร้างพื้นฐานที่ใช้ดำเนินงานปัญญาประดิษฐ์

๓) ด้านบุคลากรและการกำกับดูแล (People & Governance)

- ร้อยละของระบบปัญญาประดิษฐ์ที่ผ่านการทำ Threat Modeling ก่อนนำไปใช้งานจริง ตัวเลขที่ต่ำบ่งชี้ถึงความเสี่ยงที่ไม่ได้ถูกประเมินในขั้นตอนการออกแบบ
- ระยะเวลาเฉลี่ยในการตอบสนองและกู้คืน (Mean Time To Respond/Recover - MTTR) จากเหตุเจาะจงเฉพาะส่วนของปัญญาประดิษฐ์ วัดประสิทธิภาพของทีมในการรับมือเมื่อเกิดปัญหา
- ร้อยละของผู้พัฒนาและผู้ใช้งานที่ผ่านการอบรม AI Security Awareness วัดระดับความตระหนักรู้ด้านความปลอดภัยของบุคลากร ซึ่งเป็นหนึ่งในแนวป้องกันที่สำคัญที่สุด

๔.๓ การปฏิบัติตามกฎหมายและข้อบังคับของไทย

ระบบปัญญาประดิษฐ์ไม่ได้ทำงานอยู่ในสุญญากาศทางกฎหมาย แต่ต้องอยู่ภายใต้กรอบของกฎหมายและกฎระเบียบที่มีผลบังคับใช้ในปัจจุบัน การละเลยการปฏิบัติตามข้อกำหนดเหล่านี้อาจนำไปสู่ความรับผิดทางกฎหมาย โทษปรับ และความเสียหายต่อชื่อเสียงขององค์กรอย่างรุนแรง ดังนั้น การบูรณาการข้อพิจารณาด้านกฎหมายจึงเป็นองค์ประกอบที่ขาดไม่ได้ของการกำกับดูแลปัญญาประดิษฐ์ที่มีประสิทธิภาพ

กฎหมายสำคัญของประเทศไทยสองฉบับที่มีผลกระทบโดยตรงต่อการพัฒนาและใช้งานระบบปัญญาประดิษฐ์ คือ

๑) พระราชบัญญัติการรักษาความมั่นคงปลอดภัยไซเบอร์ พ.ศ. ๒๕๖๒

กฎหมายฉบับนี้มีวัตถุประสงค์เพื่อจัดตั้งกลไกการป้องกัน รับมือ และลดความเสี่ยงจากภัยคุกคามทางไซเบอร์ที่อาจส่งผลกระทบต่อเสถียรภาพและความมั่นคงของประเทศ โดยมุ่งเน้นไปที่การกำกับดูแลโครงสร้างพื้นฐานสำคัญทางสารสนเทศ (Critical Information Infrastructure - CII) ซึ่งเป็นระบบคอมพิวเตอร์ที่จำเป็นต่อการรักษาหน้าที่สำคัญของรัฐและการบริการสาธารณะ

การประยุกต์ใช้กับระบบปัญญาประดิษฐ์ หากระบบปัญญาประดิษฐ์ ถูกนำไปใช้เป็นส่วนหนึ่งหรือใช้ในการควบคุมจัดการระบบที่จัดเป็น CII ระบบปัญญาประดิษฐ์ นั้นจะตกอยู่ภายใต้ข้อบังคับของพระราชบัญญัตินี้โดยอัตโนมัติ องค์กรที่เป็นเจ้าของ (หน่วยงานควบคุมหรือกำกับดูแล) จะต้องปฏิบัติตามมาตรฐานและแนวปฏิบัติขั้นต่ำที่ คณะกรรมการการรักษาความมั่นคงปลอดภัยไซเบอร์แห่งชาติ (กมช.) กำหนด

ตัวอย่างที่เกี่ยวข้องกับความมั่นคงปลอดภัยปัญญาประดิษฐ์

<u>สถานการณ์</u> บริษัทด้านพลังงาน (หน่วยงาน CII) นำระบบปัญญาประดิษฐ์ มาใช้ในการบริหารจัดการโครงข่ายไฟฟ้าอัจฉริยะ เพื่อคาดการณ์ความต้องการใช้ไฟฟ้าและตรวจจับความผิดปกติ

- ข้อกำหนด ระบบปัญญาประดิษฐ์นี้ต้องผ่านการประเมินความเสี่ยงตามกรอบที่ กมช. กำหนด จะต้องมีแผนเผชิญเหตุ (Incident Response Plan) ที่ครอบคลุมสถานการณ์การโจมตีที่มุ่งเป้ามายังปัญญาประดิษฐ์โดยเฉพาะ เช่น การโจมตีแบบ Adversarial Attack ที่พยายามหลอกให้ปัญญาประดิษฐ์ตัดสินใจผิดพลาดจนเกิดไฟฟ้าดับเป็นวงกว้าง และต้องมีการทดสอบเจาะระบบ (Penetration Testing) เป็นประจำ

<u>สถานการณ์</u> สถาบันการเงิน (หน่วยงาน CII) ใช้ระบบปัญญาประดิษฐ์ในการตรวจจับธุรกรรมที่น่าสงสัยและป้องกันการฉ้อโกง

- ข้อกำหนด การเปลี่ยนแปลงหรืออัปเดตโมเดลปัญญาประดิษฐ์ใด ๆ จะต้องผ่านกระบวนการบริหารจัดการการเปลี่ยนแปลง (Change Management) ที่รัดกุมและตรวจสอบได้ การเข้าถึงระบบปัญญาประดิษฐ์ และข้อมูลที่เกี่ยวข้องจะต้องถูกควบคุมและบันทึกอย่างเข้มงวดเพื่อใช้ในการตรวจสอบทางนิติวิทยาศาสตร์ดิจิทัล (Digital Forensics) หากเกิดเหตุการณ์ขึ้น

๒) พระราชบัญญัติคุ้มครองข้อมูลส่วนบุคคล พ.ศ. ๒๕๖๒

PDPA เป็นกฎหมายที่ให้สิทธิแก่บุคคลธรรมดา (เจ้าของข้อมูล) ในการควบคุมการเก็บ ใช้ และเปิดเผยข้อมูลส่วนบุคคลของตน โดยกำหนดหน้าที่และความรับผิดชอบให้แก่องค์กร (ผู้ควบคุมข้อมูลและผู้ประมวลผลข้อมูล) ในการจัดการข้อมูลอย่างโปร่งใส ปลอดภัย และชอบด้วยกฎหมาย

การประยุกต์ใช้กับระบบปัญญาประดิษฐ์ วงจรชีวิตของระบบปัญญาประดิษฐ์ทั้งหมดมีความเกี่ยวข้องกับ PDPA อย่างแนบแน่น ตั้งแต่การรวบรวมข้อมูลเพื่อ "ฝึกสอน" โมเดล ไปจนถึงการนำข้อมูลของผู้ใช้มา "ประมวลผล" เพื่อให้ได้ผลลัพธ์ซึ่งทุกขั้นตอนถือเป็นการประมวลผลข้อมูลส่วนบุคคล

๒.๑) ตัวอย่างที่เกี่ยวข้องกับความมั่นคงปลอดภัยปัญญาประดิษฐ์

- ฐานความชอบด้วยกฎหมายในการประมวลผล

<u>สถานการณ์</u> บริษัท E-commerce ต้องการนำประวัติการซื้อของลูกค้ามาใช้ฝึกสอนโมเดลแนะนำสินค้า

- ข้อกำหนด องค์กรไม่สามารถนำข้อมูลดังกล่าวมาใช้ได้ทันที แต่ต้องมีฐานทางกฎหมายรองรับ ซึ่งโดยทั่วไปคือ **"ความยินยอม"** โดยหนังสือขอความยินยอมจะต้องระบุอย่างชัดเจนว่าจะนำข้อมูลไปใช้ "เพื่อพัฒนาและฝึกสอนระบบปัญญาประดิษฐ์สำหรับแนะนำสินค้า" ซึ่งเป็นวัตถุประสงค์ที่เฉพาะเจาะจงและแตกต่างจากการประมวลผลเพื่อส่งสินค้าตามปกติ

- มาตรการรักษาความมั่นคงปลอดภัยของข้อมูล

<u>สถานการณ์</u> โรงพยาบาลพัฒนาปัญญาประดิษฐ์จากข้อมูลประวัติการรักษาของผู้ป่วย ซึ่งเป็นข้อมูลที่มีความอ่อนไหวสูง

- ข้อกำหนด PDPA มาตรา ๓๗ (๑) กำหนดให้ต้องมีมาตรการรักษาความปลอดภัยที่เหมาะสม ในบริบทของปัญญาประดิษฐ์ "ความเหมาะสม" อาจหมายถึง:
    - การใช้เทคนิค Pseudonymization (การทำนามแฝง) กับข้อมูลผู้ป่วยก่อนนำเข้าสู่กระบวนการฝึกสอน
    - การเข้ารหัสลับ ชุดข้อมูลและโมเดลที่ฝึกสอนแล้วขณะจัดเก็บ
    - การออกแบบโมเดลให้ทนทานต่อการโจมตีแบบ Model Inversion ที่อาจทำให้ข้อมูลฝึกสอนรั่วไหลออกมาได้

- สิทธิ์ของเจ้าของข้อมูล

<u>สถานการณ์</u> ผู้ใช้บริการแพลตฟอร์มหนึ่งต้องการใช้สิทธิ์ในการลบข้อมูล

- ข้อกำหนด สิ่งนี้สร้างความท้าทายทางเทคนิคอย่างยิ่งสำหรับปัญญาประดิษฐ์ที่เรียนรู้จากข้อมูลนั้นไปแล้ว องค์กรต้องเตรียมแนวทางรับมือตั้งแต่ขั้นตอนการออกแบบ เช่น การพัฒนาเทคนิค “Machine

Unlearning” เพื่อพยายามลบอิทธิพลของข้อมูลนั้นออกจากโมเดลหรือการกำหนดนโยบายฝึกสอนโมเดลใหม่ทั้งหมดเป็นระยะ ๆ จากชุดข้อมูลที่ผ่านการลบข้อมูลตามคำขอแล้ว

๒.๒) ข้อพิจารณาการประยุกต์ใช้ พ.ร.บ. คุ้มครองข้อมูลส่วนบุคคลกับเทคโนโลยีสมัยใหม่

๒.๒.๑) **การจัดการข้อมูลจากกล้องวงจรปิดในระบบบ้านอัจฉริยะ** ข้อมูลภาพและวิดีโอจากอุปกรณ์กล้องวงจรปิดภายในเคหสถานถือเป็นข้อมูลส่วนบุคคล โดยปกติถือว่าเป็นข้อมูลส่วนบุคคลทั่วไป อย่างไรก็ดี หากมีการสร้างหรือใช้ข้อมูลชีวมิติเพื่อระบุตัวบุคคล เช่น การรู้จำใบหน้าแบบสร้างเทมเพลตหรือการประมวลผล ตามมาตรา ๒๖ ข้อมูลดังกล่าวจะถือเป็นข้อมูลส่วนบุคคลอ่อนไหวและอยู่ภายใต้ข้อกำหนดที่เข้มงวดกว่า เช่น ต้องได้รับความยินยอมโดยชัดแจ้ง เว้นแต่มีข้อยกเว้นตามกฎหมาย ทั้งนี้ การใช้กล้องเพื่อวัตถุประสงค์ส่วนตัวภายในครัวเรือนอาจเข้าข้อยกเว้นกิจกรรมส่วนบุคคลหรือครัวเรือน ตามมาตรา ๔ แต่ผู้ให้บริการหรือแพลตฟอร์มที่ประมวลผลหรือจัดเก็บข้อมูลดังกล่าว เช่น ผู้ให้บริการระบบคลาวด์สำหรับเก็บข้อมูลหรือผู้ผลิตอุปกรณ์ ยังอยู่ภายใต้ข้อบังคับของพระราชบัญญัติคุ้มครองข้อมูลส่วนบุคคล พ.ศ. ๒๕๖๒ สำหรับการโอนข้อมูลไปต่างประเทศ ต้องปฏิบัติตามประกาศคณะกรรมการคุ้มครองข้อมูลส่วนบุคคล เรื่อง หลักเกณฑ์การให้ความคุ้มครองข้อมูลส่วนบุคคลที่ส่งหรือโอนไปยังต่างประเทศ ตามมาตรา ๒๘ แห่งพระราชบัญญัติคุ้มครองข้อมูลส่วนบุคคล พ.ศ. ๒๕๖๒ พ.ศ. ๒๕๖๖ มีผลใช้บังคับตั้งแต่ ๒๔ มีนาคม ๒๕๖๗ โดยอาศัยกลไก เช่น เขตอำนาจหรือองค์การที่มีมาตรฐานคุ้มครองเพียงพอ ตามมาตรา ๒๘ มาตรการคุ้มครองที่เหมาะสม ตามมาตรา ๒๙ และนโยบายคุ้มครองข้อมูลส่วนบุคคลในเครือกิจการหรือเครือธุรกิจเดียวกัน จึงแนะนำให้พิจารณาจัดเก็บประมวลผลในประเทศ เพื่อลดความเสี่ยงการปฏิบัติตามกฎหมาย สำหรับระบบปัญญาประดิษฐ์ที่เกี่ยวข้องกับการรู้จำใบหน้าหรือการวิเคราะห์ที่อาจทำให้เปิดเผยข้อมูลส่วนบุคคลอ่อนไหว ควรกำหนดเป็นกิจกรรม ความเสี่ยงสูงและดำเนินการการประเมินผลกระทบด้านการคุ้มครองข้อมูลส่วนบุคคล ตั้งแต่ระยะเริ่มแรกของโครงการ พร้อมกำหนดมาตรการคุ้มครองที่เหมาะสมอย่างเพียงพอ

๒.๒.๒) **การจัดการข้อมูลจากยานยนต์ไฟฟ้าอัจฉริยะ** ยานยนต์ไฟฟ้าสมัยใหม่เป็นแหล่งกำเนิดข้อมูลปริมาณมหาศาล ซึ่งมีทั้งข้อมูลเชิงเทคนิคของยานพาหนะและข้อมูลที่สามารถระบุถึงตัวบุคคลได้ อาทิ พฤติกรรมการขับขี่ ตำแหน่งทางภูมิศาสตร์ และเส้นทางการเดินทาง ซึ่งถือเป็นข้อมูลส่วนบุคคลภายใต้พระราชบัญญัติคุ้มครองข้อมูลส่วนบุคคล อีกทั้งยังมีศักยภาพที่จะถูกจัดให้เป็นส่วนหนึ่งของโครงสร้างพื้นฐานสำคัญทางสารสนเทศในกลุ่มการคมนาคมและขนส่ง อันจะทำให้อยู่ภายใต้การกำกับของพระราชบัญญัติการรักษาความมั่นคงปลอดภัยไซเบอร์ฯ ในอนาคต ด้วยเหตุนี้ ผู้ประกอบการจึงพึงยึดถือหลักการออกแบบอย่างมั่นคงปลอดภัยเป็นสำคัญในการพัฒนาระบบนิเวศของยานยนต์ไฟฟ้าและพึงพิจารณาเลือกใช้สถาปัตยกรรมที่ส่งเสริมอธิปไตยทางข้อมูลของประเทศ เพื่อสร้างความเชื่อมั่นและเตรียมความพร้อมต่อข้อกำหนดทางกฎหมายที่อาจมีความเข้มข้นขึ้นในอนาคต

การปฏิบัติตามกฎหมายไม่ใช่ทางเลือกแต่เป็นสิ่งจำเป็น องค์กรควรให้ทีมกฎหมาย ทีมกำกับการปฏิบัติตาม และฝ่ายคุ้มครองข้อมูลส่วนบุคคล เข้ามามีส่วนร่วมตั้งแต่ระยะขั้นตอนแนวคิดของโครงการปัญญาประดิษฐ์ เพื่อให้แน่ใจว่านวัตกรรมและการเติบโตขององค์กรจะดำเนินไปบนรากฐานทางกฎหมายที่มั่นคง

๓) พระราชบัญญัติว่าด้วยการกระทำความผิดเกี่ยวกับคอมพิวเตอร์ พ.ศ. ๒๕๕๐ และฉบับแก้ไขเพิ่มเติม

พระราชบัญญัติฉบับนี้เป็นกฎหมายหลักที่กำหนดฐานความผิดและบทลงโทษเกี่ยวกับการใช้ระบบคอมพิวเตอร์ในทางที่มิชอบ โดยมุ่งเน้นการป้องกันและปราบปรามการนำเข้าหรือเผยแพร่เนื้อหาที่ผิดกฎหมาย อันอาจส่งผลกระทบต่อความมั่นคงของประเทศและศีลธรรมอันดีของประชาชน

การประยุกต์ใช้กับระบบปัญญาประดิษฐ์ ทั้งข้อมูลนำเข้า เช่น Prompt และข้อมูลผลลัพธ์ เช่น ข้อความ รูปภาพ วิดีโอ ถือเป็นข้อมูลคอมพิวเตอร์ภายใต้พระราชบัญญัตินี้ ดังนั้นทั้งผู้ให้บริการที่พัฒนาหรือเปิดให้ใช้ระบบปัญญาประดิษฐ์และผู้ใช้งานที่สร้าง Prompt เพื่อให้ระบบปัญญาประดิษฐ์สร้างเนื้อหา ต่างก็มีความรับผิดทางกฎหมายหากผลลัพธ์ที่เกิดขึ้นเข้าข่ายเป็นความผิด

โดยสามารถจำแนกความรับผิดชอบที่อาจเกิดขึ้นได้ดังนี้

๓.๑) ความรับผิดของผู้ให้บริการ โดยหลักการแล้ว ผู้ให้บริการหรือเจ้าของแพลตฟอร์มไม่ต้องรับผิดต่อการกระทำของผู้ใช้งานโดยอัตโนมัติ อย่างไรก็ตาม **มาตรา ๑๕** ได้กำหนดข้อยกเว้นไว้ว่า หากผู้ให้บริการให้ความร่วมมือ ยินยอม หรือรู้เห็นเป็นใจ ให้มีการกระทำความผิดเกิดขึ้นในระบบของตนจะต้องร่วมรับผิดด้วย ตัวอย่างเช่น หากแพลตฟอร์ม Generative AI ถูกนำไปใช้สร้าง Deepfake เพื่อโจมตีบุคคลอื่นอย่างแพร่หลาย การที่ผู้ให้บริการขาดซึ่งมาตรการป้องกันทางเทคนิคที่สมเหตุสมผล เช่น การกรองผลลัพธ์หรือการขาดกลไกป้องกันผลลัพธ์ที่เป็นอันตราย หรือได้รับแจ้งจากพนักงานเจ้าหน้าที่ให้ลบข้อมูลที่ผิดกฎหมายแล้วแต่ไม่ดำเนินการ ก็อาจถูกตีความได้ว่าเป็นการยินยอมให้เกิดการกระทำความผิดขึ้นได้

๓.๒) ความรับผิดของผู้ใช้งาน ผู้ใช้งานเป็นผู้รับผิดชอบโดยตรงต่อเนื้อหาที่ตนสั่งให้ระบบปัญญาประดิษฐ์สร้างขึ้น และไม่สามารถอ้างว่าระบบปัญญาประดิษฐ์เป็นผู้สร้างเพื่อปฏิเสธความรับผิดได้ โดยมีตัวอย่างการกระทำความผิดที่สำคัญ ดังนี้

๓.๒.๑) การนำเข้าข้อมูลเท็จ หากผู้ใช้งานจงใจสร้าง Prompt เพื่อให้ระบบปัญญาประดิษฐ์สร้างเนื้อหาอันเป็นเท็จที่น่าจะเกิดความเสียหายต่อความมั่นคงทางเศรษฐกิจของประเทศ หรือก่อให้เกิดความตื่นตระหนกแก่ประชาชน แล้วนำเนื้อหานั้นไปเผยแพร่ จะเป็นความผิดโดยตรงตาม**มาตรา ๑๔ (๑)** นำเข้าสู่ระบบคอมพิวเตอร์ซึ่งข้อมูลคอมพิวเตอร์ที่บิดเบือน หรือปลอมไม่ว่าทั้งหมดหรือบางส่วน หรือข้อมูลคอมพิวเตอร์อันเป็นเท็จ โดยประการที่น่าจะเกิดความเสียหายแก่ประชาชนอันมิใช่การกระทำความผิดฐานหมิ่นประมาทตามประมวลกฎหมายอาญา และ **มาตรา ๑๔ (๒)** นำเข้าสู่ระบบคอมพิวเตอร์ซึ่งข้อมูลคอมพิวเตอร์อันเป็นเท็จ โดยประการที่น่าจะเกิดความเสียหายต่อการรักษาความมั่นคงปลอดภัยของประเทศ ความปลอดภัยสาธารณะ ความมั่นคง

ในทางเศรษฐกิจของประเทศ หรือโครงสร้างพื้นฐานอันเป็นประโยชน์สาธารณะของประเทศ หรือก่อให้เกิดความตื่นตระหนกแก่ประชาชน

๓.๒.๒) การตัดต่อภาพผู้อื่น หากผู้ใช้งานนำภาพของบุคคลอื่นเข้าสู่ระบบ Generative AI พร้อมสั่งการให้ตัดต่อหรือดัดแปลงในลักษณะที่น่าจะทำให้บุคคลนั้นเสียชื่อเสียง ถูกดูหมิ่น หรือถูกเกลียดชัง แล้วนำภาพที่สร้างขึ้นนั้นไปเผยแพร่ การกระทำดังกล่าวจะเข้าข่ายเป็นความผิดตาม **มาตรา ๑๖** เข้าสู่ระบบคอมพิวเตอร์ที่ประชาชนทั่วไปอาจเข้าถึงได้ซึ่งข้อมูลคอมพิวเตอร์ที่ปรากฏเป็นภาพของผู้อื่น และภาพนั้นเป็นภาพที่เกิดจากการสร้างขึ้น ตัดต่อ เติม หรือดัดแปลง ด้วยวิธีการทางอิเล็กทรอนิกส์หรือวิธีการอื่นใด โดยประการที่น่าจะทำให้ผู้อื่นนั้นเสียชื่อเสียง ถูกดูหมิ่น ถูกเกลียดชัง หรือได้รับความอับอาย

๓.๒.๓) ความผิดเกี่ยวกับความมั่นคงแห่งราชอาณาจักร หากผู้ใช้งานนำพระบรมฉายาลักษณ์ พระฉายาลักษณ์ หรือภาพของพระบรมวงศานุวงศ์ เข้าสู่ระบบ Generative AI พร้อมสั่งการให้สร้าง ดัดแปลง หรือตัดต่อภาพในลักษณะที่เป็นการหมิ่นประมาท ดูหมิ่น หรือแสดงความอาฆาตมาดร้าย แล้วนำไปเผยแพร่ การกระทำดังกล่าวถือเป็นความผิดร้ายแรง เนื่องจากเนื้อหาที่สร้างขึ้นนั้นเข้าข่ายเป็นข้อมูลคอมพิวเตอร์อันเป็นความผิดเกี่ยวกับความมั่นคงแห่งราชอาณาจักรตามประมวลกฎหมายอาญา ซึ่งจะมีความผิดตาม **มาตรา ๑๔ (๓)** นำเข้าสู่ระบบคอมพิวเตอร์ซึ่งข้อมูลคอมพิวเตอร์ใด ๆ อันเป็นความผิดเกี่ยวกับความมั่นคง แห่งราชอาณาจักรหรือความผิดเกี่ยวกับการก่อการร้ายตามประมวลกฎหมายอาญา

นอกจากนี้ องค์กรที่อนุญาตให้สามารถใช้งานระบบปัญญาประดิษฐ์ เช่น Generative AI ควรมีกระบวนการป้องกันหรือตรวจสอบเนื้อหาใด ๆ ที่ถูกสร้างขึ้นเพื่อไม่ขัดต่อพระราชบัญญัติว่าด้วยการกระทำความผิดเกี่ยวกับคอมพิวเตอร์ พ.ศ. ๒๕๕๐ และฉบับแก้ไขเพิ่มเติม โดยเฉพาะเนื้อหาอันเป็นเท็จซึ่งอาจส่งผลกระทบต่อความมั่นคงของประเทศ หรือเนื้อหาที่กระทบต่อสถาบันหลักของชาติ เนื้อหาลักษณะดังกล่าวควรถูกจัดให้อยู่ในระดับความเสี่ยงที่ยอมรับไม่ได้ ซึ่งต้องมีนโยบายและมาตรการควบคุมเพื่อห้ามการดำเนินการโดยสิ้นเชิง ดังนั้น การออกแบบกลไกทางเทคนิคเพื่อป้องกันผลลัพธ์ที่เป็นอันตรายและการกรองผลลัพธ์ ประกอบกับการบังคับใช้นโยบายการใช้เทคโนโลยีที่ยอมรับได้ (Acceptable Use Policy: AUP) จึงเป็นมาตรการควบคุมที่จำเป็น

๔) ข้อพิจารณาพิเศษ การรักษาอธิปไตยทางข้อมูลในระบบปัญญาประดิษฐ์

นอกเหนือจากมาตรการความมั่นคงปลอดภัยเชิงเทคนิคแล้ว การกำกับดูแลระบบปัญญาประดิษฐ์จำเป็นต้องคำนึงถึงมิติของ **"อธิปไตยทางข้อมูล"** อย่างเคร่งครัด หมายถึง **อำนาจในการควบคุมและบังคับใช้กฎหมายของประเทศไทยเหนือข้อมูลที่ถูกสร้าง จัดเก็บ และประมวลผลภายในประเทศ** หลักการนี้มีความสำคัญอย่างยิ่งยวดต่อระบบปัญญาประดิษฐ์ซึ่งมีข้อมูลเป็นหัวใจสำคัญในการทำงาน เพื่อให้มั่นใจว่าการใช้ประโยชน์จากปัญญาประดิษฐ์จะไม่นำมาซึ่งความเสี่ยงต่อความมั่นคงและผลประโยชน์ของชาติ

โดยอาศัยอำนาจตามประกาศของคณะกรรมการการรักษาความมั่นคงปลอดภัยไซเบอร์แห่งชาติ (สกมช.) องค์กรพึงพิจารณาหลักการสำคัญ ๓ ประการในการรักษาอธิปไตยทางข้อมูลในระบบปัญญาประดิษฐ์ ดังนี้

๔.๑) หลักการด้านที่ตั้งของข้อมูล

หัวใจหลักของอธิปไตยทางข้อมูล คือการกำหนดให้ข้อมูลที่สำคัญของชาติถูกจัดเก็บและประมวลผลอยู่ภายในอาณาเขตของประเทศไทย ซึ่งเป็นข้อบังคับที่ชัดเจนตาม **ประกาศคณะกรรมการการรักษาความมั่นคงปลอดภัยไซเบอร์แห่งชาติ เรื่อง มาตรฐานด้านการรักษาความมั่นคงปลอดภัยไซเบอร์ระบบคลาวด์ พ.ศ. ๒๕๖๗**

- **ข้อกำหนด** สำหรับหน่วยงานของรัฐและหน่วยงานโครงสร้างพื้นฐานสำคัญทางสารสนเทศ (CII) ที่ใช้บริการคลาวด์สาธารณะ **ต้องใช้ศูนย์ข้อมูลหลักที่ตั้งอยู่ในประเทศไทย**และมีข้อกำหนดสำหรับศูนย์ข้อมูลสำรองให้อยู่ในประเทศหรือภูมิภาคที่กำหนด

- **นัยยะต่อระบบปัญญาประดิษฐ์** หมายความว่า **ชุดข้อมูลที่ใช้ฝึกสอน**และ**ข้อมูลที่เกิดจากการทำงาน**ของระบบปัญญาประดิษฐ์ที่มีความสำคัญต่อองค์กร จะต้องถูกจัดเก็บและประมวลผลบนโครงสร้างพื้นฐานที่อยู่ภายใต้เขตอำนาจกฎหมายไทย เพื่อป้องกันการเข้าถึงข้อมูลโดยไม่ได้รับอนุญาตจากหน่วยงานต่างชาติ และสร้างความเชื่อมั่นว่าข้อมูลของคนไทยจะถูกคุ้มครองตามกฎหมายไทย

๔.๒) หลักการจำแนกข้อมูลตามระดับผลกระทบ

อธิปไตยทางข้อมูลไม่ได้หมายความว่าข้อมูลทุกประเภทต้องถูกควบคุมด้วยความเข้มข้นระดับสูงสุดเท่ากันทั้งหมด แต่ต้องจำแนกตามความสำคัญและผลกระทบที่อาจเกิดขึ้น ซึ่งสอดคล้องกับประกาศคณะกรรมการการรักษาความมั่นคงปลอดภัยไซเบอร์แห่งชาติ เรื่อง มาตรฐานการกำหนดคุณลักษณะความมั่นคงปลอดภัยไซเบอร์ให้แก่ข้อมูลหรือระบบสารสนเทศ พ.ศ. ๒๕๖๖ ดังนี้

๔.๒.๑) กระบวนการ องค์กรมีหน้าที่ประเมินและจัดระดับผลกระทบของข้อมูลและระบบสารสนเทศออกเป็น **ระดับต่ำ กลาง และสูง** โดยพิจารณาจากวัตถุประสงค์ด้านการรักษาความลับ ความถูกต้องครบถ้วน และสภาพพร้อมใช้งาน

๔.๒.๒) นัยยะต่อระบบปัญญาประดิษฐ์ การจำแนกข้อมูลนี้ทำให้องค์กรสามารถกำหนดแนวทางการรักษาอธิปไตยทางข้อมูลได้อย่างเหมาะสม ตัวอย่างเช่น

○ ข้อมูลผลกระทบระดับสูง เช่น ข้อมูลสาธารณสุขที่ใช้ปัญญาประดิษฐ์วินิจฉัยโรค หรือข้อมูลทางการเงินของประชาชนจะต้องใช้มาตรการด้านที่ตั้งของข้อมูลและการควบคุมการเข้าถึงที่เข้มงวดสูงสุด

○ ข้อมูลส่วนบุคคล ตามประกาศฯ เรื่องมาตรฐานระบบคลาวด์กำหนดให้ข้อมูลส่วนบุคคลต้องถูกจัดระดับผลกระทบด้านการรักษาความลับ**ไม่น้อยกว่าระดับกลาง** ซึ่งหมายความว่าปัญญาประดิษฐ์ที่ประมวลผลข้อมูลส่วนบุคคลจะต้องอยู่ภายใต้มาตรการควบคุมที่รัดกุมเป็นอย่างน้อย

○ ข้อมูลผลกระทบระดับต่ำ เช่น ข้อมูลสาธารณะเพื่อใช้ใน Chatbot ทั่วไป อาจมีความยืดหยุ่นในการเลือกใช้บริการคลาวด์ได้มากกว่า แต่ยังคงต้องอยู่ภายใต้ข้อตกลงที่ชัดเจน

๔.๓) หลักการความรับผิดชอบร่วมกันในบริการคลาวด์

เมื่อองค์กรใช้บริการระบบปัญญาประดิษฐ์บนแพลตฟอร์มคลาวด์สาธารณะ โดยเฉพาะจากผู้ให้บริการระดับสากล การรักษาอธิปไตยทางข้อมูลจะกลายเป็น**ความรับผิดชอบร่วมกัน** ระหว่างผู้ใช้บริการ (CSC) และผู้ให้บริการ (CSP)

๔.๓.๑) ข้อกำหนด ผู้ใช้บริการคลาวด์ (องค์กร) มีหน้าที่ต้อง**กำหนดข้อตกลงและตรวจสอบ**ให้แน่ใจว่า ผู้ให้บริการคลาวด์สามารถปฏิบัติตามกฎหมายและข้อบังคับของประเทศไทยได้ ในขณะที่ผู้ให้บริการคลาวด์มีหน้าที่ต้อง**แจ้งให้ผู้ใช้บริการทราบอย่างโปร่งใส**ถึงที่ตั้งทางภูมิศาสตร์ที่สามารถจัดเก็บข้อมูลได้ และเขตอำนาจศาลที่ควบคุมบริการนั้น ๆ

๔.๓.๒) นัยยะต่อระบบปัญญาประดิษฐ์ ก่อนนำข้อมูลสำคัญขององค์กรไปใช้กับบริการปัญญาประดิษฐ์ ของผู้ให้บริการคลาวด์ใด ๆ องค์กรจำเป็นต้อง

○ ตรวจสอบสัญญาและข้อตกลง อย่างละเอียดว่ามีการระบุถึงการปฏิบัติตามกฎหมายไทยและหลักการรักษาอธิปไตยทางข้อมูลหรือไม่

○ ยืนยันที่ตั้งทางภูมิศาสตร์ ของการจัดเก็บและประมวลผลข้อมูล เพื่อให้มั่นใจว่าเป็นไปตามข้อกำหนดของสำนักงานคณะกรรมการการรักษาความมั่นคงปลอดภัยไซเบอร์แห่งชาติ (สกมช.)

○ กำหนดสิทธิ์ในการเข้าถึงข้อมูล ของบุคลากรผู้ให้บริการคลาวด์อย่างชัดเจนในสัญญา

การผนวกหลักการอธิปไตยทางข้อมูลเข้ากับแนวปฏิบัติด้านความมั่นคงปลอดภัยของปัญญาประดิษฐ์ ไม่ใช่การสร้างอุปสรรคต่อการพัฒนาเทคโนโลยี แต่เป็นการ**สร้างกรอบธรรมาภิบาลที่แข็งแกร่ง** เพื่อให้องค์กรและประเทศไทยสามารถใช้ประโยชน์จากปัญญาประดิษฐ์ได้อย่างเต็มศักยภาพ โดยยังคงไว้ซึ่งอำนาจในการปกป้องข้อมูลอันเป็นสมบัติของชาติให้มั่นคงและยั่งยืน การพิจารณาแหล่งที่ตั้งในการสร้าง เก็บรวบรวม และประมวลผลข้อมูลซึ่งเป็นความสำคัญของการรักษาอธิปไตยทางข้อมูลจะต้องพิจารณาจากหลักการสำคัญ ๓ ข้อข้างต้น ไม่ใช่พิจารณาจากระดับความเสี่ยงของระบบปัญญาประดิษฐ์

๔.๔ การตรวจสอบและการรับรอง

การตรวจสอบและการรับรองเป็นกลไกสำคัญในการสร้าง ความเชื่อมั่นและความไว้วางใจให้แก่ผู้มีส่วนได้ส่วนเสียทั้งหมด ไม่ว่าจะเป็นลูกค้า คู่ค้า หน่วยงานกำกับดูแล หรือสาธารณชน

ในขณะที่การปฏิบัติตามแนวทางภายในเป็นสิ่งจำเป็น การตรวจสอบซึ่งสามารถดำเนินการได้ทั้งจากทีมตรวจสอบภายในเพื่อทวนสอบการปฏิบัติงานและโดยผู้ประเมินอิสระจากภายนอกเพื่อยืนยันอย่างเป็นกลาง จะช่วยยืนยันว่ากระบวนการกำกับดูแล การบริหารความเสี่ยง และการควบคุมความมั่นคงปลอดภัยของระบบปัญญาประดิษฐ์นั้นมีอยู่จริงและมีประสิทธิผล

กระบวนการนี้เปรียบเสมือนการพิสูจน์ความรับผิดชอบและเป็นเครื่องมือในการยกระดับวุฒิภาวะด้านการจัดการปัญญาประดิษฐ์ขององค์กร

๑) บทบาทและขอบเขตของการตรวจสอบระบบปัญญาประดิษฐ์

การตรวจสอบระบบปัญญาประดิษฐ์ คือ กระบวนการที่เป็นระบบ เป็นอิสระ และมีการจัดทำเอกสารเพื่อรวบรวมหลักฐานและประเมินอย่างเป็นกลางว่าระบบปัญญาประดิษฐ์ และกระบวนการที่เกี่ยวข้องสอดคล้องกับเกณฑ์ที่กำหนดไว้ เช่น นโยบายองค์กร กฎหมาย มาตรฐานสากลหรือไม่ **โดยเฉพาะอย่างยิ่ง หากระบบปัญญาประดิษฐ์ถูกจัดเป็นส่วนหนึ่งของโครงสร้างพื้นฐานสำคัญทางสารสนเทศ จะต้องปฏิบัติตามข้อกำหนดของสำนักงานคณะกรรมการการรักษาความมั่นคงปลอดภัยไซเบอร์ (สกมช.) อย่างเคร่งครัด** ขอบเขตการตรวจสอบสำหรับปัญญาประดิษฐ์มีความจำเพาะและกว้างกว่าการตรวจสอบความปลอดภัยของซอฟต์แวร์ทั่วไป โดยครอบคลุมถึง

๑.๑) การตรวจสอบตามกรอบการดำเนินงาน

เพื่อให้การตรวจสอบครอบคลุมตลอดวงจรชีวิตของระบบปัญญาประดิษฐ์ ควรมีการกระบวนการตรวจสอบความปลอดภัยของปัญญาประดิษฐ์ ที่สอดคล้องกับกรอบการรักษาความมั่นคงปลอดภัยในแต่ละระยะดังนี้

ระยะที่ ๐ ขั้นตอนแนวคิด

- **วัตถุประสงค์การตรวจสอบ** เพื่อให้แน่ใจว่าโครงการระบบปัญญาประดิษฐ์ถูกริเริ่มขึ้น โดยมีการพิจารณาด้านความมั่นคงปลอดภัย จริยธรรม และข้อกฎหมายอย่างรอบด้านตั้งแต่แรกเริ่ม
- **รายการที่ตรวจสอบ**

- ตรวจสอบ เอกสารระบุวัตถุประสงค์ทางธุรกิจและขอบเขตการใช้งานระบบปัญญาประดิษฐ์ว่าชัดเจนหรือไม่

- ทวนสอบ ว่ามีการประเมินผลกระทบเบื้องต้นด้านจริยธรรมและความเป็นส่วนตัวแล้วหรือไม่

- ตรวจสอบ ว่ามีการระบุข้อกฎหมายและข้อบังคับที่เกี่ยวข้อง เช่น พระราชบัญญัติคุ้มครองข้อมูลส่วนบุคคล พ.ศ. ๒๕๖๒ และพระราชบัญญัติการรักษาความมั่นคงปลอดภัยไซเบอร์ พ.ศ. ๒๕๖๒ ตั้งแต่ต้นหรือไม่

ระยะที่ ๑ การออกแบบอย่างมั่นคงปลอดภัย (Secure Design)

• **วัตถุประสงค์การตรวจสอบ** เพื่อยืนยันว่าสถาปัตยกรรมของระบบปัญญาประดิษฐ์ถูกออกแบบโดยฝังกลไกความปลอดภัย ความเป็นธรรม และความโปร่งใสเข้าไปในตัว

• **รายการที่ตรวจสอบ:**

- ตรวจสอบ ผลลัพธ์จากการทำ Threat Modeling ที่จำเพาะเจาะจงกับปัญญาประดิษฐ์ เช่น ความเสี่ยงด้าน Data Poisoning Model Evasion

- ทวนสอบ การออกแบบสถาปัตยกรรมว่าสอดคล้องกับหลักการ Security & Privacy by Design หรือไม่ เช่น Data Minimization

- ประเมิน การออกแบบกลไกเพื่อความเป็นธรรมและความสามารถในการอธิบายผล สำหรับระบบที่มีความเสี่ยงสูง

ระยะที่ ๒ การพัฒนาอย่างมั่นคงปลอดภัย (Secure Development)

• **วัตถุประสงค์การตรวจสอบ** เพื่อให้มั่นใจว่ากระบวนการสร้างและฝึกสอนโมเดลปัญญาประดิษฐ์ รวมถึงส่วนประกอบอื่น ๆ เป็นไปอย่างมั่นคงปลอดภัย

• **รายการที่ตรวจสอบ**

- **ทวนสอบ** แหล่งที่มาและความถูกต้องของข้อมูลที่ใช้ในการฝึกสอน

- **ตรวจสอบ** มาตรการป้องกันการโจมตีห่วงโซ่อุปทาน เช่น การสแกนช่องโหว่ไลบรารีและ Pre-trained Model

- **ประเมิน** ประสิทธิผลของมาตรการป้องกัน Data Poisoning และกระบวนการจัดการข้อมูลส่วนบุคคลให้ปลอดภัย

- **ตรวจสอบ** การใช้เครื่องมือวิเคราะห์โคด (SAST) สำหรับส่วนประกอบที่เป็นซอฟต์แวร์

ระยะที่ ๓ การทวนสอบด้านความมั่นคงปลอดภัย (Secure Verification)

• **วัตถุประสงค์การตรวจสอบ** เพื่อทดสอบและยืนยันความทนทานของระบบปัญญาประดิษฐ์ต่อการโจมตีและสภาวะที่ไม่พึงประสงค์ก่อนการนำไปใช้งานจริง

• **รายการที่ตรวจสอบ**

- **ทวนสอบ** รายงานผลการทดสอบเจาะระบบและการทดสอบโดย Red Team

- **ตรวจสอบ** หลักฐานการทดสอบความทนทานของโมเดลต่อการโจมตีแบบ Adversarial Attacks

- **ประเมิน** ผลการทดสอบความเป็นธรรมของโมเดลว่าเป็นไปตามเกณฑ์ที่กำหนดไว้หรือไม่

ระยะที่ ๔ การนำไปใช้งานอย่างมั่นคงปลอดภัย (Secure Deployment)

• **วัตถุประสงค์การตรวจสอบ** เพื่อให้แน่ใจว่าสภาพแวดล้อมที่ใช้ในการดำเนินงานระบบปัญญาประดิษฐ์มีการตั้งค่าที่มั่นคงปลอดภัยและรัดกุม

• **รายการที่ตรวจสอบ**

- **ตรวจสอบ** การตั้งค่าความปลอดภัยของโครงสร้างพื้นฐานและคอนเทนเนอร์
- **ทวนสอบ** กระบวนการจัดการข้อมูลสำคัญ เช่น API Keys และ Credentials
- **ประเมิน** ความรัดกุมของกระบวนการอนุมัติและนำโมเดลขึ้นใช้งานจริง

ระยะที่ ๕ การดำเนินงานและการบำรุงรักษา

• **วัตถุประสงค์การตรวจสอบ** เพื่อยืนยันว่าองค์กรมีกระบวนการเฝ้าระวังตอบสนองต่อเหตุการณ์ และบำรุงรักษาระบบปัญญาประดิษฐ์อย่างต่อเนื่อง

• **รายการที่ตรวจสอบ**

- **ประเมิน** ความครอบคลุมของระบบติดตามและเฝ้าระวังทั้งในส่วนของประสิทธิภาพโมเดล และความปลอดภัย
- **ทวนสอบ** แผนการตอบสนองต่อเหตุการณ์ที่ครอบคลุมสถานการณ์เฉพาะของปัญญาประดิษฐ์
- **ตรวจสอบ** กระบวนการอัปเดตและฝึกสอนโมเดลใหม่ว่ามีความมั่นคงปลอดภัยหรือไม่

ระยะที่ ๖ กระบวนการกำจัดและทำลาย

• **วัตถุประสงค์การตรวจสอบ** เพื่อให้แน่ใจว่าเมื่อระบบปัญญาประดิษฐ์ถูกเลิกใช้งาน ข้อมูลและส่วนประกอบทั้งหมดจะถูกทำลายอย่างมั่นคงปลอดภัยตามนโยบาย

• **รายการที่ตรวจสอบ**

- **ตรวจสอบ** ว่ามีกระบวนการในการปลดระวางระบบปัญญาประดิษฐ์อย่างเป็นทางการ
- **ทวนสอบ** หลักฐานการทำลายข้อมูลที่เกี่ยวข้อง เช่น ชุดข้อมูลฝึกสอน Log ว่าสอดคล้องกับนโยบายการเก็บรักษาข้อมูลและ PDPA
- **ยืนยัน** ว่าส่วนประกอบของระบบทั้งหมด เช่น Model Artifacts Container Images ถูกลบออกจากระบบอย่างถาวร

๑.๒) การตรวจสอบเชิงลึกเฉพาะทาง

๑.๒.๑) การตรวจสอบการกำกับดูแลข้อมูล

- วัตถุประสงค์ เพื่อทวนสอบว่าข้อมูลที่ใช้ในวงจรชีวิตของระบบปัญญาประดิษฐ์ถูกจัดการอย่างถูกต้องและมีจริยธรรม
- ตัวอย่างการตรวจสอบ ผู้ตรวจสอบจะขอหลักฐาน แหล่งที่มาของข้อมูล การบันทึกการขอความยินยอมจากเจ้าของข้อมูลตามข้อกำหนด PDPA และทวนสอบกระบวนการ Data Anonymization เพื่อให้แน่ใจว่าข้อมูลส่วนบุคคลได้รับการคุ้มครองอย่างเหมาะสม

๑.๒.๒) การตรวจสอบความเป็นธรรมและอคติ

- วัตถุประสงค์ เพื่อประเมินว่าผลลัพธ์ของโมเดลปัญญาประดิษฐ์ไม่มีการเลือกปฏิบัติหรือมีอคติต่อกลุ่มประชากรกลุ่มใดกลุ่มหนึ่งอย่างไม่เป็นธรรม
- ตัวอย่างการตรวจสอบ ผู้ตรวจสอบอาจใช้ชุดข้อมูลทดสอบอิสระเพื่อประเมินผลลัพธ์ของโมเดล เช่น โมเดลอนุมัติสินเชื่อ โดยเปรียบเทียบอัตราการอนุมัติระหว่างกลุ่มประชากรต่าง ๆ เช่น เพศ อายุ และวิเคราะห์ว่ามีความแตกต่างอย่างมีนัยสำคัญทางสถิติหรือไม่

๑.๒.๓) การตรวจสอบความทนทานและความมั่นคงปลอดภัย

- วัตถุประสงค์ เพื่อประเมินความสามารถของโมเดลในการทนทานต่อการโจมตีที่จำเพาะเจาะจงกับปัญญาประดิษฐ์ และตรวจสอบว่าสถาปัตยกรรมโดยรอบมีความมั่นคงปลอดภัยเพียงพอ
- ตัวอย่างการตรวจสอบ
  - ผู้ตรวจสอบจะทบทวนผลลัพธ์จากการทดสอบโดย Red Team (ระยะที่ ๓) ตรวจสอบว่าองค์กรมีกระบวนการทดสอบโมเดลด้วย Adversarial Samples หรือไม่ และประเมินความรัดกุมของการตั้งค่า API Endpoint ที่ให้บริการโมเดล
  - กระบวนการทดสอบเชิงรุกตรวจสอบว่า องค์กรมีกระบวนการทดสอบโมเดลด้วย Adversarial Samples อย่างสม่ำเสมอหรือไม่ เพื่อประเมินภูมิต้านทานของโมเดลต่อการโจมตีแบบ Evasion หรือ Data Poisoning
  - การตั้งค่า API Endpoint ประเมินความรัดกุมของการตั้งค่า API ที่ให้บริการโมเดล รวมถึงกลไกการพิสูจน์ตัวตน การให้สิทธิ์ และการจำกัดอัตราการเรียกใช้งาน
  - การตรวจสอบโมเดลแบบ Agentic สำหรับโมเดลที่มีความสามารถในการดำเนินการการตรวจสอบจะเพิ่มขอบเขตในเชิงลึก ได้แก่
    - ■ การวิเคราะห์ขอบเขตการดำเนินการประเมินว่าขอบเขตของ Action ที่ Agent สามารถทำได้นั้นถูกจำกัดตามหลักการให้สิทธิ์น้อยที่สุดหรือไม่และมีกลไกป้องกันการดำเนินการที่อยู่นอกเหนือขอบเขตที่กำหนดไว้อย่างไร
    - ■ การประเมินความทนทานของการเชื่อมต่อ ตรวจสอบการเชื่อมต่อไปยังระบบหรือเซิร์ฟเวอร์ภายนอกที่มีความสำคัญ เช่น Mission-Critical Server ว่ามีมาตรการป้องกันที่รัดกุมหรือไม่ เช่น การเข้ารหัสลับการสื่อสาร การพิสูจน์ตัวตนระหว่างระบบ และความสามารถในการทนทานต่อการโจมตีที่อาจส่งผลกระทบต่อความพร้อมใช้งาน

๑.๒.๔) การตรวจสอบความโปร่งใสและคำอธิบาย

- วัตถุประสงค์ เพื่อทวนสอบว่าองค์กรมีความสามารถในการอธิบายการตัดสินใจของระบบปัญญาประดิษฐ์ที่มีความสำคัญสูงได้

- ตัวอย่างการตรวจสอบ สำหรับระบบปัญญาประดิษฐ์วินิจฉัยโรค ผู้ตรวจสอบจะประเมินว่าระบบมีกลไก Explainable AI (XAI) ที่สามารถชี้แจงได้ว่าปัจจัยใดที่โมเดลใช้ในการวินิจฉัย เพื่อให้แพทย์สามารถทำความเข้าใจและทวนสอบความถูกต้องได้

๑.๒.๕) การตรวจสอบโมเดลแบบ Agentic สำหรับโมเดลที่มีความสามารถในการดำเนินการ การตรวจสอบจะเพิ่มขอบเขตในเชิงลึก ได้แก่

- การวิเคราะห์ขอบเขตการดำเนินการ ประเมินว่าขอบเขตของ Action ที่ Agent สามารถทำได้นั้นถูกจำกัดตามหลักการให้สิทธิ์น้อยที่สุดหรือไม่ และมีกลไกป้องกันการดำเนินการที่อยู่นอกเหนือขอบเขตที่กำหนดไว้อย่างไร
- การประเมินความทนทานของการเชื่อมต่อตรวจสอบการเชื่อมต่อไปยังระบบหรือเซิร์ฟเวอร์ภายนอกที่มีความสำคัญ เช่น Mission-Critical Server ว่ามีมาตรการป้องกันที่รัดกุมหรือไม่ เช่น การเข้ารหัสลับการสื่อสาร การพิสูจน์ตัวตนระหว่างระบบ และความสามารถในการทนทานต่อการโจมตีที่อาจส่งผลกระทบต่อความพร้อมใช้งาน

๒) การรับรองตามมาตรฐานสากล

การได้รับการรับรองเป็นเครื่องหมายแสดงการยอมรับอย่างเป็นทางการจากหน่วยงานที่น่าเชื่อถือว่าระบบการจัดการขององค์กรสอดคล้องกับมาตรฐานสากล ซึ่งถือเป็นการสร้างความเชื่อมั่นในระดับสูงสุด

๒.๑) ISO/IEC 42001:2023 (Artificial Intelligence — Management System)

เป็นมาตรฐานสากลฉบับแรกของโลกสำหรับ ระบบการจัดการ ซึ่งมีโครงสร้างคล้ายคลึงกับมาตรฐานที่รู้จักกันดีอย่าง ISO/IEC 27001:2022 (Information Security Management System)

การรับรองหมายถึง องค์กรที่ได้รับการรับรองตามมาตรฐานนี้ได้พิสูจน์ว่ามี กระบวนการที่เป็นระบบและครบวงจร ในการบริหารจัดการปัญญาประดิษฐ์ ตั้งแต่การกำหนดนโยบาย การประเมินความเสี่ยง การจัดการวงจรชีวิตของระบบปัญญาประดิษฐ์ การประเมินผลกระทบ ไปจนถึงการสร้างความสัมพันธ์กับผู้มีส่วนได้ส่วนเสีย การรับรองนี้แสดงให้เห็นว่าองค์กรมีวุฒิภาวะในการกำกับดูแลปัญญาประดิษฐ์

๒.๒) ISO/IEC 23894:2023 (AI — Guidance on risk management)

ไม่ใช่มาตรฐานสำหรับให้การรับรองโดยตรง แต่เป็นมาตรฐานคำแนะนำที่ให้รายละเอียดเชิงลึกเกี่ยวกับกระบวนการและเทคนิคในการบริหารจัดการความเสี่ยงที่จำเพาะเจาะจงกับปัญญาประดิษฐ์

บทบาทในการตรวจสอบ ในระหว่างการตรวจสอบเพื่อรับรองมาตรฐาน ISO/IEC 42001:2023 ผู้ตรวจสอบจะคาดหวังให้องค์กรแสดงหลักฐานว่ากระบวนการประเมินความเสี่ยงปัญญาประดิษฐ์ของตนมีความรัดกุมและครอบคลุม องค์กรสามารถใช้ ISO/IEC 23894:2023 เป็นแนวทางหลักในการดำเนินกระบวนการดังกล่าว เพื่อระบุและประเมินความเสี่ยงที่เป็นเอกลักษณ์ของปัญญาประดิษฐ์ เช่น Data Poisoning, Model Evasion และ Fairness ได้อย่างเป็นระบบ

๓) ประโยชน์ของการตรวจสอบและการรับรอง

การลงทุนในการตรวจสอบและขอการรับรองระบบปัญญาประดิษฐ์ ก่อให้เกิดประโยชน์ที่สำคัญต่อองค์กร ดังนี้

๓.๑) การสร้างความไว้วางใจ เพิ่มความเชื่อมั่นให้แก่ลูกค้าและคู่ค้าว่าระบบปัญญาประดิษฐ์ขององค์กรถูกพัฒนาและใช้งานอย่างมีความรับผิดชอบ

๓.๒) ความได้เปรียบทางการแข่งขัน การได้รับการรับรองตามมาตรฐานสากลสามารถใช้เป็นเครื่องมือทางการตลาดที่สำคัญ

๓.๓) การปรับปรุงอย่างต่อเนื่อง ผลลัพธ์จากการตรวจสอบให้ข้อมูลเชิงลึกที่มีค่าสำหรับนำไปปรับปรุงกระบวนการและมาตรการควบคุมให้ดียิ่งขึ้น

๓.๔) การพิสูจน์ความรับผิดชอบ เป็นหลักฐานที่จับต้องได้ว่าองค์กรได้ใช้ความพยายามอย่างเหมาะสมในการบริหารจัดการความเสี่ยง ซึ่งอาจเป็นประโยชน์อย่างยิ่งในกรณีที่เกิดข้อพิพาททางกฎหมาย

๔.๕ การสื่อสาร การฝึกอบรม และการสร้างวัฒนธรรมความมั่นคงปลอดภัย

เพื่อให้การกำกับดูแลและบริหารจัดการความเสี่ยงด้านปัญญาประดิษฐ์เกิดประสิทธิผลสูงสุด องค์กรพึงดำเนินการด้านการพัฒนาบุคลากร การเตรียมความพร้อมต่อเหตุการณ์ฉุกเฉิน และการสร้างวัฒนธรรมความมั่นคงปลอดภัยอย่างเป็นระบบ ดังนี้

๑) การพัฒนาบุคลากรและสร้างความตระหนักรู้

กำหนดให้มีการจัดทำแผนการฝึกอบรมและสร้างความตระหนักรู้ด้านความมั่นคงปลอดภัยของระบบปัญญาประดิษฐ์สำหรับบุคลากรทุกระดับที่เกี่ยวข้อง โดยมีวัตถุประสงค์เพื่อลดความเสี่ยงอันเกิดจากปัจจัยมนุษย์และเสริมสร้างความเข้าใจในภัยคุกคามรูปแบบใหม่

แนวทางการดำเนินงาน

- **จัดทำหลักสูตร** พัฒนาเนื้อหาหลักสูตรให้สอดคล้องกับบทบาทและความรับผิดชอบของบุคลากร เช่น หลักสูตรสำหรับผู้บริหาร นักวิทยาศาสตร์ข้อมูล นักพัฒนา และผู้ใช้งานทั่วไป
- **เนื้อหาการอบรม** หลักสูตรต้องครอบคลุมประเด็นสำคัญ อาทิ ภัยคุกคามที่จำเพาะเจาะจงกับระบบปัญญาประดิษฐ์ หลักการออกแบบและพัฒนาอย่างมั่นคงปลอดภัย ข้อกำหนดด้านการคุ้มครองข้อมูลส่วนบุคคล และแนวปฏิบัติในการใช้งานระบบปัญญาประดิษฐ์อย่างมีจริยธรรม
- **การประเมินผล** จัดให้มีการประเมินความรู้ความเข้าใจของบุคลากรหลังการฝึกอบรมเพื่อให้มั่นใจว่าบุคลากรมีความพร้อมในการปฏิบัติงานอย่างมั่นคงปลอดภัย

๒) การจัดการและการตอบสนองต่อเหตุการณ์ฉุกเฉิน

จัดทำและทบทวนกระบวนการจัดการเหตุการณ์ฉุกเฉินทางความมั่นคงปลอดภัยไซเบอร์ให้ครอบคลุมสถานการณ์ที่เกิดจากระบบปัญญาประดิษฐ์โดยเฉพาะ เพื่อให้องค์กรสามารถตรวจจับ ตอบสนอง และฟื้นคืนระบบได้อย่างทันท่วงทีและมีประสิทธิภาพ

แนวทางการดำเนินงาน

- **จัดทำคู่มือการปฏิบัติ** พัฒนาคู่มือการปฏิบัติเพื่อตอบสนองต่อเหตุการณ์ฉุกเฉินสำหรับสถานการณ์เฉพาะทางของปัญญาประดิษฐ์ เช่น

- **กรณี Data Poisoning** แนวทางการตรวจจับข้อมูลที่เป็นพิษ การกักกันชุดข้อมูล และกระบวนการฝึกสอนโมเดลใหม่

- **กรณี Model Evasion/Adversarial Attacks** แนวทางการระบุ บล็อกแหล่งที่มาของการโจมตี และการปรับปรุงโมเดลให้มีความทนทาน

- **กรณี Prompt Injection** แนวทางการจำกัดผลกระทบ การปรับปรุงตัวกรอง และการสื่อสารกับผู้ใช้งาน

- **กรณีตรวจพบอคติรุนแรง** แนวทางการปิดระบบชั่วคราว การสืบสวนสาเหตุ และการสื่อสารต่อสาธารณะ

- **การซักซ้อม** กำหนดให้มีการซักซ้อมตามคู่มือการปฏิบัติอย่างสม่ำเสมอ เพื่อทดสอบประสิทธิภาพของกระบวนการและสร้างความคุ้นเคยให้แก่ทีมงาน

๓) การส่งเสริมวัฒนธรรมความมั่นคงปลอดภัย

ส่งเสริมและสนับสนุนให้การรายงานช่องโหว่ ความผิดพลาด หรือเหตุการณ์ที่น่าสงสัยเป็นส่วนหนึ่งของวัฒนธรรมองค์กร เพื่อสร้างความไว้วางใจและทำให้สามารถตรวจจับภัยคุกคามได้อย่างรวดเร็วก่อนที่จะเกิดความเสียหายในวงกว้าง

แนวทางการดำเนินงาน

- **ประกาศใช้นโยบาย** กำหนดและสื่อสารนโยบายการรายงานเหตุการณ์โดยสมัครใจที่จะไม่มีผลในเชิงลบต่อผู้รายงานอย่างเป็นทางการ

- **จัดตั้งช่องทางการรายงาน** สร้างช่องทางการรายงานที่ชัดเจน เข้าถึงง่าย และสามารถรักษาความลับของผู้รายงานได้

- **การนำข้อมูลไปใช้** เน้นย้ำว่าข้อมูลที่ได้รับจากการรายงานจะถูกนำไปใช้เพื่อการวิเคราะห์และปรับปรุงกระบวนการเท่านั้น ไม่ใช่เพื่อตำหนิหรือลงโทษบุคลากรอันจะนำไปสู่การพัฒนาระบบให้มีความมั่นคงปลอดภัยสูงขึ้นอย่างต่อเนื่อง

บรรณานุกรม

AI Governance Center (AIGC) by ETDA. (2024). *Generative AI governance guideline for organizations*. สำนักงานพัฒนาธุรกรรมทางอิเล็กทรอนิกส์. https://www.etda.or.th/th/Our-Service/AIGC/Knowledge.aspx

ENISA. (2023a). *Cybersecurity of AI and standardisation*. European Union Agency for Cybersecurity.

ENISA. (2023b). *Multilayer framework for good cybersecurity practices for AI*. European Union Agency for Cybersecurity.

International Organization for Standardization. (2022). *ISO/IEC 27001:2022 Information security, cybersecurity and privacy protection — Information security management systems — Requirements*. ISO.

International Organization for Standardization. (2023a). *ISO/IEC 23894:2023 Information technology — Artificial intelligence — Guidance on risk management*. ISO.

International Organization for Standardization. (2023b). *ISO/IEC 42001:2023 Information technology — Artificial intelligence — Management system*. ISO.

International Telecommunication Union. (2022). *ITU-T XSTR-SEC-AI: Guidelines for security management of using artificial intelligence technology*. ITU.

OWASP Foundation. (2023). *OWASP top 10 for large language model applications*. https://owasp.org/www-project-top-10-for-large-language-model-applications/

U.S. Department of Homeland Security. (2024). *Mitigating artificial intelligence (AI) risk: Safety and security guidelines for critical infrastructure owners and operators*. DHS.

พระราชบัญญัติการรักษาความมั่นคงปลอดภัยไซเบอร์ พ.ศ. 2562. (2562, 27 พฤษภาคม). *ราชกิจจานุเบกษา*, 136(69 ก).

พระราชบัญญัติคุ้มครองข้อมูลส่วนบุคคล พ.ศ. 2562. (2562, 27 พฤษภาคม). *ราชกิจจานุเบกษา*, 136(69 ก).

ภาคผนวก

ผนวก ก นิยามศัพท์ที่เกี่ยวข้องกับกระบวนการตรวจสอบและประเมินผล

เอกสารฉบับนี้ได้นิยามศัพท์ที่เกี่ยวข้องกับกระบวนการตรวจสอบและประเมินผลตามแนวทางของมาตรฐานสากล เพื่อสร้างความเข้าใจที่ตรงกันและเป็นมาตรฐานในการนำไปปฏิบัติ โดยมีรายละเอียดดังต่อไปนี้

๑) นิยามศัพท์เชิงกระบวนการ

๑.๑) การเฝ้าระวัง

หมายถึง กระบวนการพิจารณาและกำกับดูแลสถานะของระบบ กระบวนการ หรือกิจกรรมอย่างเป็นระบบและต่อเนื่อง เพื่อรวบรวมข้อมูลสถานะการดำเนินงานสำหรับการตรวจจับความผิดปกติ ความเบี่ยงเบนจากเกณฑ์ที่กำหนด หรือแนวโน้มการเปลี่ยนแปลงที่อาจส่งผลกระทบต่อความมั่นคงปลอดภัย การเฝ้าระวังเป็นกิจกรรมพื้นฐานที่จำเป็นสำหรับการควบคุมการปฏิบัติงานและใช้เป็นข้อมูลตั้งต้นในการตอบสนองต่อเหตุการณ์

เอกสารอ้างอิง แนวคิดปรากฏในมาตรฐานระบบการจัดการ เช่น ISO 9001:2015 ข้อ ๙.๑ การเฝ้าระวัง การวัด การวิเคราะห์ และการประเมินผล

๑.๒) การทบทวน

หมายถึง กิจกรรมที่ถูกกำหนดขึ้นตามวาระ เพื่อพิจารณาและตัดสินความเหมาะสม ความเพียงพอ และประสิทธิผลของระบบ นโยบาย หรือมาตรการควบคุมที่ได้นำไปปฏิบัติแล้ว โดยพิจารณาจากข้อมูลที่รวบรวมจากการเฝ้าระวังและการตรวจประเมิน เพื่อให้ฝ่ายบริหารหรือผู้มีอำนาจตัดสินใจสามารถกำหนดทิศทางเชิงกลยุทธ์ อนุมัติการจัดสรรทรัพยากร และปรับปรุงนโยบายให้บรรลุวัตถุประสงค์ที่ตั้งไว้

เอกสารอ้างอิง ISO 9000:2015 ข้อ ๓.๑๑.๒ และเป็นข้อกำหนดหลักในมาตรฐาน เช่น ISO 9001:2015 ข้อ ๙.๓ การทบทวนของฝ่ายบริหาร

๑.๓) การประเมิน

หมายถึง กระบวนการเชิงวิเคราะห์ในการรวบรวม สังเคราะห์ และพิจารณาหลักฐานและข้อมูลเพื่อกำหนดระดับหรือค่าของคุณลักษณะที่สนใจเทียบกับเกณฑ์ที่กำหนด ในบริบทของความมั่นคงปลอดภัยปัญญาประดิษฐ์ การประเมินมักหมายถึง การประเมินความเสี่ยงเพื่อระบุ วิเคราะห์ และประเมินระดับความเสี่ยง หรือ การประเมินผลกระทบเพื่อทำความเข้าใจถึงผลสืบเนื่องที่อาจเกิดขึ้นจากระบบปัญญาประดิษฐ์ต่อองค์กร บุคคล และสังคม

เอกสารอ้างอิง แนวคิดเป็นหัวใจสำคัญของมาตรฐาน ISO 31000:2018 การบริหารความเสี่ยง — แนวทาง (Risk management — Guidelines)

๑.๔) การตรวจประเมิน

หมายถึง กระบวนการที่เป็นระบบ เป็นอิสระ และจัดทำเป็นเอกสาร สำหรับการได้มาซึ่งหลักฐานการตรวจประเมินและการประเมินหลักฐานนั้นอย่างเที่ยงธรรม เพื่อตัดสินระดับความสอดคล้องกับหลักเกณฑ์การตรวจประเมินที่กำหนดไว้ล่วงหน้า เช่น ข้อกำหนดของมาตรฐาน นโยบายองค์กรหรือข้อกฎหมาย วัตถุประสงค์หลักของการตรวจประเมินคือการยืนยันความสอดคล้องและสร้างความเชื่อมั่นให้แก่ผู้มีส่วนได้ส่วนเสีย

เอกสารอ้างอิง ISO 19011:2018 แนวทางสำหรับระบบการจัดการการตรวจประเมิน (Guidelines for auditing management systems)

ตารางที่ ๑๐ สรุป Monitoring, Review, Assessment, Audit

กิจกรรม	คำนิยามหลัก	ความถี่	วัตถุประสงค์หลัก
Monitoring	การเฝ้าดูสถานะอย่างต่อเนื่อง	ต่อเนื่อง / Real-Time	เพื่อควบคุมการปฏิบัติงาน (Operational Control)
Review	การตัดสินความเหมาะสม เพียงพอ ประสิทธิผล	ตามช่วงเวลา (Periodic)	เพื่อตัดสินใจเชิงกลยุทธ์ (Strategic Decision)
Assessment	การรวบรวมข้อมูลเพื่อประเมินความเสี่ยง/ความสามารถ	ตามความจำเป็น	เพื่อทำความเข้าใจเชิงลึก (In-depth Understanding)
Audit	การตรวจสอบความสอดคล้องอย่างเป็นระบบและเป็นอิสระ	ตามแผนที่กำหนด	เพื่อยืนยันความสอดคล้อง (Verify Conformity)

๒) ประเภทของการตรวจประเมิน

การตรวจประเมินสามารถจำแนกตามแนวทางของมาตรฐาน ISO 19011:2018 ได้เป็น ๓ ประเภทหลัก ดังนี้

๒.๑) การตรวจประเมินโดยบุคคลที่หนึ่ง (First-Party Audit) หรือ การตรวจประเมินภายใน (Internal Audit)

หมายถึง การตรวจประเมินที่ดำเนินการโดยบุคลากรขององค์กรเอง หรือโดยบุคคลภายนอกในนามขององค์กร มีวัตถุประสงค์หลักเพื่อการทวนสอบประสิทธิผลของระบบการควบคุมภายในการค้นหาโอกาสในการปรับปรุง และการเตรียมการสำหรับการตรวจประเมินโดยหน่วยงานภายนอก

๒.๒) การตรวจประเมินโดยบุคคลที่สอง (Second-Party Audit)

หมายถึง การตรวจประเมินที่ดำเนินการโดย หรือในนามของผู้มีส่วนได้ส่วนเสียที่มีผลประโยชน์โดยตรงกับองค์กร เช่น ลูกค้า คู่สัญญา หรือหน่วยงานกำกับดูแล วัตถุประสงค์หลักคือการบริหารจัดการห่วงโซ่อุปทาน (Supply Chain Management) การทวนสอบการปฏิบัติตามข้อกำหนดในสัญญา และการสร้างความเชื่อมั่นระหว่างองค์กรกับคู่ค้า

๒.๓) การตรวจประเมินโดยบุคคลที่สาม (Third-Party Audit)

หมายถึง การตรวจประเมินที่ดำเนินการโดยองค์กรอิสระภายนอก ซึ่งโดยทั่วไปคือหน่วยงานให้การรับรอง (Certification Body - CB) ที่ได้รับการขึ้นทะเบียนรับรองคุณสมบัติ วัตถุประสงค์หลักคือการให้การรับรอง (Certification) ตามมาตรฐานสากล เช่น ISO/IEC 27001:2022, ISO/IEC 42001:2023 หรือเพื่อยืนยันการปฏิบัติตามกฎหมายและข้อบังคับที่เกี่ยวข้อง

ตารางที่ ๑๑ สรุปประเภทของการตรวจประเมิน (Audit Types)

ประเภท	ผู้ตรวจประเมิน	วัตถุประสงค์หลัก
1st Party	องค์กรตรวจตนเอง	การปรับปรุงภายใน / เตรียมความพร้อม
2nd Party	ลูกค้า / ผู้มีส่วนได้ส่วนเสีย	การประเมินคู่ค้า / ตรวจสอบตามสัญญา
3rd Party	หน่วยงานรับรองอิสระ (CB)	การให้การรับรองมาตรฐาน / การปฏิบัติตามกฎหมาย

ผนวก ข ตัวอย่างรายการตรวจสอบความมั่นคงปลอดภัยสำหรับระบบปัญญาประดิษฐ์ (AI Security Checklist Example)

รายการตรวจสอบนี้ถูกออกแบบมาเพื่อเป็นเครื่องมือสำหรับทีมพัฒนา ทีมความมั่นคงปลอดภัย และผู้มีส่วนได้ส่วนเสียในการทวนสอบว่ากิจกรรมด้านความมั่นคงปลอดภัยที่จำเป็นได้ถูกนำไปปฏิบัติในทุกระยะของวงจรชีวิตของระบบปัญญาประดิษฐ์หรือไม่ องค์กรควรปรับแก้และเพิ่มเติมรายการตรวจสอบเหล่านี้ให้สอดคล้องกับบริบท ระดับความเสี่ยง และข้อกำหนดของแต่ละโครงการ

ระยะที่ ๐ ขั้นตอนแนวคิด (Concept Phase)

วัตถุประสงค์ เพื่อวางรากฐานด้านการกำกับดูแลและความมั่นคงปลอดภัยตั้งแต่ก่อนเริ่มโครงการ

รหัส	รายการตรวจสอบ	สถานะ
CPT-01	ได้กำหนดวัตถุประสงค์ ขอบเขต และข้อจำกัดการใช้งานของระบบปัญญาประดิษฐ์ อย่างชัดเจนแล้ว	☐
CPT-02	ได้ระบุและจัดทำรายการสินทรัพย์ที่สำคัญทั้งหมด เช่น ข้อมูล โมเดล ชื่อเสียง แล้ว	☐
CPT-03	ได้ดำเนินการประเมินความเสี่ยงเบื้องต้นสำหรับระบบปัญญาประดิษฐ์ ตามกรอบการทำงาน เช่น ISO 23894:2023 แล้ว	☐
CPT-04	ได้ดำเนินการประเมินผลกระทบด้านการคุ้มครองข้อมูลส่วนบุคคล (DPIA) (กรณีที่เกี่ยวข้องกับข้อมูลส่วนบุคคล) แล้ว	☐
CPT-05	ได้กำหนดบทบาทและความรับผิดชอบ (Roles & Responsibilities) ด้านความปลอดภัยสำหรับโครงการปัญญาประดิษฐ์แล้ว	☐

ระยะที่ ๑ การออกแบบอย่างมั่นคงปลอดภัย (Secure Design Phase)

วัตถุประสงค์ เพื่อแปรเปลี่ยนข้อกำหนดด้านความปลอดภัยให้กลายเป็นสถาปัตยกรรมที่รัดกุม

รหัส	รายการตรวจสอบ	สถานะ
DSG-01	ได้จัดทำแบบจำลองภัยคุกคาม (Threat Model) ที่จำเพาะเจาะจงกับปัญญาประดิษฐ์และได้รับการทบทวนแล้ว	☐
DSG-02	สถาปัตยกรรมได้รวมมาตรการควบคุมเพื่อลดความเสี่ยงที่ระบุในแบบจำลองภัยคุกคาม (Threat Model) แล้ว เช่น กลไกป้องกัน Prompt Injection	☐
DSG-03	การออกแบบได้ระบุถึงการใช้เทคนิค Privacy-Preserving ML เช่น Data Anonymization, Differential Privacy แล้ว	☐
DSG-04	การออกแบบสิทธิ์การเข้าถึงข้อมูลและ API เป็นไปตามหลักการสิทธิ์น้อยที่สุด (Principle of Least Privilege)	☐
DSG-05	สถาปัตยกรรมมีการแบ่งแยกสภาพแวดล้อมระหว่าง Training และ Inference อย่างชัดเจน	☐
DSG-06	ได้กำหนดค่าเริ่มต้นที่ปลอดภัย (Secure Defaults) สำหรับ API ทั้งหมด เช่น บังคับใช้ Authentication, Rate Limiting แล้ว	☐
DSG-07	ได้มีการออกแบบและจำกัดขอบเขตการดำเนินการ (Action Space) ของ Agentic AI ตามหลักการสิทธิ์น้อยที่สุด (Principle of Least Privilege) แล้ว	☐

ระยะที่ ๒ การพัฒนาอย่างมั่นคงปลอดภัย (Secure Development Phase)

วัตถุประสงค์ เพื่อสร้างระบบปัญญาประดิษฐ์ จากพิมพ์เขียวที่ปลอดภัยและจัดการความเสี่ยงของส่วนประกอบทั้งหมด

รหัส	รายการตรวจสอบ	สถานะ
DEV-01	ได้จัดทำผลลัพธ์เชิงเอกสารที่ระบุส่วนประกอบซอฟต์แวร์ทั้งหมดสำหรับไลบรารีและส่วนประกอบของระบบปัญญาประดิษฐ์ตามแนวปฏิบัติสากลหรือแนวทางที่องค์กรได้นำมากำหนดและประยุกต์ใช้แล้ว	☐
DEV-02	ไลบรารีของบุคคลที่สามทั้งหมดได้ผ่านการสแกนหาช่องโหว่ (SCA - Software Composition Analysis) แล้ว	☐
DEV-03	ซอร์สโคดที่พัฒนาขึ้นได้ผ่านการสแกนหาช่องโหว่ (SAST - Static Application Security Testing) แล้ว	☐
DEV-04	สินทรัพย์ปัญญาประดิษฐ์ทั้งหมด (ข้อมูล โมเดล โคด) ถูกจัดเก็บในที่ปลอดภัยและมีการควบคุมเวอร์ชัน	☐
DEV-05	ข้อมูลลับ (Secrets) เช่น API Keys และ System Prompts ถูกจัดเก็บใน Vault และไม่ได้ถูกเขียนลงในโคดโดยตรง	☐
DEV-06	หากมีการใช้ Pre-trained Model ให้มีการทวนสอบแหล่งที่มาและความสมบูรณ์ของโมเดล	☐

ระยะที่ ๓ การทวนสอบด้านความมั่นคงปลอดภัย (Secure Verification Phase)

วัตถุประสงค์ เพื่อทดสอบและยืนยันว่าระบบที่สร้างขึ้นมีความมั่นคงปลอดภัยจริงก่อนนำไปใช้งาน

รหัส	รายการตรวจสอบ	สถานะ
VER-01	มาตรการควบคุมความปลอดภัยที่ระบุในระยะออกแบบ ได้รับการทวนสอบว่าถูกนำไปใช้อย่างถูกต้องและมีประสิทธิผล	☐
VER-02	โมเดลได้ผ่านการทดสอบความทนทานต่อการโจมตีแบบ Evasion Attacks (Adversarial Testing) แล้ว	☐
VER-03	ได้มีการทดสอบเจาะระบบ LLM ด้วยเทคนิค Prompt Injection และ Jailbreaking โดยเฉพาะแล้ว	☐
VER-04	ได้มีการตรวจสอบอคติและความเป็นธรรม (Bias and Fairness Audit) ของโมเดลแล้ว	☐
VER-05	ได้มีการทดสอบเพื่อหาความเสี่ยงการรั่วไหลของข้อมูลฝึกสอน เช่น การโจมตีแบบ Model Inversion แล้ว	☐
VER-06	ระบบโดยรวมได้ผ่านการทดสอบเจาะระบบ (Penetration Test) โดยผู้เชี่ยวชาญแล้ว	☐
VER-07	ได้มีการทดสอบเจาะระบบเพื่อพยายามควบคุม Agentic AI ให้ออกคำสั่งที่เป็นอันตราย (Malicious Action) นอกเหนือขอบเขตที่ได้รับอนุญาตแล้ว	☐

ระยะที่ ๔ การนำไปใช้งานอย่างมั่นคงปลอดภัย (Secure Deployment Phase)

วัตถุประสงค์ เพื่อส่งมอบระบบที่ผ่านการทวนสอบแล้วเข้าสู่สภาพแวดล้อม Production อย่างปลอดภัย

รหัส	รายการตรวจสอบ	สถานะ
DEP-01	สภาพแวดล้อม Production ได้รับการเสริมความแข็งแกร่ง (Hardening) ตามมาตรฐานความปลอดภัยขององค์กร	☐
DEP-02	ไฟล์โมเดลและข้อมูลที่ละเอียดอ่อนทั้งหมดถูกเข้ารหัสลับขณะจัดเก็บ (Encryption at Rest)	☐
DEP-03	สิทธิ์การเข้าถึง (IAM Roles) ใน Production ถูกกำหนดตามหลักการสิทธิ์น้อยที่สุด (Least Privilege)	☐
DEP-04	กระบวนการนำขึ้นระบบเป็นแบบอัตโนมัติและผ่านกระบวนการบริหารจัดการการเปลี่ยนแปลง (Change Management) ที่เป็นทางการ	☐
DEP-05	ได้มีการตรวจสอบการตั้งค่าความปลอดภัย (Configuration Audit) ของสภาพแวดล้อม Production ครั้งสุดท้ายก่อน Go-live	☐

ระยะที่ ๕ การดำเนินงานและการบำรุงรักษาอย่างมั่นคงปลอดภัย (Secure Operation & Maintenance Phase)

วัตถุประสงค์ เพื่อรักษาและปรับปรุงสถานะความมั่นคงปลอดภัยของระบบปัญญาประดิษฐ์ ตลอดอายุการใช้งาน

รหัส	รายการตรวจสอบ	สถานะ
OPS-01	มีระบบเฝ้าระวัง (Monitoring) พฤติกรรมของโมเดล เช่น Data Drift และความปลอดภัยของระบบอย่างต่อเนื่อง	☐
OPS-02	มีการบันทึก Log การทำงานและการเข้าถึงทั้งหมดไปยังระบบที่ปลอดภัยและตรวจสอบได้	☐
OPS-03	มีกระบวนการ MLOps ที่ปลอดภัยสำหรับการฝึกสอนและอัปเดตโมเดลใหม่	☐
OPS-04	มีแผนและมีการซ้อมรับมือเหตุการณ์ (Incident Response Plan) ที่ครอบคลุมสถานการณ์ของปัญญาประดิษฐ์	☐
OPS-05	มีกระบวนการจัดการแพตช์ (Patch Management) สำหรับช่องโหว่ที่ถูกค้นพบในส่วนประกอบต่าง ๆ ของระบบ	☐
OPS-06	มีระบบเฝ้าระวังและบันทึก Log การดำเนินการ (Actions) ทั้งหมดที่ Agentic AI กระทำโดยอัตโนมัติ เพื่อการตรวจสอบย้อนหลัง	☐

ระยะที่ ๖ กระบวนการกำจัดและทำลาย (Disposal Phase)

วัตถุประสงค์ เพื่อปลดระวางระบบและสินทรัพย์ทั้งหมดอย่างปลอดภัยเพื่อป้องกันความเสี่ยงคงค้าง

รหัส	รายการตรวจสอบ	สถานะ
DSP-01	มีแผนการปลดระวางระบบปัญญาประดิษฐ์ และสินทรัพย์ที่เกี่ยวข้องที่ได้รับอนุมัติแล้ว	☐
DSP-02	ข้อมูลฝึกสอน ข้อมูลผู้ใช้ และไฟล์โมเดลทั้งหมดถูกทำลายอย่างปลอดภัยตามมาตรฐานการทำลายข้อมูล	☐
DSP-03	มีการออกใบรับรองการทำลายข้อมูล (Certificate of Destruction) และจัดเก็บเป็นหลักฐาน	☐
DSP-04	ทรัพยากรบนคลาวด์และฮาร์ดแวร์ที่เกี่ยวข้องถูกปลดระวางและเคลียร์ข้อมูลอย่างปลอดภัย	☐
DSP-05	สิทธิ์การเข้าถึง รวมทั้ง API Keys และบัญชีผู้ใช้บริการที่เกี่ยวข้องกับระบบถูกเพิกถอนทั้งหมด	☐
DSP-06	เอกสารทั้งหมดของโครงการได้ถูกจัดเก็บอย่างปลอดภัยตามนโยบายการเก็บรักษาข้อมูล	☐

ประกอบผนวก ข. รายการตรวจสอบการประเมินภัยคุกคาม (AI Threat Modeling Checklist)

วัตถุประสงค์ เพื่อใช้เป็นแนวทางในการระบุ ประเมิน และวางแผนรับมือภัยคุกคามที่อาจเกิดขึ้นกับระบบปัญญาประดิษฐ์ตั้งแต่ในระยะออกแบบ

ชื่อโปรเจกต์/ระบบปัญญาประดิษฐ์: _____ **วันที่ประเมิน:** _____

ผู้เข้าร่วม: _____

ขั้นตอน	หัวข้อการตรวจสอบ	สถานะ/มาตรการควบคุม (OK/Needs Action)	หมายเหตุ / ภัยคุกคามที่พบ / การประเมินระดับความเสี่ยงหลังมาตรการควบคุม (Likelihood x Impact)
๑) การวิเคราะห์องค์ประกอบ (Decomposition)	☐ ระบุแหล่งข้อมูล ท่อส่งข้อมูล (Data Pipeline) สภาพแวดล้อมที่ใช้ฝึกสอน API และแอปพลิเคชันที่เกี่ยวข้องครบถ้วน		
	☐ จัดทำแผนภาพ Data Flow Diagram (DFD) ของระบบปัญญาประดิษฐ์		
๒) การระบุภัยคุกคาม (Threat Identification)	**ห่วงโซ่อุปทานข้อมูล (Data Supply Chain)**		

ขั้นตอน	หัวข้อการตรวจสอบ	สถานะ/ มาตรการควบคุม (OK/Needs Action)	หมายเหตุ / ภัยคุกคามที่พบ / การประเมินระดับความเสี่ยงหลังมาตรการควบคุม (Likelihood x Impact)
	☐ มีความเสี่ยงจาก Data Poisoning หรือไม่ เช่น การแฝงข้อมูลที่เป็นพิษในชุดข้อมูล		
	☐ มีการควบคุมการเข้าถึงแหล่งข้อมูลดิบอย่างรัดกุมหรือไม่		
	การพัฒนาและฝึกสอนโมเดล (Model Development)		
	☐ มีความเสี่ยงที่โมเดลหรือทรัพย์สินทางปัญญา (IP) จะถูกขโมยหรือไม่ (Model Theft)		
	☐ สภาพแวดล้อมที่ใช้ฝึกสอนโมเดลมีการป้องกันการเข้าถึงโดยไม่ได้รับอนุญาตหรือไม่		
	การให้บริการโมเดล (Model Inference)		
	☐ มีความเสี่ยงจากการโจมตีแบบ Model Evasion หรือไม่ เช่น Adversarial Attacks		

ขั้นตอน	หัวข้อการตรวจสอบ	สถานะ/ มาตรการ ควบคุม (OK/Needs Action)	หมายเหตุ / ภัยคุกคามที่พบ / การประเมินระดับ ความเสี่ยงหลัง มาตรการควบคุม (Likelihood x Impact)
	☐ มีความเสี่ยงจากการโจมตีแบบ Prompt Injection หรือไม่ (สำหรับ LLMs)		
	☐ มีความเสี่ยงที่ข้อมูลอ่อนไหวจะรั่วไหลผ่านผลลัพธ์ของโมเดลหรือไม่ (Data Leakage)		
	☐ มีความเสี่ยงจากการโจมตีแบบ **Denial of Service** ที่ API หรือไม่		
๓) การกำหนดมาตรการควบคุม (Mitigation Planning)	☐ ภัยคุกคามที่พบทั้งหมดมีมาตรการควบคุมที่เหมาะสมแล้วหรือไม่ ☐ ระบุมาตรการควบคุมที่ใช้และความเสี่ยงที่ถูกควบคุม		
	☐ ระบุผู้รับผิดชอบและกำหนดเวลาในการแก้ไขช่องโหว่ที่พบ		

ผนวก ค รายการตรวจสอบเพื่อการพัฒนาอย่างมั่นคงปลอดภัย
(Secure Coding Checklist)

คำแนะนำในการใช้งาน รายการตรวจสอบนี้ออกแบบมาสำหรับนักพัฒนาซอฟต์แวร์และนักวิทยาศาสตร์ข้อมูล เพื่อใช้เป็นแนวปฏิบัติในระหว่างการเขียนโคดและการพัฒนาระบบ ควรถูกผนวกเป็นส่วนหนึ่งของกระบวนการ Code Review และการตรวจสอบอัตโนมัติ (CI/CD Pipeline)

วัตถุประสงค์ เพื่อลดช่องโหว่ด้านความปลอดภัยตั้งแต่ขั้นตอนการพัฒนาซอฟต์แวร์

หลักการ

รายการตรวจสอบนี้ถูกพัฒนาขึ้นโดยยึดตามหลักการพื้นฐานด้านความมั่นคงปลอดภัยทางไซเบอร์และธรรมาภิบาลปัญญาประดิษฐ์ที่สำคัญ ดังต่อไปนี้

๑) หลักการออกแบบอย่างมั่นคงปลอดภัย (Secure by Design) เป็นการฝังกระบวนการและมาตรการด้านความมั่นคงปลอดภัยเข้าไปในทุกขั้นตอนของวงจรการพัฒนา ตั้งแต่การออกแบบ การพัฒนา ไปจนถึงการทดสอบ แทนที่จะเป็นการตรวจสอบความปลอดภัยในขั้นตอนสุดท้ายเพียงอย่างเดียว ซึ่งเป็นแนวทางหลักของมาตรฐานสากล เช่น OWASP และ ISO/IEC 27001:2022

๒) หลักการป้องกันเชิงลึก (Defense in Depth) เป็นการใช้มาตรการควบคุมความมั่นคงปลอดภัยที่หลากหลายและซ้ำซ้อนกันในแต่ละชั้นของสถาปัตยกรรม (Data, Application, Host, Network) เพื่อให้หากมีมาตรการใดผิดพลาดไป ก็ยังมีมาตรการอื่นคอยป้องกันอยู่ ซึ่ง Checklist นี้ครอบคลุมตั้งแต่ระดับโคด ข้อมูลนำเข้า ไปจนถึงโครงสร้างพื้นฐาน

๓) หลักการสิทธิ์น้อยที่สุด (Principle of Least Privilege) การกำหนดสิทธิ์ให้แก่ผู้ใช้งาน กระบวนการทำงาน หรือระบบ ในการเข้าถึงข้อมูลและฟังก์ชันต่าง ๆ เท่าที่จำเป็นต่อการปฏิบัติงานเท่านั้น เพื่อจำกัดความเสียหายในกรณีที่เกิดการบุกรุก

๔) หลักการจัดการความเสี่ยง (Risk-Based Approach) การประเมินและเลือกใช้มาตรการควบคุมความมั่นคงปลอดภัยให้เหมาะสมกับระดับความเสี่ยงของระบบปัญญาประดิษฐ์แต่ละระบบ ระบบที่มีความเสี่ยงสูง เช่น ระบบปัญญาประดิษฐ์ทางการแพทย์ ย่อมต้องการการควบคุมที่เข้มงวดกว่าระบบที่มีความเสี่ยงต่ำ ซึ่งเป็นหัวใจสำคัญของมาตรฐาน ISO/IEC 42001:2023

๕) หลักการความมั่นคงปลอดภัยตลอดวงจรชีวิต (Full Lifecycle Security) การบูรณาการความมั่นคงปลอดภัยตลอดวงจรชีวิตของระบบปัญญาประดิษฐ์ ตั้งแต่การจัดหาและเตรียมข้อมูล การฝึกสอนและทดสอบโมเดล การนำไปใช้งาน การเฝ้าระวัง ไปจนถึงการนำระบบออกจากบริการ

รายการตรวจสอบการพัฒนาอย่างมั่นคงปลอดภัย (Secure Coding Checklist)

หมวดหมู่	รายการตรวจสอบ	สถานะ (/)	ตัวอย่างหลักฐานที่เกี่ยวข้อง	หมายเหตุ
๑) การจัดการโคด ข้อมูล และการเรียนรู้	๑.๑ โคดทั้งหมดได้ถูกพัฒนาตามมาตรฐาน Secure Coding Standard ที่องค์กรกำหนดและเป็นที่ยอมรับ	☐	• ชื่อเอกสาร Secure Coding Standard ขององค์กร • ผลการตรวจสอบจากเครื่องมือ Static Analysis (SAST)	
	๑.๒ ไม่มีการฝังข้อมูลยืนยันตัวตน (Credentials) รหัสผ่าน หรือ API Keys ไว้ในซอร์สโคดโดยตรง และมีการใช้เครื่องมือจัดการข้อมูลลับที่เหมาะสม	☐	• Path ไปยังไฟล์ Config ใน Secrets Management Tool • รายงานผลสแกนหา Secrets ในซอร์สโคด	
	๑.๓ มีการใช้ระบบควบคุมเวอร์ชัน เช่น Git สำหรับซอร์สโคด และใช้ Model Registry เพื่อติดตามและบริหารจัดการการเปลี่ยนแปลงของโมเดล	☐	• URL ของ Git Repository • URL ของ Model Registry เช่น MLflow	
	๑.๔ การคุ้มครองความสมบูรณ์ของข้อมูลฝึกสอน (Training Data Integrity)	☐	• บันทึกการเข้าถึง (Access Log) ชุดข้อมูลฝึกสอน • เอกสารแสดงกระบวนการตรวจสอบข้อมูล (Data Validation Pipeline)	

หมวดหมู่	รายการตรวจสอบ	สถานะ (/)	ตัวอย่างหลักฐานที่เกี่ยวข้อง	หมายเหตุ
๒) การจัดการข้อมูลนำเข้า ผลลัพธ์ และปฏิสัมพันธ์	๒.๑ มีการติดตั้งกลไกตรวจสอบความถูกต้องของข้อมูลที่รับมาจากผู้ใช้งานหรือระบบอื่นอย่างครบถ้วน (Input Validation)	☐	• ระบุชื่อไฟล์และฟังก์ชันที่ทำ Input Validation • ผลลัพธ์จาก Unit Test ที่เกี่ยวข้อง	
	๒.๒ สำหรับ LLMs มีการใช้เทคนิค Prompt Sanitization เพื่อกรองและลบคำสั่งที่อาจเป็นอันตรายออกจากข้อมูลนำเข้า	☐	• ระบุชื่อไฟล์และฟังก์ชันที่ทำ Prompt Sanitization • Test Case ที่แสดงการจัดการกับ Malicious Prompt	
	๒.๓ มีการใช้เครื่องมือสแกนหาช่องโหว่ในไลบรารีและมีการอัปเดตอย่างสม่ำเสมอ	☐	• รายงานผลสแกนจากเครื่องมือ เช่น OWASP Dependency-Check, Snyk • ภาพหน้าจอจาก CI/CD Pipeline	
	๒.๔ การจัดการผลลัพธ์ที่ไม่มั่นคงปลอดภัย	☐	• ระบุชื่อฟังก์ชันที่ทำ Output Encoding/Sanitization • กฎของ Content Filtering ที่ใช้งาน	
	๒.๕ การออกแบบปลั๊กอินอย่างมั่นคงปลอดภัย	☐	• เอกสารระบุขอบเขตสิทธิ์ของปลั๊กอิน • Log การยืนยันตัวตนและการเรียกใช้งาน API ของปลั๊กอิน	

หมวดหมู่	รายการตรวจสอบ	สถานะ (/)	ตัวอย่างหลักฐานที่เกี่ยวข้อง	หมายเหตุ
๓) สถาปัตยกรรมและความทนทานของระบบ	๓.๑ มีการแบ่งแยกสภาพแวดล้อมระหว่างการพัฒนา การทดสอบ และการใช้งานจริง อย่างชัดเจน	☐	• แผนภาพสถาปัตยกรรม • ไฟล์ Infrastructure as Code เช่น Terraform	
	๓.๒ มีการจำกัดสิทธิ์การเข้าถึงโมเดลและข้อมูลตามหลักการสิทธิ์น้อยที่สุด	☐	• เอกสาร IAM Policies หรือ Role Definitions • ภาพหน้าจอการตั้งค่า Access Control List (ACL)	
	๓.๓ การจัดการข้อผิดพลาดและข้อยกเว้น	☐	• การตั้งค่าให้แสดงข้อความแสดงข้อผิดพลาดทั่วไป • Log ที่บันทึกรายละเอียดของ Exception สำหรับผู้ดูแลระบบ	
	๓.๔ การจำกัดทรัพยากรและการควบคุมอัตรา	☐	• การตั้งค่า Rate Limiting บน API Gateway • การตั้งค่า Timeout และ Resource Quota ในโคดหรือคอนฟิก	
๔) การติดตาม ตรวจสอบ และกำกับดูแล	๔.๑ มีการติดตั้งกลไกการบันทึก Log และเฝ้าระวังที่ครอบคลุมการเรียกใช้งานและการเข้าถึงโมเดล	☐	• ภาพหน้าจอของ Dashboard • ตัวอย่าง Log ที่บันทึกการเรียกใช้งานโมเดล	

หมวดหมู่	รายการตรวจสอบ	สถานะ (/)	ตัวอย่างหลักฐานที่เกี่ยวข้อง	หมายเหตุ
	๔.๒ มีการดำเนินการทดสอบโมเดลด้วยเทคนิค Adversarial Testing เพื่อประเมินความทนทานของโมเดล	☐	• แผนการทดสอบสำหรับ Adversarial Testing • รายงานสรุปผลการทดสอบจากทีม Security	
	๔.๓ การสนับสนุนความสามารถในการอธิบายได้ (Support for Explainability - XAI)	☐	• เอกสารการออกแบบที่ระบุการบันทึกพารามิเตอร์ที่ใช้ตัดสินใจ • ตัวอย่าง Log ที่สามารถใช้สืบย้อนการทำงานของโมเดลได้	

ภาคผนวก ง ตารางอ้างอิงภัยคุกคามและมาตรฐานควบคุมที่เกี่ยวข้อง (AI Threat and Control Framework Mapping)

ตารางที่ ๑๒ แสดงรายการภัยคุกคามที่สำคัญของระบบปัญญาประดิษฐ์และการอ้างอิงถึงมาตรการควบคุมตามมาตรฐานสากลที่เกี่ยวข้อง

หมวดหมู่ภัยคุกคาม	ภัยคุกคามต่อระบบปัญญาประดิษฐ์	คำอธิบาย	การอ้างอิงมาตรการควบคุม
๑) Data-related Threats	Data Poisoning	การแทรกหรือบิดเบือนข้อมูลฝึกเพื่อทำให้โมเดลเรียนรู้ผิดพลาด	ISO/IEC 27001:2022 A.8.25; ISO/IEC 42001:2023 Data quality & governance; OWASP ML Top 10: ML03
๒) Data-related Threats	Data Leakage / PII Exposure	การรั่วไหลของข้อมูลส่วนบุคคลหรือข้อมูลลับออกจากชุดข้อมูลหรือผลลัพธ์ของโมเดล	ISO/IEC 27001:2022 A.8.12, A.8.24; ISO/IEC 27701:2019 Privacy & PII protection; ISO/IEC 42001:2023 Privacy-by-design; OWASP ML Top 10: ML09
๓) Data-related Threats	Data Quality & Label Errors	การใช้ข้อมูลที่สกปรกป้ายกำกับผิดพลาด ทำให้โมเดลด้อยคุณภาพ	ISO/IEC 27001:2022 A.5.13, A.8.29; ISO/IEC 42001:2023 Data quality requirements; OWASP ML Top 10: ML04

หมวดหมู่ภัยคุกคาม	ภัยคุกคามต่อระบบปัญญาประดิษฐ์	คำอธิบาย	การอ้างอิงมาตรการควบคุม
๔) Model-related Threats	Adversarial Examples	การป้อนอินพุตที่ปรับแต่งเล็กน้อยเพื่อหลอกโมเดลให้ทำนายผิด	ISO/IEC 27001:2022 A.8.27; ISO/IEC 42001:2023 Robustness of AI systems; OWASP ML Top 10: ML01
๕) Model-Related Threats	Membership Inference	การคาดเดาว่าตัวอย่างข้อมูลอยู่ในชุดฝึกหรือไม่	ISO/IEC 27001:2022 A.8.24; ISO/IEC 42001:2023 Privacy safeguards; OWASP ML Top 10: ML08
๖) Model-Related Threats	Model Inversion	การกู้คืนข้อมูลต้นฉบับจากพารามิเตอร์หรือเอาต์พุตของโมเดล	ISO/IEC 27001:2022 A.8.3, A.8.24; ISO/IEC 42001:2023 Privacy & model protection; OWASP ML Top 10: ML07
๗) Model-Related Threats	Model Theft / Extraction	การคัดลอกพฤติกรรมของโมเดลผ่านการยิงคำถามจำนวนมาก	ISO/IEC 27001:2022 A.8.3, A.8.16; ISO/IEC 42001:2023 AI asset protection; OWASP ML Top 10: ML06

หมวดหมู่ภัยคุกคาม	ภัยคุกคามต่อระบบปัญญาประดิษฐ์	คำอธิบาย	การอ้างอิงมาตรการควบคุม
๘) LLM-Specific Threats	Prompt Injection	การฝังคำสั่งในข้อความ/เอกสารเพื่อบังคับ LLM ทำงานผิดวัตถุประสงค์	ISO/IEC 27001:2022 A.8.26, A.8.25; ISO/IEC 42001:2023 AI misuse prevention; OWASP LLM Top 10: LLM01
๙) LLM-Specific Threats	Jailbreak / Abuse Content	การหาวิธีหลีกเลี่ยงข้อจำกัดเพื่อให้ LLM สร้างคำตอบที่ผิดนโยบาย	ISO/IEC 27001:2022 A.5.36, A.8.16; ISO/IEC 42001:2023 AI safety & human oversight; OWASP LLM Top 10: LLM02
๑๐) LLM-Specific Threats	Hallucination & Misinformation	การที่ LLM ให้ข้อมูลที่แต่งขึ้นหรือไม่ถูกต้อง	ISO/IEC 27001:2022 A.5.36; ISO/IEC 42001:2023 Transparency & accuracy of outputs; OWASP LLM Top 10: LLM02
๑๑) Ethics & Governance Threats	Bias & Fairness	โมเดลมีอคติจากข้อมูลที่ไม่สมดุล ทำให้ผลลัพธ์ไม่ยุติธรรม	ISO/IEC 42001:2023 Fairness & non-discrimination; OWASP ML Top 10: ML10
๑๒) Ethics & Governance Threats	Lack of Transparency / Traceability	ขาดการบันทึกและหลักฐานในการตัดสินใจ ทำให้ตรวจสอบไม่ได้	ISO/IEC 27001:2022 A.5.37, A.8.15;

หมวดหมู่ภัยคุกคาม	ภัยคุกคามต่อระบบปัญญาประดิษฐ์	คำอธิบาย	การอ้างอิงมาตรการควบคุม
			ISO/IEC 42001:2023 Traceability & auditability
๑๓) Ethics & Governance Threats	Shadow AI / Unauthorized Tools	การที่พนักงานใช้ AI ภายนอกโดยไม่ได้รับอนุญาต	ISO/IEC 27001:2022 A.5.23, A.5.36; ISO/IEC 42001:2023 AI usage policy governance
๑๔) Infrastructure & Pipeline Threats	Malicious/Compromised Dependencies	การใช้ library/โมเดลภายนอกที่ถูกฝังโคดอันตราย	ISO/IEC 27001:2022 A.5.21, A.5.22, A.8.25; ISO/IEC 42001:2023 Third-party AI components; OWASP ML Top 10: ML05
๑๕) Infrastructure & Pipeline Threats	MLOps Pipeline Compromise	การเข้าถึงหรือยึดระบบ CI/CD, registry, feature store	ISO/IEC 27001:2022 A.5.15, A.5.17, A.8.2; ISO/IEC 42001:2023 AI operations security
๑๖) Infrastructure & Pipeline Threats	API Abuse & Excessive Access	การเรียกใช้งาน API เกินสิทธิ์หรือเจาะระบบ API	ISO/IEC 27001:2022 A.8.20, A.8.21, A.8.5; OWASP API Top 10
๑๗) Monitoring & Operations Threats	Model Drift / Concept Drift	โมเดลเสื่อมสมรรถนะเมื่อข้อมูล/สภาพแวดล้อมเปลี่ยน	ISO/IEC 27001:2022 A.8.16, A.8.15; ISO/IEC 42001:2023 Performance

หมวดหมู่ภัยคุกคาม	ภัยคุกคามต่อระบบปัญญาประดิษฐ์	คำอธิบาย	การอ้างอิงมาตรการควบคุม
			monitoring & continual improvement; OWASP ML Top 10: ML04
๑๘) Monitoring & Operations Threats	Inadequate Monitoring & Logging	ขาดการติดตาม log และตรวจสอบความผิดปกติ ทำให้ภัยคุกคามไม่ถูกค้นพบ	ISO/IEC 27001:2022 A.8.15, A.8.16; ISO/IEC 42001:2023 Monitoring & Audit
๑๙) Agent-Related Threats	Unauthorized Agentic Action	Agentic AI ถูกหลอกลวงหรือทำงานผิดพลาดจนดำเนินการที่เป็นอันตรายต่อระบบอื่น เช่น การลบข้อมูล การเปลี่ยนแปลงค่าคอนฟิก	ISO/IEC 42001:2023 Human oversight, AI safety; ISO/IEC 27001:2022 A.5.15, A.8.2

ผนวก จ รูปแบบรายงานเหตุการณ์ฉุกเฉิน (AI-Specific Incident Response Template)

คำแนะนำในการใช้งาน แม่แบบนี้ควรถูกจัดเก็บในที่ที่ทีมตอบสนองเหตุการณ์สามารถเข้าถึงได้ง่าย ควรมีการซักซ้อมโดยใช้แม่แบบนี้เป็นประจำ เพื่อให้ทีมงานคุ้นเคยและสามารถกรอกข้อมูลสำคัญได้อย่างรวดเร็วและครบถ้วนเมื่อเกิดเหตุการณ์จริง

วัตถุประสงค์ เพื่อให้มีกระบวนการและขั้นตอนการตอบสนองต่อเหตุการณ์ฉุกเฉินด้านความมั่นคงปลอดภัยของปัญญาประดิษฐ์ที่ชัดเจนและเป็นมาตรฐาน

แบบฟอร์มรายงานเหตุการณ์ฉุกเฉินด้านความมั่นคงปลอดภัยระบบปัญญาประดิษฐ์

รหัสเหตุการณ์	IR-AI-[YYYYMMDD]-[NN]	**สถานะ**	Open / In Progress / Closed
วันที่ตรวจพบ		**เวลาที่ตรวจพบ**	
ผู้รายงาน		**ระบบ/โมเดลที่ได้รับผลกระทบ**	

๑) สรุปสำหรับผู้บริหาร (Executive Summary) *(สรุปภาพรวมของเหตุการณ์ ผลกระทบทางธุรกิจ และสถานะการจัดการล่าสุด)*

๒) ประเภทของเหตุการณ์ (Incident Type) *(เลือกข้อที่เกี่ยวข้องที่สุด)*

[ ] Data Poisoning

[ ] Prompt Injection / Jailbreak

[ ] Model Evasion / Adversarial Attack

[ ] Data Privacy Breach

[ ] ตรวจพบอคติรุนแรง (Severe Bias)

[ ] การรั่วไหลของข้อมูลจากโมเดล

[ ] การขโมยโมเดล (Model Theft)

[ ] การดำเนินการโดย Agent ที่ไม่ได้รับอนุญาต (Unauthorized Agentic Action)

[ ] อื่น ๆ : _____

๓) การประเมินผลกระทบ (Impact Assessment)

- **ขอบเขตของเหตุการณ์** ระบุระบบและข้อมูลที่ได้รับผลกระทบ
- **ผลกระทบต่อธุรกิจ** ด้านการเงิน ชื่อเสียง การดำเนินงาน
- **ผลกระทบต่อกฎหมาย** เช่น การละเมิด PDPA
- **ผลกระทบต่อโมเดล** ความแม่นยำ ความเป็นธรรม ความพร้อมใช้งาน
- **ผลกระทบต่อผู้ใช้และบริการ** _____
- **ระดับความรุนแรง** สูง กลาง ต่ำ

๔) ลำดับเหตุการณ์และการตอบสนอง (Timeline and Response Actions) *(บันทึกการดำเนินการต่าง ๆ ตามลำดับเวลา)*

วัน/เวลา	การดำเนินการ	ผู้ดำเนินการ	กำหนดเสร็จ
	☐ **การกักกันเหตุการณ์ (Containment)** *(ตัวอย่าง: API ของโมเดลถูกระงับการใช้งานชั่วคราว)* *(ตัวอย่างการกักกันสำหรับ Agentic AI: เพิกถอน API Keys และระงับสิทธิ์การเข้าถึงระบบอื่นของ Agent ทันที)* ☐ **การแก้ไข (Eradication)** *(ตัวอย่าง: วิเคราะห์และลบข้อมูลที่เป็นพิษ (Poisoned Data) ออกจากชุดข้อมูลฝึกสอนหลัก)* ☐ **การกู้คืน (Recovery)** *(ตัวอย่าง: ทำการ Rollback กลับไปใช้โมเดลเวอร์ชันก่อนหน้า)* *(ตัวอย่าง: ฝึกสอนโมเดลใหม่ (Retrain) ด้วยชุดข้อมูลที่สะอาด และนำเวอร์ชันที่ผ่านการทวนสอบแล้วขึ้นใช้งานแทน)*		

๕) การวิเคราะห์สาเหตุที่แท้จริง (Root Cause Analysis) *(ระบุสาเหตุที่แท้จริงของปัญหา เช่น ช่องโหว่ทางเทคนิค ข้อบกพร่องของกระบวนการ หรือปัจจัยมนุษย์)*

๖) บทเรียนและแผนการแก้ไข (Lessons Learned & Remediation Plan) *(ระบุมาตรการที่จะดำเนินการเพื่อป้องกันไม่ให้เกิดเหตุการณ์ซ้ำ)*

ลำดับ	รายการที่ต้องแก้ไข	ผู้รับผิดชอบ	กำหนดเสร็จ
๑	*(ตัวอย่าง: พัฒนาและติดตั้งระบบ Input Sanitization สำหรับ LLM)*		
๒	*(ตัวอย่าง: ทบทวนและปรับปรุงชุดข้อมูลเพื่อลดอคติ)*		

ผนวก ฉ กรณีศึกษาการประยุกต์ใช้ปัญญาประดิษฐ์ในบริบทประเทศไทย

เพื่อสร้างความเข้าใจในการนำแนวทางไปปรับใช้ให้เกิดผลเป็นรูปธรรม จึงขอยกตัวอย่างกรณีศึกษาการประยุกต์ใช้ปัญญาประดิษฐ์ ความเสี่ยงที่จำเพาะเจาะจง และแนวทางการควบคุมในภาคส่วนสำคัญของประเทศไทย ดังนี้

๑) ภาคการเงิน

ตัวอย่างการประยุกต์ใช้

๑.๑) การตรวจจับการฉ้อโกง (Fraud Detection) ใช้ Machine Learning เพื่อวิเคราะห์พฤติกรรมการทำธุรกรรมและแจ้งเตือนรายการที่ผิดปกติ

๑.๒) การวิเคราะห์สินเชื่อ (Credit Scoring) ใช้ปัญญาประดิษฐ์วิเคราะห์ข้อมูลและพฤติกรรมทางการเงินเพื่อประกอบการพิจารณาวงเงินสินเชื่อ

ความเสี่ยงที่จำเพาะเจาะจง

๑.๓) การบิดเบือนข้อมูล (Data Poisoning) ผู้ไม่หวังดีอาจแทรกข้อมูลธุรกรรมปลอมเข้าไปในชุดข้อมูลฝึกสอน (เปรียบเสมือนการแอบใส่ "วัตถุดิบผิด" ลงในสูตรอาหาร) ส่งผลให้โมเดลตัดสินใจผิดพลาด

๑.๔) อคติของโมเดล (Model Bias) โมเดลอาจให้น้ำหนักกับปัจจัยบางอย่างมากเกินไป ทำให้การพิจารณาสินเชื่อไม่เป็นธรรมต่อประชากรกลุ่มใดกลุ่มหนึ่ง

๑.๕) การขโมยโมเดล (Model Theft) อาชญากรไซเบอร์อาจขโมยโมเดลเพื่อนำไปวิเคราะห์และสร้างระบบเลียนแบบเพื่อหาช่องโหว่

แนวทางการควบคุม

๑.๖) ใช้กระบวนการตรวจสอบความถูกต้องของข้อมูลและใช้ข้อมูลจากแหล่งที่เชื่อถือได้

๑.๗) ประยุกต์ใช้เทคนิค Explainable AI (XAI) (AI ที่อธิบายได้) เพื่อสร้างความโปร่งใสและตรวจสอบกระบวนการอนุมัติสินเชื่อได้

๑.๘) ติดตั้ง Rate Limiting (การจำกัดจำนวนคำขอ) และ API Gateway เพื่อป้องกันการโจมตีด้วยการส่งคำร้องจำนวนมาก

๒) ภาคสาธารณสุข

ตัวอย่างการประยุกต์ใช้

๒.๑) การวินิจฉัยทางการแพทย์ ปัญญาประดิษฐ์ช่วยแพทย์วิเคราะห์ภาพถ่ายทางการแพทย์ เช่น X-ray หรือ MRI เพื่อตรวจหาความผิดปกติ

๒.๒) การแพทย์แม่นยำ ปัญญาประดิษฐ์วิเคราะห์ข้อมูลพันธุกรรมเพื่อเสนอแนวทางการรักษาที่เหมาะสมกับผู้ป่วยแต่ละราย

ความเสี่ยงที่จำเพาะเจาะจง

๒.๓) การรั่วไหลของข้อมูลส่วนบุคคล ข้อมูลสุขภาพของผู้ป่วยเป็นข้อมูลอ่อนไหวสูง หากรั่วไหลจะส่งผลกระทบอย่างรุนแรง

๒.๔) การโจมตีแบบลวง (Adversarial Attack) การปรับแก้ข้อมูลในภาพถ่ายทางการแพทย์เพียงเล็กน้อยที่ตามนุษย์อาจมองไม่เห็น อาจทำให้โมเดลวินิจฉัยผลผิดพลาดได้

๒.๕) การชี้นำพรอมต์ (Prompt Injection) หากใช้ Chatbot ให้คำปรึกษาทางการแพทย์ อาจถูกโจมตีด้วยพรอมต์ที่แฝงคำสั่งอันตราย เปรียบเสมือนการ "แอบบอก" บอกให้ระบบตอบกลับในสิ่งที่ไม่ควร เพื่อให้ระบบให้คำแนะนำที่เป็นอันตราย

แนวทางการควบคุม

๒.๖) ปฏิบัติตามพระราชบัญญัติคุ้มครองข้อมูลส่วนบุคคลอย่างเคร่งครัด

๒.๗) ใช้เทคโนโลยีการเข้ารหัสลับและการควบคุมสิทธิ์การเข้าถึงในการจัดเก็บข้อมูลผู้ป่วย

๒.๘) จัดให้มีการทดสอบแบบ Adversarial Testing เพื่อประเมินความทนทานของโมเดลทางการแพทย์

ผนวก ช ตัวอย่าง นโยบายการใช้เทคโนโลยี Generative AI ที่ยอมรับได้
(Acceptable Use Policy: Generative AI)

(ตัวอย่าง)

ตราสัญลักษณ์หน่วยงาน

หน่วยงาน.....*(ระบุ)*.....

---

๑) บทนำ

ปัจจุบันเทคโนโลยี Generative AI มีความสามารถในการสร้างสรรค์เนื้อหาได้หลากหลายรูปแบบ เช่น ข้อความ ภาพ วิดีโอ เสียง ซอร์สโคด เป็นต้น โดยการสั่งการผ่านข้อความ หรือคำสั่ง (Prompt) ที่ผู้ใช้งานเป็นผู้กำหนด อีกทั้งยังสามารถสร้างเนื้อหาได้อย่างสมจริงและโต้ตอบกับผู้ใช้งานได้คล้ายคลึงกับมนุษย์ อันสามารถช่วยลดระยะเวลาและเพิ่มประสิทธิภาพในการดำเนินงานได้ตั้งแต่การเขียนบันทึกข้อความ การเขียนบทความ การเขียนโปรแกรม การแก้ไขปัญหาด้านเทคนิค การสร้างเอกสารนำเสนอประกอบการประชุม การสร้างสื่อประชาสัมพันธ์ หรือการสร้างเนื้อหาอื่นใดตามแต่ผู้ใช้งานจะสร้างสรรค์

อย่างไรก็ตาม เมื่อมีการนำเทคโนโลยี Generative AI มาใช้ในการรังสรรค์จนเกิดข้อมูลเท็จที่ไม่ได้ตั้งใจให้เกิดความเข้าใจผิด (Misinformation) รวมทั้งข้อมูลเท็จที่จงใจให้เกิดความเข้าใจผิด (Disinformation) กรณีเช่นนี้จะถูกจัดเป็นหนึ่งในความเสี่ยงระดับโลกที่รุนแรงที่สุด ตามรายงาน Global Risks Report ๒๐๒๓ - ๒๐๒๔ ของ World Economic Forum ที่เนื้อหาที่ได้จากการสร้างสรรค์จากเทคโนโลยี Generative AI ที่อาจทำให้ผู้ใช้งานเชื่อได้ว่าข้อมูลนั้นถูกต้อง มีความน่าเชื่อถือและสามารถนำไปใช้งานได้ทันที โดยไม่จำเป็นต้องพิจารณาถึงความเหมาะสมและข้อจำกัดของเทคโนโลยี โดยเฉพาะอย่างยิ่งข้อจำกัดทั้งด้านภาษาไทย ความเข้าใจในบริบทหรือวัฒนธรรมของประเทศไทย หรือบริบทของหน่วยงาน ซึ่งอาจก่อให้เกิดผลกระทบในเชิงลบต่อบุคคล องค์กร สังคมและประเทศชาติ เช่น

๑.๑) ข้อมูลเท็จที่ไม่ได้ตั้งใจให้เกิดความเข้าใจผิด (Misinformation) และข้อมูลเท็จที่จงใจให้เกิดความเข้าใจผิด (Disinformation)

๑.๒) เนื้อหาไม่ถูกต้องตามข้อเท็จจริง (Confabulation)

๑.๓) เนื้อหามีความอันตรายหรือรุนแรง (Dangerous or Violent Recommendations)

๑.๔) เนื้อหาส่งผลกระทบต่อความเป็นส่วนตัวของข้อมูล (Data Privacy)

๑.๕) เนื้อหาส่งผลกระทบต่อประเด็นที่มีความอ่อนไหว และมีเนื้อหาที่ขัดต่อหลักจริยธรรม (Sensitive and Ethics Context)

๑.๖) เนื้อหามีข้อมูลที่ไม่ถูกต้องครบถ้วนหรือถูกทำให้บิดเบือน (Information Integrity)

๑.๗) เนื้อหาที่สร้างก่อให้เกิดอคติ แบ่งแยกและการเลือกปฏิบัติ (Bias and Discrimination)

๑.๘) ข้อมูลส่วนบุคคลรั่วไหล (Personal Data Leakage)

๑.๙) การละเมิดทรัพย์สินทางปัญญา (Intellectual Property Infringement)

๑.๑๐) ความมั่นคงปลอดภัยของข้อมูล (Information Security)

๑.๑๑) ผลกระทบอื่น ๆ ที่อาจเกิดขึ้นจากเทคโนโลยี Generative AI หรือเทคโนโลยีอื่นที่มีลักษณะเดียวกัน

ดังนั้น จึงสมควรกำหนดนโยบายการใช้เทคโนโลยี Generative AI ที่ยอมรับได้ (Acceptable Use Policy: Generative AI) เพื่อนำมารังสรรค์งานของ.....*(ระบุชื่อหน่วยงาน)*.....ขึ้น

๒) วัตถุประสงค์

นโยบายนี้จัดทำขึ้นโดยมีเนื้อหาสาระสำคัญ ที่สามารถนำไปเป็นแนวทางการจัดฝึกอบรม การให้ความรู้ การทราบถึงโทษของการไม่ปฏิบัติตามนโยบาย และตัวอย่างพฤติกรรมที่ควรทำและไม่ควรทำ ของการใช้งานเทคโนโลยี Generative AI รวมถึงแอปพลิเคชันหรือบริการ Generative AI ที่สำนักงานแนะนำ เพื่อให้พนักงาน ลูกจ้าง ผู้ปฏิบัติงานและผู้รับจ้างของสำนักงานสามารถนำเทคโนโลยี Generative AI ไปใช้งานได้อย่างถูกต้องเหมาะสมมีความรับผิดชอบและสามารถนำเทคโนโลยีดังกล่าวไปรังสรรค์งานได้อย่างปลอดภัย มีประสิทธิภาพ ตลอดจนป้องกันความเสี่ยงที่อาจเกิดขึ้นจากการใช้งานที่ไม่เหมาะสม ในการนี้ สำนักงานจึงได้กำหนดวัตถุประสงค์ของนโยบายดังกล่าวไว้ ดังนี้

๒.๑) เพื่อให้พนักงาน ลูกจ้าง ผู้ปฏิบัติงาน และผู้รับจ้างของสำนักงาน มีแนวปฏิบัติในการใช้เทคโนโลยี Generative AI งานอย่างถูกต้อง เหมาะสมและมีประสิทธิภาพ

๒.๒) เพื่อให้การใช้เทคโนโลยี Generative AI เป็นไปตามกฎหมาย กฎ ระเบียบ ประกาศ ข้อบังคับ หรืออื่น ๆ ที่เกี่ยวข้อง ตลอดจนป้องกันความเสี่ยงที่เกิดจากการใช้งานผิดวัตถุประสงค์ที่ก่อให้เกิดผลกระทบที่เกิดขึ้นกับบุคคล สำนักงาน สังคม และประเทศชาติ

๒.๓) เพื่อเสริมสร้างความตระหนักรู้และความเข้าใจเกี่ยวกับการใช้เทคโนโลยี Generative AI อย่างมีความรับผิดชอบ รอบครอบ สามารถใช้งานได้อย่างถูกต้องเหมาะสมและเป็นไปตามแนวทางที่สำนักงานกำหนดไว้

๓) ขอบเขต

ขอบเขตของนโยบายฉบับนี้มีเนื้อหาครอบคลุมสำหรับการนำเทคโนโลยี Generative AI มาใช้เพื่อการรังสรรค์งานตามภารกิจของสำนักงานที่กำหนดให้พนักงาน ลูกจ้าง และผู้ปฏิบัติงาน รวมถึงผู้รับจ้างที่ได้รับมอบหมายให้ดำเนินงานตามภารกิจของสำนักงาน ซึ่งต่อไปนี้เรียกว่า “ผู้ใช้งาน” ที่ใช้งานเทคโนโลยี Generative AI ในการดำเนินการตามภารกิจที่กฎหมายกำหนดไว้

๔) คำนิยาม *(ระบุคำนิยามของหน่วยงานตามความเหมาะสม)*

“สำนักงาน/หน่วยงาน” หมายถึง .....*(ระบุ)*.....

“ผู้บริหารสูงสุด” หมายถึง .....*(ระบุ)*.....

“ผู้บังคับบัญชา” หมายถึง .....*(ระบุ)*.....

“ผู้ปฏิบัติงาน” หมายถึง .....*(ระบุ)*.....

“ผู้รับจ้าง” หมายถึง บุคคลหรือนิติบุคคลที่ลงนามเป็นคู่สัญญากับสำนักงาน รวมทั้งตัวแทน หรือลูกจ้าง หรือผู้รับจ้างช่วง ที่อยู่ในความรับผิดชอบของผู้รับจ้างตามสัญญานั้น ๆ

“นโยบาย” หมายถึง นโยบายการประยุกต์ใช้เทคโนโลยี Generative AI ที่ยอมรับได้

“ปัญญาประดิษฐ์ (AI)” หมายถึง เทคโนโลยีที่ถูกพัฒนาขึ้นเพื่อให้ระบบประมวลผลของคอมพิวเตอร์ หุ่นยนต์ เครื่องจักร หรืออุปกรณ์อิเล็กทรอนิกส์ต่าง ๆ มีคุณสมบัติหรือพฤติกรรมใกล้เคียงมนุษย์ตามวัตถุประสงค์ที่มนุษย์กำหนด เช่น การเรียนรู้ การรับรู้ และตอบสนองต่อสภาพแวดล้อม การให้เหตุผล และการแก้ไขปัญหา เป็นต้น

“เทคโนโลยี Generative AI” หมายถึง เทคโนโลยี AI ประเภทหนึ่งที่มีความสามารถในการสร้างเนื้อหาใหม่ในหลากหลายรูปแบบตามข้อความหรือคำสั่ง (Prompt) ที่มนุษย์เป็นผู้กำหนด เช่น ข้อความ ภาพ วิดีโอ ซอร์สโคด หรือรูปแบบอื่น เป็นต้น

“ข้อมูลส่วนบุคคล” หมายถึง ข้อมูลเกี่ยวกับบุคคลธรรมดาซึ่งทำให้สามารถระบุตัวบุคคลนั้นได้ ไม่ว่าทางตรงหรือทางอ้อม แต่ไม่รวมถึงข้อมูลของผู้ถึงแก่กรรมโดยเฉพาะ

“ข้อมูลส่วนบุคคลอ่อนไหว” หมายถึง ข้อมูลส่วนบุคคลตามที่ถูกบัญญัติไว้ในมาตรา ๒๖ แห่งพระราชบัญญัติคุ้มครองข้อมูลส่วนบุคคล พ.ศ. ๒๕๖๒ ซึ่งได้แก่ ข้อมูลเชื้อชาติ เผ่าพันธุ์ ความคิดเห็นทางการเมือง ความเชื่อในลัทธิ ศาสนาหรือปรัชญา พฤติกรรมทางเพศ ประวัติอาชญากรรม ข้อมูลสุขภาพ ความพิการ ข้อมูลสหภาพแรงงาน ข้อมูลพันธุกรรม ข้อมูลชีวภาพ หรือข้อมูลอื่นใดซึ่งกระทบต่อเจ้าของข้อมูลส่วนบุคคล ในทำนองเดียวกันตามที่คณะกรรมการคุ้มครองข้อมูลส่วนบุคคลประกาศกำหนด

๕) แนวทางการใช้เทคโนโลยี Generative AI ที่ยอมรับได้ *(หน่วยงานสามารถเพิ่มเติมเนื้อหาได้ตามความเหมาะสมเพื่อให้สอดคล้องตามภารกิจของหน่วยงาน)*

๕.๑) การใช้เทคโนโลยี Generative AI ที่ยอมรับได้ สำหรับการดำเนินงานตามภารกิจของสำนักงาน

๕.๑.๑) ผู้ใช้งานต้องทำการศึกษาข้อมูลเทคโนโลยี Generative AI แต่ละประเภท เพื่อสร้างความเข้าใจเกี่ยวกับข้อมูลพื้นฐาน ศักยภาพ ประโยชน์ ความเสี่ยง และข้อจำกัดเพื่อให้ผู้ใช้งานสามารถประยุกต์ใช้เทคโนโลยี Generative AI ได้อย่างเหมาะสม มีประสิทธิภาพและสอดคล้องกับเป้าหมายของงานแต่ละประเภทที่ได้รับมอบหมาย

๕.๑.๒) ผู้ใช้งานต้องประยุกต์ใช้เทคโนโลยี Generative AI เพื่อช่วยสนับสนุนในการดำเนินงานในบริบทที่กำหนด ดังต่อไปนี้

๑) การวิเคราะห์ข้อมูลและสร้างรายงาน

๒) การเขียนชุดคำสั่งหรือโปรแกรม

๓) การจัดทำร่างหนังสือหรือเอกสารต่าง ๆ เช่น บันทึกข้อความ นโยบาย หรือคำแนะนำ เป็นต้น

๔) การให้คำแนะนำหรือแนวทางเบื้องต้นในการแก้ปัญหาด้านต่าง ๆ

๕) การสร้างเนื้อหาสำหรับการสื่อสารภายในองค์กร

๖) การสร้างสื่อประชาสัมพันธ์

๗) การสนับสนุนงานวิจัยและวิชาการ เช่น การช่วยทบทวนวรรณกรรม การวิเคราะห์ข้อมูลเบื้องต้น และการช่วยร่างเอกสารงานวิจัย

๘) ดำเนินงานอื่น ๆ ที่ได้รับมอบหมายจากผู้บังคับบัญชา ทั้งนี้ การประยุกต์ใช้เทคโนโลยี Generative AI ต้องเป็นไปตามภารกิจและเพื่อประโยชน์ของสำนักงานเท่านั้น

๕.๑.๓ ผู้ใช้งานต้องใช้เทคโนโลยี Generative AI อย่างมีธรรมาภิบาลเหมาะสมมีความมั่นคงปลอดภัยและสอดคล้องกับกฎหมาย ข้อบังคับ ระเบียบ ประกาศ หรือคำสั่งของสำนักงานหรืออื่น ๆ ที่เกี่ยวข้อง

๕.๒) ข้อห้ามในการใช้เทคโนโลยี Generative AI

แม้ว่าเทคโนโลยี Generative AI จะเป็นเครื่องมือที่ช่วยสนับสนุนและเพิ่มประสิทธิภาพในการดำเนินงานให้แก่ผู้ใช้งานอยู่หลายประการ แต่อย่างไรก็ตาม การใช้เทคโนโลยี Generative AI จำเป็นต้องอยู่ในขอบเขตที่เหมาะสมและไม่ก่อให้เกิดความเสียหายใด ๆ ต่อบุคคล สำนักงาน สังคม หรือประเทศชาติ ในการนี้ สำนักงานจึงกำหนดข้อห้ามสำหรับผู้ใช้งานเทคโนโลยี Generative AI ไว้ดังนี้

๕.๒.๑) ห้ามใช้แทนการตัดสินใจของผู้ใช้งานในกรณีที่มีความเสี่ยงสูง เช่น การตัดสินใจทางกฎหมาย ทางการแพทย์ ทางการเงิน หรือการตัดสินใจที่อาจส่งผลกระทบต่อชีวิต ทรัพย์สิน และสิทธิของบุคคล

๕.๒.๒) ห้ามใช้เพื่อสร้างข้อมูลอันเป็นเท็จ หรือสร้างเนื้อหาที่อาจก่อให้เกิดความเสียหายต่อบุคคล สำนักงาน หรือสังคม อันอาจนำไปสู่ความเข้าใจผิด หรือสร้างความขัดแย้งในสังคม

๕.๒.๓) ห้ามใช้หรือเปิดเผยข้อมูลที่เป็นความลับของสำนักงาน รวมถึงข้อมูลภายในเอกสารสำคัญหรือข้อมูลที่อาจส่งผลกระทบต่อการดำเนินงานของสำนักงาน

๕.๒.๔) ห้ามใช้ข้อมูลส่วนบุคคลที่ไม่ได้รับความยินยอมจากเจ้าของข้อมูลส่วนบุคคลหรือใช้ข้อมูลที่อาจเป็นความผิดตามพระราชบัญญัติคุ้มครองข้อมูลส่วนบุคคล พ.ศ. ๒๕๖๒ หรือกฎหมายอื่นที่เกี่ยวข้อง

๕.๒.๕) ห้ามใช้ในทางที่ขัดต่อหลักธรรมาภิบาล คุณธรรม จริยธรรม ศีลธรรม หรือมีเจตนาแอบแฝงโดยไม่สุจริต

๕.๒.๖) ห้ามใช้เพื่อสร้างเนื้อหาที่ละเมิดลิขสิทธิ์หรือทรัพย์สินทางปัญญา รวมถึงการทำซ้ำ คัดลอก หรือ ดัดแปลงซึ่งเนื้อหาที่เป็นของบุคคลหรือหน่วยงานอื่นโดยไม่ได้รับอนุญาต

๕.๒.๗) ห้ามใช้เพื่อสร้างเนื้อหาที่ส่งเสริมการเหยียดเชื้อชาติ ศาสนา เพศ วัย ความพิการ หรือสถานะทางสังคม ซึ่งอาจขัดต่อกฎหมายและหลักสิทธิมนุษยชน

๕.๒.๘) ห้ามใช้ในการดำเนินการใด ๆ ที่อาจเป็นการกระทำความผิดตามกฎหมาย กฎ ระเบียบ ข้อบังคับ ประกาศ คำสั่ง หลักเกณฑ์ หรืออื่น ๆ ที่เกี่ยวข้อง

๕.๒.๙) ห้ามใช้เพื่อสร้างเครื่องมือหรือเนื้อหาที่เป็นอันตรายโดยตรง เช่น การสร้างสื่อสังเคราะห์ปลอม (Deepfake) เพื่อโจมตีผู้อื่น การเขียนโคดเพื่อพัฒนามัลแวร์ หรือการสร้างข้อความเพื่อใช้ในการหลอกลวง

๕.๓) การรักษาความลับและคุ้มครองข้อมูลส่วนบุคคล

การใช้งานเทคโนโลยี Generative AI โดยเฉพาะอย่างยิ่งที่มีการให้บริการแบบไม่มีค่าใช้จ่าย อาจมีความเสี่ยงที่ข้อมูลที่ผู้ใช้งานระบุหรือโต้ตอบกับระบบ จะถูกเข้าถึงหรือบันทึกและนำไปใช้ในการฝึกสอนโมเดล Generative AI เพื่อปรับปรุงและพัฒนาประสิทธิภาพการทำงานของระบบ ดังนั้น เพื่อเป็นการป้องกันมิให้เกิดการรั่วไหลของข้อมูล การละเมิดสิทธิของบุคคล หรือก่อให้เกิดความเสียหายต่อสำนักงาน หน่วยงานภายนอก หรือบุคคลที่เกี่ยวข้อง ผู้ใช้งานจึงต้องมีความระมัดระวังและต้องปฏิบัติ ดังนี้

๕.๓.๑) ห้ามนำข้อมูลภายในสำนักงาน และข้อมูลที่มีชั้นความลับ เช่น รหัสผ่าน เอกสารสัญญา เอกสารหรือหนังสือที่ประทับข้อความลับ เอกสารหรือข้อมูลเกี่ยวกับโครงการภายในสำนักงาน เป็นต้น ไปใช้งานร่วมกับเทคโนโลยี Generative AI

๕.๓.๒) ในกรณีที่ต้องใช้เทคโนโลยี Generative AI ร่วมกับข้อมูลภายในองค์กรข้อมูลที่มีชั้นความลับ ข้อมูลส่วนบุคคล หรือข้อมูลส่วนบุคคลที่อ่อนไหว ผู้ใช้งานจะต้องใช้เทคโนโลยี Generative AI ที่สำนักงานประกาศกำหนดเท่านั้น และผู้ใช้งานต้องได้รับอนุมัติจาก.....*(ให้ระบุผู้มีอำนาจอนุมัติได้ตามความเหมาะสม เช่น ผู้อำนวยการ เลขาธิการ หรือผู้บริหารสูงสุดของสำนักงาน)*..... ก่อนนำมาใช้งานทุกกรณี

๕.๓.๓) ในกรณีที่มีข้อสงสัยเกี่ยวกับการรักษาความมั่นคงปลอดภัยของข้อมูลหรือข้อกำหนดตามกฎหมาย ผู้ใช้งานต้องปรึกษาหารือกับผู้บังคับบัญชาหรือเจ้าหน้าที่ที่เกี่ยวข้องก่อนดำเนินการใด ๆ

๕.๔) การรักษาความมั่นคงปลอดภัยของระบบสารสนเทศ กรณีนำเทคโนโลยี Generative AI มาใช้งาน

๕.๔.๑) ห้ามนำข้อมูลที่อาจส่งผลกระทบต่อความมั่นคงปลอดภัยของระบบสารสนเทศ เช่น รหัสผ่าน ข้อมูล API Key รายละเอียดการตั้งค่าของระบบ เป็นต้น ไปใช้งานร่วมกับเทคโนโลยี Generative AI

๕.๔.๒) ต้องตรวจสอบซอร์สโคดที่สร้างโดยเทคโนโลยี Generative AI ก่อนนำมาใช้งาน โดยพิจารณาถึงความถูกต้อง และการตรวจสอบช่องโหว่อย่างถี่ถ้วน

๕.๔.๓) หากพบเหตุการณ์ละเมิดความมั่นคงปลอดภัยที่เกิดจากการประยุกต์ใช้เทคโนโลยี Generative AI ผู้ใช้งานต้องแจ้งให้บังคับบัญชาตามลำดับชั้นทราบ และต้องปฏิบัติตามขั้นตอนการปฏิบัติงาน (Work Procedure) เกี่ยวกับการบริหารจัดการเหตุการณ์ความมั่นคงปลอดภัยโดยทันที

๕.๔.๔) การใช้งานเทคโนโลยี Generative AI ผู้ใช้งานต้องดำเนินการให้มีความมั่นคงปลอดภัยไซเบอร์ ความมั่นคงปลอดภัยของระบบสารสนเทศ ความมั่นคงปลอดภัยของข้อมูล และความมั่นคงปลอดภัยในด้านอื่น ๆ ที่เกี่ยวข้อง

๕.๔.๕) สำนักงานจะมีการตรวจสอบการใช้เทคโนโลยีสารสนเทศ Generative AI ของผู้ใช้งานอย่างต่อเนื่องเพื่อเป็นการรักษาความมั่นคงปลอดภัยของสำนักงาน

๕.๕) การลดหรือหลีกเลี่ยงการเกิดอคติ (Bias) และการเลือกปฏิบัติ (Discrimination) ต่อบุคคลหรือกลุ่มบุคคล

ด้วยผลลัพธ์จากการประยุกต์ใช้เทคโนโลยี Generative AI มีโอกาสที่จะสร้างเนื้อหาที่มีความคิดเชิงลบ อคติ หรือแบ่งแยก อันอาจถูกเผยแพร่เป็นวงกว้างและไม่สามารถควบคุมการแพร่กระจายได้โดยง่าย อาจก่อให้เกิดความเสียหายแก่ชื่อเสียง จิตใจของบุคคล เกิดอคติ หรือการเลือกปฏิบัติต่อบุคคลหรือกลุ่มบุคคล ดังนั้น ผู้ใช้งานจึงต้องมีความระมัดระวังและปฏิบัติ ดังนี้

๕.๕.๑) ต้องตรวจสอบเนื้อหาที่สร้างโดยเทคโนโลยี Generative AI ก่อนนำไปใช้งานหรือเผยแพร่ต่อสาธารณะ เพื่อลดหรือหลีกเลี่ยงการเกิดอคติหรือการเลือกปฏิบัติอย่างไม่เป็นธรรม

๕.๕.๒) ระมัดระวังการใช้งานเทคโนโลยี Generative AI ในการดำเนินงานที่อาจกระทบต่อสิทธิของบุคคลหรือกลุ่มบุคคล หรือมีผลกระทบต่อหลักความเสมอภาค และมาตรฐานด้านสิทธิมนุษยชน

๕.๖) หน้าที่และความรับผิดชอบ

การนำเนื้อหาที่สร้างโดยเทคโนโลยี Generative AI มาใช้งาน อาจส่งผลกระทบเชิงลบต่อบุคคล หน่วยงาน สังคม และประเทศชาติ ทั้งโดยเจตนาหรือไม่เจตนา ซึ่งสำนักงานและผู้ใช้งานไม่อาจปฏิเสธความรับผิดชอบต่อผลที่เกิดขึ้นจากการกระทำดังกล่าวได้ ดังนั้น ผู้ใช้งานและบุคคลที่เกี่ยวข้องกับการประยุกต์ใช้เทคโนโลยี Generative AI จึงมีหน้าที่และความรับผิดชอบ ดังนี้

๕.๖.๑) ผู้ใช้งานต้องแจ้งผู้บังคับบัญชาเกี่ยวกับวัตถุประสงค์ ขอบเขต และลักษณะของการทำงานร่วมกันระหว่างผู้ใช้งานกับเทคโนโลยี Generative AI (AI-Human Involvement) และเมื่อมีการใช้เทคโนโลยี Generative AI ในการปฏิบัติงาน

๕.๖.๒) ผู้ใช้งานต้องปฏิบัติตามหลักเกณฑ์และวิธีการใช้งานเทคโนโลยี Generative AI ตามที่กำหนดในนโยบายนี้ และตรวจสอบเนื้อหาที่สร้างโดยเทคโนโลยี Generative AI ก่อนนำไปใช้งานหรือเผยแพร่ และต้องมีการตรวจสอบอย่างเคร่งครัด โดยเฉพาะเนื้อหาที่เกี่ยวข้องกับกฎหมาย ข้อมูลด้านการเงิน ข้อมูลอ่อนไหว ซึ่งผู้ใช้งานจำเป็นต้องพิจารณาในประเด็นดังต่อไปนี้ ประกอบด้วย

- ความถูกต้องของเนื้อหา
- ผลการใช้งานหรือเผยแพร่เนื้อหาที่นำไปสู่การกระทำผิดทางกฎหมายรวมถึงการละเมิดลิขสิทธิ์ เครื่องหมายการค้า หรือทรัพย์สินทางปัญญาอื่น ๆ
- ความเท่าเทียมและการไม่เลือกปฏิบัติต่อบุคคลหรือกลุ่มบุคคล
- การรักษาความลับและการคุ้มครองข้อมูลส่วนบุคคล
- ผลกระทบต่อความมั่นคงปลอดภัยและความเสี่ยงที่อาจเกิดขึ้นจากการใช้งาน
- ผลกระทบเชิงลบอื่น ๆ ที่อาจเกิดขึ้นต่อบุคคล สำนักงาน สังคม และประเทศชาติ

๕.๖.๓) ผู้ใช้งานต้องรายงานให้ผู้บังคับบัญชารับทราบโดยทันที ในกรณีที่การประยุกต์ใช้ Generative AI เกิดความผิดพลาดหรือพบประเด็นปัญหา ทั้งกรณีการนำเข้าข้อมูลและแสดงผลลัพธ์ที่อาจส่งผลกระทบเชิงลบต่อบุคคล สำนักงาน สังคม และประเทศชาติ เพื่อให้สามารถดำเนินมาตรการแก้ไขได้โดยเร็ว

๕.๖.๔) ผู้บังคับบัญชาและผู้ใช้งานต้องมีการติดตาม ทบทวน และประเมินประสิทธิภาพ ประสิทธิผลของการใช้งานเทคโนโลยี Generative AI อย่างต่อเนื่อง เพื่อปรับปรุงวิธีการทำงานและการเลือกใช้เทคโนโลยี Generative AI ที่เหมาะสมกับการปฏิบัติงาน

๕.๖.๕) การนำเนื้อหาที่สร้างจากเทคโนโลยี Generative AI มาใช้งาน ต้องมีการระบุว่า “เนื้อหานี้ได้รับความช่วยเหลือจากเทคโนโลยี Generative AI” ตามความเหมาะสม เพื่อให้เกิดความโปร่งใส และป้องกันความเข้าใจผิดของผู้รับข้อมูล

๕.๖.๖) หน่วยงานที่รับผิดชอบทางด้านเทคโนโลยีมีหน้าที่ในการพิจารณาตรวจสอบรายการแอปพลิเคชันหรือบริการ Generative AI ที่ใช้งานภายในสำนักงานตามความเหมาะสมของระดับการใช้งาน และนำเสนอผู้บริหารระดับสูงของหน่วยงาน เช่น CEO CIO CISO

๕.๗) การเคารพสิทธิ์ในทรัพย์สินทางปัญญา

๕.๗.๑) การนำเนื้อหาที่สร้างโดยเทคโนโลยี Generative AI ไปใช้งานหรือเผยแพร่ต่อสาธารณะ ผู้ใช้งานต้องใช้แอปพลิเคชันหรือบริการของเทคโนโลยี Generative AI ที่สำนักงานประกาศกำหนดเท่านั้น

๕.๗.๒) การประยุกต์ใช้เทคโนโลยี Generative AI ผู้ใช้งานต้องมีความระมัดระวังไม่ให้เกิดการละเมิดลิขสิทธิ์ เครื่องหมายการค้า หรือสิทธิ์ในทรัพย์สินทางปัญญาอื่น ๆ โดยการตรวจสอบให้แน่ใจว่าเนื้อหาที่สร้างขึ้นไม่เป็นการทำซ้ำ คัดลอก ดัดแปลง หรือใช้ประโยชน์จากผลงานที่มีเจ้าของโดยไม่ได้รับอนุญาต เพื่อป้องกันการละเมิดกฎหมายและข้อพิพาทที่อาจเกิดขึ้น

๖) การจัดฝึกอบรมให้ความรู้เกี่ยวกับการประยุกต์ใช้เทคโนโลยี Generative AI

สำนักงานจะจัดให้มีการอบรมพนักงาน ลูกจ้าง และผู้ปฏิบัติงานของสำนักงานเกี่ยวกับการประยุกต์ใช้เทคโนโลยี Generative AI รวมถึงแนวปฏิบัติที่ถูกต้อง ความเสี่ยง ข้อจำกัดของเทคโนโลยีและผลกระทบที่อาจเกิดขึ้น อย่างน้อยปีละ ๑ ครั้ง เป็นประจำทุกปีงบประมาณ (หน่วยงานสามารถปรับจำนวนครั้ง/ความถี่ของการจัดอบรมได้ตามความเหมาะสม)

๗) โทษของการไม่ปฏิบัติตามนโยบายการใช้เทคโนโลยี Generative AI

กรณีสำหรับพนักงาน ลูกจ้าง ผู้ปฏิบัติ หรือผู้รับจ้างของสำนักงานผู้ใดไม่ปฏิบัติตามนโยบายฉบับนี้ อาจมีผลเป็นความผิดและถูกดำเนินการทางวินัยตามข้อบังคับ ระเบียบ หรือประกาศของสำนักงาน หรือตามข้อตกลงในสัญญาจ้าง ทั้งนี้ ตามแต่กรณีและความสัมพันธ์ที่ผู้ใช้งานมีต่อสำนักงาน และอาจได้รับโทษตามที่กำหนดในกฎหมายลำดับรอง กฎ ระเบียบ หรือคำสั่งที่เกี่ยวข้อง

กรณีผู้ใช้งานพบการใช้งาน Generative AI ที่ไม่ถูกต้อง ไม่เหมาะสมหรือมีความเสี่ยงอันอาจจะก่อให้เกิดความเสียหายใด ๆ ต้องรายงานต่อผู้บังคับบัญชา หรือเจ้าหน้าที่ที่รับผิดชอบทราบโดยทันที

๘) ตัวอย่างพฤติกรรมที่ควรทำและไม่ควรทำของการใช้งานเทคโนโลยี Generative AI

(หน่วยงานสามารถเพิ่มเติมพฤติกรรมที่ควรทำและไม่ควรทำได้ตามความเหมาะสมเพื่อให้สอดคล้องกับภารกิจของหน่วยงาน)

พฤติกรรมที่ควรทำ	พฤติกรรมที่ไม่ควรทำ
✓ ตรวจสอบความถูกต้องของข้อมูลก่อนนำไปใช้ ✓ ไม่นำข้อมูลที่เป็นความลับของสำนักงานไปใช้งานร่วมกับเทคโนโลยี Generative AI ✓ เผยแพร่ข้อมูลที่ถูกต้องและไม่บิดเบือน ✓ อ้างอิงแหล่งที่มาของข้อมูลเสมอ ✓ หลีกเลี่ยงการสร้างเนื้อหาที่มีอคติหรือเลือกปฏิบัติ ✓ ใช้เครื่องมือที่ได้รับอนุญาตจากสำนักงาน ✓ หลีกเลี่ยงการใช้เนื้อหาที่อาจละเมิดลิขสิทธิ์หรือทรัพย์สินทางปัญญา ✓ ปฏิบัติตามนโยบายและแนวทางของสำนักงาน ✓ หลีกเลี่ยงการใช้งานเทคโนโลยี Generative AI ในทางที่ผิดกฎหมายหรือขัดต่อหลักธรรมาภิบาล คุณธรรมและจริยธรรม	✗ นำข้อมูลไปใช้ทันทีโดยไม่ได้มีการตรวจสอบความถูกต้อง ✗ นำข้อมูลที่เป็นความลับของสำนักงานไปใช้งานร่วมกับเทคโนโลยี Generative AI ✗ เผยแพร่ข้อมูลที่ไม่ถูกต้อง (Misinformation) หรือบิดเบือน (Disinformation) ✗ ไม่มีการอ้างอิงแหล่งที่มาของข้อมูล ✗ สร้างเนื้อหาที่มีอคติหรือเลือกปฏิบัติ ✗ ใช้เครื่องมือที่ไม่ได้รับอนุญาตจากสำนักงาน ✗ ใช้เนื้อหาที่อาจละเมิดลิขสิทธิ์หรือทรัพย์สินทางปัญญา ✗ ไม่ปฏิบัติตามนโยบายและแนวทางของสำนักงาน ✗ ใช้งานเทคโนโลยี Generative AI ในทางที่ผิดกฎหมายหรือขัดต่อหลักธรรมาภิบาล คุณธรรม และจริยธรรม

๙) แอปพลิเคชันหรือบริการ Generative AI

ประเภทแอปพลิเคชันหรือบริการ	ตัวอย่างแอปพลิเคชันหรือบริการ
เครื่องมือสร้างข้อความ	
เครื่องมือสร้างภาพ	
เครื่องมือสร้างโคด	
เครื่องมือสร้างเสียงและวิดีโอ	

ประกาศ ณ วันที่ เดือน..... พ.ศ. ....

ลงนาม.....

(...ระบุชื่อ-สกุลของผู้บริหารสูงสุดของหน่วยงาน..)

ตำแหน่ง .....

ผนวก ซ ตัวอย่าง นโยบายการใช้เทคโนโลยีระบบปัญญาประดิษฐ์ที่ยอมรับได้ (Acceptable Use Policy)

(ตัวอย่าง)

ตราสัญลักษณ์หน่วยงาน

หน่วยงาน.....(โปรดระบุ).....

---

๑) บทนำ

เทคโนโลยีระบบปัญญาประดิษฐ์ ได้เข้ามามีบทบาทสำคัญในการเพิ่มประสิทธิภาพการปฏิบัติงานขององค์กร ตั้งแต่การวิเคราะห์ข้อมูลเชิงลึก การคาดการณ์แนวโน้ม การช่วยตัดสินใจในกระบวนการทางธุรกิจ ไปจนถึงการสร้างสรรค์เนื้อหาใหม่ ๆ ผ่านระบบปัญญาประดิษฐ์ (AI System)อย่างไรก็ตาม การนำระบบปัญญาประดิษฐ์มาใช้งานย่อมมาพร้อมกับความเสี่ยงและข้อควรพิจารณาที่กว้างขวาง ไม่ว่าจะเป็นความเสี่ยงจากการตัดสินใจที่มีอคติ ความผิดพลาดของแบบจำลอง การขาดความโปร่งใสในการให้เหตุผล ความมั่นคงปลอดภัยของข้อมูลและโมเดล ใช้ระบบปัญญาประดิษฐ์ในทางที่ผิดและการละเมิดข้อมูลส่วนบุคคล

ดังนั้น เพื่อให้การประยุกต์ใช้ระบบปัญญาประดิษฐ์เป็นไปอย่างมีความรับผิดชอบและมั่นคงปลอดภัย จึงสมควรกำหนดนโยบายการใช้เทคโนโลยีระบบปัญญาประดิษฐ์ที่ยอมรับได้ขึ้น

๒) วัตถุประสงค์

นโยบายนี้จัดทำขึ้นเพื่อให้ผู้ใช้งานสามารถนำเทคโนโลยีระบบปัญญาประดิษฐ์ไปใช้งานได้อย่างถูกต้องเหมาะสม มีความรับผิดชอบ และมั่นคงปลอดภัย โดยมีวัตถุประสงค์ ดังนี้

๒.๑) เพื่อให้มีแนวปฏิบัติในการใช้เทคโนโลยีระบบปัญญาประดิษฐ์อย่างถูกต้อง เหมาะสม และมีประสิทธิภาพ

๒.๒) เพื่อให้การใช้เทคโนโลยีระบบปัญญาประดิษฐ์สอดคล้องกับกฎหมาย ข้อบังคับ และหลักจริยธรรมที่เกี่ยวข้อง พร้อมทั้งป้องกันความเสี่ยงที่อาจเกิดขึ้นกับบุคคล องค์กร และสังคม

๒.๓) เพื่อเสริมสร้างความตระหนักรู้และความเข้าใจเกี่ยวกับการใช้เทคโนโลยีระบบปัญญาประดิษฐ์อย่างมีความรับผิดชอบและรอบคอบ

๓) ขอบเขต

นโยบายฉบับนี้ครอบคลุมการนำเทคโนโลยีระบบปัญญาประดิษฐ์ทุกประเภทที่องค์กรจัดหาหรืออนุญาตให้ใช้มาสนับสนุนการดำเนินงานตามภารกิจขององค์กร สำหรับพนักงาน ลูกจ้าง ผู้ปฏิบัติงาน และผู้รับจ้างขององค์กร ซึ่งต่อไปนี้เรียกว่า “ผู้ใช้งาน”

๔) คำนิยาม

๔.๑) “ระบบปัญญาประดิษฐ์ (AI System)” หมายถึง ระบบที่ทำงานบนเครื่องจักรซึ่งถูกออกแบบมาให้ทำงานด้วยระดับความเป็นอิสระที่แตกต่างกัน และทำการอนุมานจากข้อมูลที่ได้รับ เพื่อสร้างผลลัพธ์ เช่น การคาดการณ์ เนื้อหา คำแนะนำ หรือการตัดสินใจ ที่สามารถมีอิทธิพลต่อสภาพแวดล้อมได้ ซึ่งรวมถึงแต่ไม่จำกัดเพียง Generative AI, Predictive Analytics และ Computer Vision

๔.๒) “Generative AI” หมายถึง เทคโนโลยี AI ประเภทหนึ่งที่มีความสามารถในการสร้างเนื้อหาใหม่ เช่น ข้อความ รูปภาพ หรือซอร์สโคด ตามคำสั่ง (Prompt) ที่มนุษย์กำหนด

๔.๓) “ข้อมูลส่วนบุคคล” หมายถึง ข้อมูลเกี่ยวกับบุคคลธรรมดาซึ่งทำให้สามารถระบุตัวบุคคลนั้นได้ ไม่ว่าทางตรงหรือทางอ้อม แต่ไม่รวมถึงข้อมูลของผู้ถึงแก่กรรมโดยเฉพาะ

๔.๔) “หน่วยงาน” หมายถึง [โปรดระบุชื่อหน่วยงาน]

- ***(คำนิยามอื่นๆ สามารถระบุเพิ่มเติมได้ตามความเหมาะสม)***

๕) แนวทางการใช้เทคโนโลยีระบบปัญญาประดิษฐ์ที่ยอมรับได้

๕.๑) การใช้งานที่ยอมรับได้ ผู้ใช้งานต้องประยุกต์ใช้เทคโนโลยีระบบปัญญาประดิษฐ์เพื่อช่วยสนับสนุนการดำเนินงานในบริบทที่กำหนด ดังต่อไปนี้

๕.๑.๑) การวิเคราะห์ข้อมูล การคาดการณ์แนวโน้ม และการสร้างรายงาน

๕.๑.๒) การเขียนและทวนสอบชุดคำสั่งหรือโปรแกรม

๕.๑.๓) การจัดทำร่างหนังสือหรือเอกสารต่าง ๆ

๕.๑.๔) การให้คำแนะนำหรือเป็นผู้ช่วยในการแก้ปัญหาเบื้องต้น

๕.๑.๕) การสนับสนุนงานวิจัยและวิชาการ เช่น การช่วยทบทวนวรรณกรรม

๕.๑.๖) การสร้างเนื้อหาเพื่อการสื่อสารและการประชาสัมพันธ์

๕.๑.๗) ดำเนินงานอื่น ๆ ที่ได้รับมอบหมายจากผู้บังคับบัญชา

๕.๒) ข้อห้ามในการใช้งาน

๕.๒.๑) ห้ามใช้ปัญญาประดิษฐ์ในการตัดสินใจที่สำคัญ (High-Stakes Decisions) เช่น การอนุมัติสินเชื่อ การวินิจฉัยทางการแพทย์ หรือการพิจารณาด้านวินัย โดยปราศจากการตรวจสอบและอนุมัติโดยมนุษย์

๕.๒.๒) ห้ามป้อนข้อมูลที่ไม่ได้รับการตรวจสอบความถูกต้อง หรือข้อมูลที่อาจบิดเบือนการทำงานของโมเดล (Data Poisoning) เข้าสู่ระบบปัญญาประดิษฐ์

๕.๒.๓) ห้ามใช้ระบบปัญญาประดิษฐ์ในทางที่ผิด (Misuse) หรือจงใจใช้เป็นเครื่องมือเพื่อสร้างการโจมตีทางไซเบอร์ โดยเด็ดขาด ซึ่งรวมถึงแต่ไม่จำกัดเพียงการใช้ AI เพื่อค้นหาช่องโหว่ในระบบอื่น

การสร้างโคดสำหรับมัลแวร์ การสร้างสื่อสังเคราะห์ปลอม (Deepfake) เพื่อใช้ในการหลอกลวง หรือการเขียนข้อความฟิชชิ่ง (Phishing) ที่มีความแนบเนียนสูง

๕.๒.๔) ห้ามป้อนข้อมูลที่เป็นความลับขององค์กร ข้อมูลส่วนบุคคล หรือข้อมูลอ่อนไหวเข้าสู่ระบบปัญญาประดิษฐ์ สาธารณะ หรือระบบที่องค์กรไม่ได้รับรองความปลอดภัย

๕.๒.๕) ห้ามใช้เพื่อสร้างเนื้อหาที่ผิดกฎหมาย ละเมิดศีลธรรม หรืออาจสร้างความเสียหายต่อบุคคล องค์กรและสังคม

๕.๒.๖) ห้ามใช้เพื่อสร้างเนื้อหาที่ละเมิดลิขสิทธิ์หรือทรัพย์สินทางปัญญาของผู้อื่น

๕.๓) การรักษาความลับและคุ้มครองข้อมูลส่วนบุคคล

๕.๓.๑) ห้ามนำข้อมูลภายในองค์กรและข้อมูลที่มีชั้นความลับ เช่น รหัสผ่าน เอกสารสัญญา ข้อมูลโครงการไปใช้งานร่วมกับเทคโนโลยีระบบปัญญาประดิษฐ์ที่ไม่ได้รับอนุญาต

๕.๓.๒) กรณีที่จำเป็นต้องใช้ระบบปัญญาประดิษฐ์ร่วมกับข้อมูลภายในหรือข้อมูลส่วนบุคคล จะต้องใช้ระบบที่หน่วยงานจัดหาหรือรับรองเท่านั้น และต้องได้รับอนุมัติจาก [*ระบุผู้มีอำนาจอนุมัติ*] ก่อนทุกครั้ง

๕.๔) การรักษาความมั่นคงปลอดภัยของระบบสารสนเทศ

๕.๔.๑) ห้ามนำข้อมูลที่อาจส่งผลกระทบต่อความมั่นคงปลอดภัยของระบบ เช่น รหัสผ่าน API Key ไปใช้งานร่วมกับระบบปัญญาประดิษฐ์

๕.๔.๒) ต้องตรวจสอบซอร์สโคดหรือชุดคำสั่งที่สร้างโดยปัญญาประดิษฐ์ ก่อนนำมาใช้งานจริงเสมอ เพื่อหาช่องโหว่และข้อผิดพลาด

๕.๔.๓) หากพบพฤติกรรมของปัญญาประดิษฐ์ ที่ผิดปกติ น่าสงสัย หรืออาจเป็นช่องโหว่ด้านความปลอดภัย ต้องแจ้งผู้บังคับบัญชาหรือฝ่ายที่รับผิดชอบทันที

๕.๕) การตรวจสอบผลลัพธ์และลดอคติ

๕.๕.๑) ผู้ใช้งานต้องตรวจสอบเนื้อหาและผลลัพธ์ที่สร้างหรือวิเคราะห์โดยปัญญาประดิษฐ์ก่อนนำไปใช้งานหรือเผยแพร่เสมอ เพื่อลดหรือหลีกเลี่ยงการเกิดอคติหรือการเลือกปฏิบัติอย่างไม่เป็นธรรม

๕.๕.๒) ต้องใช้ความระมัดระวังเป็นพิเศษในการใช้ปัญญาประดิษฐ์ที่เกี่ยวข้องกับการประเมินหรือตัดสินใจเกี่ยวกับบุคคลเพื่อไม่ให้เกิดผลกระทบต่อสิทธิและความเสมอภาค

๖) หน้าที่และความรับผิดชอบ

๖.๑) ผู้ใช้งานมีหน้าที่และความรับผิดชอบต่อการตัดสินใจขั้นสุดท้ายที่เกิดขึ้นจากการใช้ข้อมูล คำแนะนำหรือผลลัพธ์จากระบบปัญญาประดิษฐ์

๖.๒) ผู้ใช้งานต้องตรวจสอบผลลัพธ์จากปัญญาประดิษฐ์ก่อนนำไปใช้งานหรือเผยแพร่เสมอ โดยต้องพิจารณาถึงความถูกต้อง ผลกระทบทางกฎหมาย ความเท่าเทียม และความมั่นคงปลอดภัย

๖.๓) ผู้ใช้งานต้องรายงานให้ผู้บังคับบัญชาทราบทันที หากพบว่าระบบปัญญาประดิษฐ์ทำงานผิดพลาด ให้ผลลัพธ์ที่มีอคติ หรืออาจส่งผลกระทบเชิงลบ

๖.๔) การนำเนื้อหาที่ได้รับความช่วยเหลือจากปัญญาประดิษฐ์มาใช้งานต้องมีการระบุที่มาตามความเหมาะสมเพื่อความโปร่งใส

๗) การจัดฝึกอบรมให้ความรู้

หน่วยงานจะจัดให้มีการอบรมเกี่ยวกับการใช้งานระบบปัญญาประดิษฐ์ ความเสี่ยง ข้อจำกัด และผลกระทบที่อาจเกิดขึ้น อย่างน้อยปีละ ๑ ครั้ง

๘) โทษของการไม่ปฏิบัติตามนโยบาย

การไม่ปฏิบัติตามนโยบายฉบับนี้ อาจมีผลเป็นความผิดและถูกดำเนินการทางวินัยตามข้อบังคับของหน่วยงาน หรือตามกฎหมายที่เกี่ยวข้อง

๙) ตัวอย่างพฤติกรรมที่ควรทำและไม่ควรทำ

พฤติกรรม ที่ควรทำ	พฤติกรรม ที่ไม่ควรทำ
✓ **ตรวจสอบความถูกต้องและความสมเหตุสมผล** ของผลลัพธ์จากปัญญาประดิษฐ์ทุกครั้ง	✗ **เชื่อผลการวิเคราะห์หรือการตัดสินใจ** ของระบบปัญญาประดิษฐ์ทันทีโดยไม่ใช้วิจารณญาณ
✓ **ใช้ปัญญาประดิษฐ์เป็นเครื่องมือสนับสนุน** โดยมนุษย์เป็นผู้ตัดสินใจขั้นสุดท้าย	✗ **ใช้ระบบปัญญาประดิษฐ์ในการตัดสินใจที่สำคัญ**โดยไม่มีการกำกับดูแลจากมนุษย์
✓ **ใช้ข้อมูลที่สะอาดและได้รับอนุญาต**ป้อนเข้าสู่ระบบปัญญาประดิษฐ์	✗ **ป้อนข้อมูลที่เป็นความลับ ข้อมูลส่วนบุคคลหรือข้อมูลที่ไม่น่าเชื่อถือ**เข้าระบบ
✓ **อ้างอิงและให้เครดิต**ว่ามีการใช้ปัญญาประดิษฐ์ช่วยในการทำงานตามความเหมาะสม	✗ **นำผลงานจากระบบปัญญาประดิษฐ์ไปใช้โดยอ้างว่าเป็นผลงานของตนเอง**ทั้งหมด
✓ **ปฏิบัติตามนโยบาย**และแนวทางขององค์กรอย่างเคร่งครัด	✗ **ใช้งานระบบปัญญาประดิษฐ์ในทางที่ผิดกฎหมาย** หรือขัดต่อหลักธรรมาภิบาล

ประกาศ ณ วันที่ เดือน..... พ.ศ. ....

ลงนาม.....

(...ระบุชื่อ-สกุลของผู้บริหารสูงสุดของหน่วยงาน..)

ตำแหน่ง .....