Building Responsible AI: Frameworks for Accountability, Audits, and Autonomy

โดย Prof. Karrie Karahalios
Professor of Media Arts and Sciences, the MIT Media Lab
วันพุธที่ 19 พ.ย. 68 ณ ห้องโกมุท ชั้น 29 สำนักงานใหญ่ สีลม เวลา 11:00 – 12:00 น.

แม้ปัญญาประดิษฐ์ (AI) จะช่วยขับเคลื่อนประสิทธิภาพและนวัตกรรมใหม่ ๆ ในหลายภาคส่วน แต่ความท้าทายด้านการกำกับดูแลกลับเพิ่มสูงขึ้นตามไปด้วย เราพบตัวอย่างจำนวนมากที่อัลกอริทึมนำไปสู่อคติ ความเหลื่อมล้ำ และการตัดสินใจที่ไม่เป็นธรรม โดยเกิดขึ้นอย่างแนบเนียนจนผู้ใช้ไม่ทันตระหนัก การพัฒนา Responsible AI จึงเป็นสิ่งจำเป็น เพื่อให้ระบบทำงานอย่างปลอดภัย โปร่งใส และสอดคล้องกับค่านิยมของสังคมในวงกว้าง

ด้วยเหตุนี้ การบรรยายหัวข้อ “Building Responsible AI: Frameworks for Accountability, Audits, and Autonomy” จึงมีวัตถุประสงค์เพื่อเสนอกรอบแนวคิดและเครื่องมือสำหรับการพัฒนาและกำกับดูแล AI อย่างรอบด้าน ตั้งแต่การตรวจสอบอัลกอริทึม การออกแบบกลไกความรับผิดชอบ ไปจนถึงการสร้างระบบที่เปิดโอกาสให้ผู้ใช้มีอำนาจควบคุมและมีส่วนร่วมเหนือเทคโนโลยีได้อย่างแท้จริง

โลกที่ AI กลายเป็นโครงสร้างพื้นฐานของสังคม

AI ไม่ได้ทำงานเพียงในฐานะเทคโนโลยีที่อยู่เบื้องหลังระบบดิจิทัล แต่ได้กลายเป็นโครงสร้างพื้นฐานที่กำหนดรูปแบบการทำงานของเศรษฐกิจ หน่วยงานรัฐ และวิถีชีวิตของประชาชน ตั้งแต่แพลตฟอร์มโซเชียลมีเดีย ระบบธนาคาร การแพทย์ ไปจนถึงสถาบันการศึกษา AI มีอิทธิพลต่อสิ่งที่ผู้คนมองเห็น การตัดสินใจในชีวิตประจำวัน และโอกาสที่สามารถเข้าถึงได้ในระบบเศรษฐกิจและสังคม

ลักษณะการแพร่กระจายของ AI เช่นนี้ทำให้งานพัฒนาระบบจำเป็นต้องตั้งอยู่บนกรอบ “Socio-technical Systems” ซึ่งมองเทคโนโลยีควบคู่ไปกับบริบททางสังคม ได้แก่ พฤติกรรมผู้ใช้ วัฒนธรรม กฎหมาย กลไกตลาด และโครงสร้างอำนาจในสังคม เนื่องจากอัลกอริทึม

เพียงตัวเดียวสามารถเปลี่ยนทิศทางของการสื่อสารสาธารณะ ส่งผลต่อโอกาสของบุคคล หรือก่อให้เกิดความเหลื่อมล้ำเชิงโครงสร้างได้โดยที่ผู้ได้รับผลกระทบอาจไม่ทันรับรู้

ดังนั้น การสร้าง AI ที่รับผิดชอบจำเป็นต้องอาศัยการทำงานข้ามสาขาวิชา นักมานุษยวิทยา นักสังคมวิทยา นักกฎหมาย นักเศรษฐศาสตร์ วิศวกรคอมพิวเตอร์ ตลอดจนนักการเมือง ล้วนมีบทบาทสำคัญ เพราะเทคโนโลยีทั้งหมดอยู่ภายใต้กฎเกณฑ์ ค่านิยม และพฤติกรรมมนุษย์ ไม่ใช่ผลผลิตที่ดำรงอยู่ในห้องปฏิบัติการอย่างโดดเดี่ยว

ศักยภาพและผลกระทบของ AI: ดาบสองคมที่สังคมต้องรับมือ

AI มีศักยภาพในการยกระดับคุณภาพชีวิตอย่างลึกซึ้ง โดยเฉพาะในงานที่ซับซ้อนหรือมีข้อจำกัดด้านเวลาและความแม่นยำของมนุษย์ ตัวอย่างสำคัญคือการแพทย์แบบเฉพาะบุคคล (personalized medicine) ซึ่งใช้ AI วิเคราะห์ภาพถ่ายเพื่อค้นหาความผิดปกติของโรคได้เร็วกว่าการประเมินของแพทย์ เช่น การตรวจมะเร็งเต้านม หรือโรคผิวหนังที่ต้องดูจากสัญญาณเฉพาะจุด ช่วยเพิ่มอัตราการวินิจฉัยที่ถูกต้องและการรักษาอย่างทันท่วงที

ในภาคเกษตรกรรม AI ถูกนำมาใช้เพื่อเพิ่มผลผลิตผ่านการวิเคราะห์ภาพถ่ายดาวเทียม เซนเซอร์ตรวจวัดความชื้นของดิน และระบบคาดการณ์ความต้องการปุ๋ย–ปริมาณน้ำที่เหมาะสม แม้ประสิทธิภาพที่เพิ่มขึ้นจะช่วยลดต้นทุนและเพิ่มรายได้ให้เกษตรกร แต่ก็ตั้งคำถามสำคัญด้านสิ่งแวดล้อมและจริยธรรม เช่น การใช้สารเคมี

เกินความจำเป็น การปล่อยคาร์บอนที่สูงขึ้นจากระบบอุตสาหกรรมเกษตร และความเสี่ยงที่ระบบ AI จะผลักให้เกษตรกรมุ่งสู่การผลิตสูงสุดโดยละเลยผลกระทบต่อคุณภาพดิน น้ำ และสุขภาพแรงงานในระยะยาว

ในตลาดแรงงาน ผลกระทบของ AI ปรากฏเด่นชัดที่สุดในกลุ่มแรงงานรายได้ปานกลาง–ต่ำ โดยเฉพาะงานที่มีรูปแบบซ้ำหรือพึ่งพาการประมวลผลข้อมูลจำนวนมาก เช่น งานเอกสาร การบริการลูกค้า หรือสายการผลิต งานวิจัยหลายชิ้นระบุว่า AI เป็นตัวเร่งสำคัญของนวัตกรรมใหม่ในสินค้าและบริการ เช่น ระบบวิเคราะห์ข้อมูลอัตโนมัติ หรือบริการแปลภาษาแบบเรียลไทม์ ซึ่งสร้างโอกาสทางเศรษฐกิจรูปแบบใหม่ อย่างไรก็ดี ผลประโยชน์จากอุตสาหกรรมเหล่านี้ไม่ได้กระจายไปยังแรงงานทุกกลุ่มอย่างเท่าเทียม แรงงานทักษะสูงมีแนวโน้มได้รับรายได้สูงขึ้น ในขณะที่แรงงานทักษะทั่วไปมีความเสี่ยงถูกแทนที่ด้วยระบบอัตโนมัติ หากไม่มีนโยบาย reskilling และมาตรการรองรับการเปลี่ยนผ่านโครงสร้างรายได้ในสังคมอาจยิ่งเหลื่อมล้ำมากขึ้น

ผลกระทบของ AI ยังปรากฏชัดในระดับทักษะมนุษย์ งานวิจัยหลายชิ้นชี้ว่า แม้ AI จะช่วยลดภาระงานและเพิ่มความสะดวก แต่กลับทำให้ความสามารถพื้นฐานบางด้านลดลงอย่างค่อยเป็นค่อยไป เช่น การใช้ระบบนำทาง (GPS) ต่อเนื่องทำให้สมองส่วนที่ใช้ในการจดจำเส้นทางและวางแผนการเดินทางทำงานน้อยลง ขณะที่ผู้เรียนที่ใช้โมเดลภาษา (LLM) เพื่อช่วยเขียนงานมัก “จำเนื้อหาของตนเองไม่ได้” และมีความรู้สึกเป็นเจ้าของผลงานต่ำเมื่อเทียบกับผู้ที่ลงมือเขียนเองหรือเพียงค้นคว้าข้อมูล นอกจากนี้ ระบบ AI ยังมีแนวโน้มเพิ่มจำนวนงานย่อยที่ต้องจัดการ เช่น งานตรวจทาน แก้ไข หรือส่งข้อมูลให้ระบบอัตโนมัติ ซึ่งนำไปสู่ภาวะ “Tyranny of Tiny Tasks” หรือการสะสมของงานเล็ก ๆ จำนวนมากจนกลายเป็นความเครียดเรื้อรังและลดคุณภาพชีวิตในระยะยาว

ปรากฏการณ์เหล่านี้สะท้อนว่า AI ไม่ได้เป็นเพียงเครื่องมือเสริมศักยภาพการทำงานเท่านั้น แต่ยังเปลี่ยนรูปแบบการคิด การเรียนรู้ และพฤติกรรมของมนุษย์อย่างลึกซึ้ง หากสังคมไม่มีมาตรการรองรับและแนวทางใช้เทคโนโลยีอย่างรอบคอบ ผลลัพธ์ที่เกิดขึ้นอาจลดทอนศักยภาพมนุษย์และเพิ่มความเหลื่อมล้ำทั้งในตลาดแรงงานและชีวิตประจำวัน

รูปแบบการใช้งาน AI และความเสี่ยงที่เกิดขึ้นจริงในชีวิตประจำวัน

การทำงานของ AI พึ่งพาฐานข้อมูลจำนวนมหาศาล ซึ่งต้องผ่านกระบวนการป้ายกำกับ (labeling) โดยแรงงานมนุษย์เป็นจำนวนมาก ขณะเดียวกันโมเดลภาษาขนาดใหญ่ (Large Language Model: LLM) ยังอาศัยข้อมูลจาก Reddit มากถึง 40% ซึ่งเป็นแหล่งข้อมูลที่มิได้ผ่านการตรวจสอบหรือคัดกรองอย่างเป็นระบบ ส่งผลให้มุมมองของโมเดลถูกกำหนดโดยวัฒนธรรมออนไลน์แบบอเมริกันและเสียงของผู้ใช้งานกลุ่มเฉพาะ มากกว่าที่จะสะท้อนความหลากหลายของสังคมโลกโดยรวม

ตัวอย่างหนึ่งที่สะท้อนปัญหานี้ได้ชัด คือกรณีที่โมเดล AI วิเคราะห์ภาพบ้านในอินเดียแล้วระบุว่ามี “ยาสีฟันยี่ห้ออเมริกัน” อยู่ในภาพ ทั้งที่ผลิตภัณฑ์ดังกล่าวไม่ได้อยู่ในบริบทนั้นจริง การตีความผิดเช่นนี้เกิดจากข้อมูลฝึกที่มี “อคติด้านวัฒนธรรม” ทำให้โมเดลมองโลกตามแบบแผนของข้อมูลที่พบมากที่สุด ไม่ใช่ตามความเป็นจริงของแต่ละสังคมซึ่งแสดงให้เห็นว่าการรับรู้ของ AI สามารถเอนเอียงได้อย่างง่ายดายตามโครงสร้างข้อมูลที่ใช้ฝึก ไม่ใช่ตามสภาพแวดล้อมโลกในความเป็นจริง

ความเสี่ยงจากการใช้งาน AI จึงเป็นสิ่งที่เกิดขึ้นจริงในหลากหลายสาขา และส่งผลโดยตรงต่อประชาชนในชีวิตประจำวัน เช่น

1) **การจับกุมผิดตัวจากระบบจดจำใบหน้า:** ชายคนหนึ่งถูกจับกุมจากข้อผิดพลาดของระบบ face recognition ซึ่งระบุผิดว่าเขาเป็นผู้ต้องสงสัย ทั้งที่ไม่ได้อยู่ในพื้นที่เกิดเหตุ ส่งผลกระทบต่อความน่าเชื่อถือของกระบวนการยุติธรรมอย่างรุนแรง
2) **อัลกอริทึมคัดกรองผู้ต้องขัง:** อัลกอริทึมคัดกรองผู้ต้องขัง ที่ประเมินว่าคนผิวสีมีความเสี่ยงก่อเหตุซ้ำสูงกว่าคนผิวขาว ทั้งที่ไม่มีข้อมูลเชิงประจักษ์สนับสนุน ผลลัพธ์ดังกล่าวถูกนำไปใช้ในการตัดสินโทษและสิทธิในการ

ประกันตัว ทำให้เกิดการเลือกปฏิบัติอย่างเป็นระบบโดยไม่ตั้งใจ

3) **ระบบตรวจการโกงข้อสอบช่วง COVID:** อัลกอริทึมกล่าวหานักเรียนที่มีภาวะสมาธิสั้น (Attention Deficit Hyperactivity Disorder: ADHD) ว่าโกง เพียงเพราะ “มองออกนอกจอ” อีกทั้งยังตรวจจับใบหน้านักเรียนผิวสีได้แย่ในสภาพแสงต่ำ ทำให้ผู้เรียนจำนวนมากถูกกล่าวหาโดยไม่เป็นธรรม
4) **Chatbot ในศูนย์ให้คำปรึกษาเด็กมีปัญหาการกิน:** องค์กรที่ให้การสนับสนุนผู้ป่วยที่มีความผิดปกติด้านการกินอาหาร (Eating Disorder) ได้นำระบบ chatbot มาใช้แทนเจ้าหน้าที่ในการให้คำปรึกษา แต่ระบบกลับให้คำแนะนำอันตราย เช่น การแนะนำให้ “ลดปริมาณอาหาร” ซึ่งเป็นสิ่งตรงข้ามกับหลักการรักษา ส่งผลให้องค์กรต้องถอนระบบออกทันที
5) **การใช้อัลกอริทึมเพื่อคำนวณผลสอบแทนการสอบจริง:** อังกฤษใช้อัลกอริทึมให้เกรดนักเรียนช่วง COVID-19 ผลปรากฏว่าคะแนนของนักเรียนถูกกำหนดโดย “ชื่อเสียงของโรงเรียน” มากกว่าความสามารถของนักเรียนจริง ทำให้นักเรียนจากโรงเรียนรัฐได้คะแนนต่ำกว่าที่ควร ขณะที่โรงเรียนเอกชนได้คะแนนสูงเกินควรจนเกิดการประท้วงทั่วประเทศภายในเวลาเพียงไม่กี่วัน
6) **การผลักภาระงานสู่ผู้ป่วยในระบบโรงพยาบาล:** AI ทำให้การทำงานบางขั้นตอนของบุคลากรลดลง แต่ในหลายกรณีกลับผลักภาระไปสู่ผู้ป่วย เช่น ต้องลงทะเบียนออนไลน์เอง กรอกข้อมูลจำนวนมาก หรือนัดหมายรักษาผ่านระบบอัตโนมัติ ซึ่งอาจเป็นอุปสรรคสำหรับกลุ่มผู้สูงอายุหรือผู้มีข้อจำกัดด้านดิจิทัล

ตัวอย่างทั้งหมดนี้สะท้อนให้เห็นว่า AI แม้ได้รับการออกแบบมาเพื่อเพิ่มประสิทธิภาพ แต่เมื่อถูกนำไปใช้งานในสภาพแวดล้อมจริงกลับสามารถสร้างความเสี่ยงทั้งในระดับบุคคลและระดับสังคมได้อย่างกว้างขวาง โดยเฉพาะเมื่อข้อมูลที่ใช้ฝึกโมเดลมีความเอนเอียง หรือเมื่อระบบขาดการตรวจสอบและกำกับดูแลที่รัดกุม

ธรรมาภิบาล AI และแนวทางอยู่ร่วมอย่างยั่งยืน

การกำกับดูแล AI จำเป็นต้องตั้งอยู่บนหลักความโปร่งใส ความรับผิดชอบ และการคุ้มครองศักดิ์ศรีความเป็นมนุษย์ เนื่องจากอัลกอริทึมเข้ามามีบทบาทสำคัญในการตัดสินใจที่ส่งผลกระทบต่อชีวิตผู้คน ตั้งแต่การคัดกรองข้อมูลบนแพลตฟอร์ม ไปจนถึงการอนุมัติสินเชื่อ การประเมินความเสี่ยง และการเข้าถึงบริการสาธารณะ การทำความเข้าใจว่าอัลกอริทึมทำงานอย่างไรและส่งผลต่อใครจึงเป็นหัวใจสำคัญของธรรมาภิบาล AI ระบบอาจมีอคติแฝงอยู่จากข้อมูลฝึกที่ไม่สมดุล การออกแบบที่ไม่ครอบคลุม หรือกลไกการคัดเลือกเนื้อหาที่ไม่โปร่งใส ขั้นตอนแรกของการกำกับดูแลจึงต้องมีเครื่องมือที่สามารถตรวจสอบและเปิดเผยพฤติกรรมของระบบได้ ซึ่งนำไปสู่แนวทางการตรวจสอบอัลกอริทึมในรูปแบบต่าง ๆ

1) **การตรวจสอบอัลกอริทึม (Algorithm Audits):** การตรวจสอบอัลกอริทึมจะช่วยเปิดเผยรูปแบบความเอนเอียงที่ซ่อนอยู่ในระบบ โดยในงานศึกษาได้ใช้วิธีการหลากหลาย เช่น sock puppets หรือบัญชีผู้ใช้จำลองที่กำหนดเพศ เชื้อชาติ อายุ และพฤติกรรม เพื่อดูว่าแพลตฟอร์มปฏิบัติต่อผู้ใช้แต่ละกลุ่มแตกต่างกันหรือไม่ ตลอดจน collective audits ที่รวบรวมข้อมูลจากผู้ใช้จำนวนมาก เพื่อหาความผิดปกติของระบบ และ micro-audits สำหรับสถาบันการเงินเพื่อตรวจสอบความเป็นธรรมในการอนุมัติสินเชื่อ ผลการตรวจสอบชี้ให้เห็นความเหลื่อมล้ำหลายรูปแบบ เช่น การแสดงบ้านราคาแพงให้ผู้หญิงบ่อยกว่าผู้ชาย เว็บไซต์โรงแรมบางแห่งเพิ่มคะแนนรีวิวสูงกว่าความจริงถึง 25% หรือโฆษณาสินเชื่อที่เอาเปรียบ (predatory loans) ถูกส่งให้กลุ่มรายได้น้อยในอัตราสูงกว่ากลุ่มรายได้สูง

   นอกจากนี้ หลักฐานเชิงประจักษ์ยังช่วยยืนยันความจำเป็นของการตรวจสอบ เช่น ใน ปี 2012 แพลตฟอร์มเสนอขาย–เช่าอสังหาริมทรัพย์ในสหรัฐพบว่าผู้ใช้เชื้อสายเอเชียได้รับการแสดงผลที่อยู่อาศัยน้อยกว่า

ผู้ใช้ผิวขาวถึง 18.8% และผู้ใช้ผิวสีถูกแสดงผลน้อยกว่าถึง 17.7% แสดงให้เห็นว่าระบบอาจมีบทบาทในการสร้างหรือขยายความเหลื่อมล้ำด้านการเข้าถึงที่อยู่อาศัยโดยไม่ตั้งใจ

2) **ความโปร่งใสของแพลตฟอร์มโซเชียลมีเดีย:** การศึกษาความโปร่งใสของแพลตฟอร์มโซเชียลมีเดีย พบว่าใน ปี 2017 เมื่อทดลองแสดงโพสต์บน Facebook แบบ “ครบทุกโพสต์” เทียบกับ “โพสต์ที่ระบบคัดเลือกให้ดู” ผู้ใช้มากกว่า 60% เพิ่งตระหนักว่าตนไม่ได้เห็นเนื้อหาทั้งหมดจากเพื่อนและครอบครัว และบางคนเห็นโพสต์จากบุคคลที่ไม่รู้จักในสัดส่วนสูงกว่าคนใกล้ชิด ปรากฏการณ์นี้ทำให้ความไว้วางใจต่อแพลตฟอร์มลดลง และชี้ให้เห็นว่าผู้ใช้ต้องการคำอธิบายเกี่ยวกับกลไกการคัดเลือกเนื้อหาเพื่อให้ผู้ใช้งานเข้าใจว่าเหตุใดจึงเห็นหรือไม่เห็นเนื้อหาบางประเภท ไม่เช่นนั้นแพลตฟอร์มอาจถูกมองว่า “ควบคุมการรับรู้” โดยผู้ใช้ไม่รู้ตัว
3) **การกำกับดูแลกิจกรรมบนโลกออนไลน์ (Online Governance):** งานศึกษาจาก Reddit พบว่าผู้ดูแลและผู้ใช้มักเห็นพ้องในหลักเกณฑ์ของชุมชน แต่เกิดความไม่สอดคล้องกันใน “วิธีการบังคับใช้กฎ” โดยเฉพาะเมื่อชุมชนเติบโตเร็วเกินไป ส่งผลให้ผู้ใช้บางกลุ่มรู้สึกว่ามีการบังคับใช้กฎอย่างเข้มงวดหรือไม่เป็นธรรม และนำไปสู่การเสื่อมความเชื่อมั่นในแพลตฟอร์ม
4) **Impact Assessment ในระดับเมือง:** ในระดับนโยบายเมืองของสหรัฐ หลายพื้นที่ได้ทดลองใช้ AI Impact Assessment เพื่อประเมินผลกระทบของระบบอัลกอริทึมต่อประชาชนในด้านต่าง ๆ เช่น สิทธิการเข้าถึงที่อยู่อาศัย การขอสินเชื่อ หรือการคัดกรองผู้ป่วย พร้อมพัฒนากลไกให้ประชาชนสามารถโต้แย้งผลลัพธ์ของอัลกอริทึมหากพบความคลาดเคลื่อนหรือความไม่เป็นธรรม แนวคิดนี้ช่วยสร้างสมดุลระหว่างการใช้เทคโนโลยีและการคุ้มครองสิทธิของผู้ได้รับผลกระทบ
5) **หลักการออกแบบ AI ที่เน้นมนุษย์เป็นศูนย์กลาง:** การออกแบบ AI ที่มีมนุษย์เป็นศูนย์กลาง เป็นหลักการที่ขาดไม่ได้ โดยระบบควรยึดคุณภาพของข้อมูลมากกว่าปริมาณ ให้ความสำคัญกับความปลอดภัย ศักดิ์ศรี และความเป็นส่วนตัวของผู้ใช้ ต้องมีความสามารถในการอธิบายเหตุผลของผลลัพธ์ และเปิดโอกาสให้ผู้ใช้ควบคุมประสบการณ์ของตนเองได้จริง ไม่ใช่เพียงภาพลวงตาของการเลือก นอกจากนี้ ต้องให้ความสำคัญกับความเป็นธรรมและความหลากหลายของตัวแทนทางวัฒนธรรม เพื่อให้ระบบสะท้อนความเป็นจริงของผู้คนหลากหลายกลุ่ม ไม่ใช่เฉพาะกลุ่มที่มีข้อมูลมากกว่า

กล่าวโดยสรุป การพัฒนา AI ในศตวรรษนี้ไม่ได้เป็นเพียงการแข่งขันเพื่อสร้าง “โมเดลที่ฉลาดที่สุด” หากแต่เป็นการแข่งขันเพื่อออกแบบระบบที่ตอบสนองความต้องการของสังคมได้ดีที่สุด การออกแบบที่ดีจำเป็นต้องเริ่มจากความเข้าใจมนุษย์ในฐานะผู้ใช้ ผู้ได้รับผลกระทบ และผู้กำกับดูแลเทคโนโลยี ระบบ AI จึงควรทำหน้าที่เป็นเครื่องมือที่เพิ่มศักยภาพและเปิดโอกาสให้มนุษย์พัฒนา ไม่ใช่ลดทอนสิทธิ ความสามารถ หรือโอกาสในชีวิต การกำกับดูแลที่โปร่งใส การตรวจสอบได้อย่างเป็นระบบ และการออกแบบที่ยึดศักดิ์ศรีความเป็นมนุษย์เป็นศูนย์กลาง คือรากฐานสำคัญที่จะทำให้สังคมสามารถอยู่ร่วมกับ AI ได้อย่างมีความรับผิดชอบและยั่งยืนในระยะยาว

Pitchayanin Wadprapan
Pitchayanin.Wadprapan@bangkokbank.com

Strategic Outlook and
Transformation Management
Office of the President