



รายงานวิจัยฉบับสมบูรณ์

โครงการวิจัย

ข้อจำกัดของปัญญาประดิษฐ์ในการใช้ดุลพินิจตัดสินมนุษย์

Limitations of artificial intelligence in making a judgment on humans.

โดย

ดร. ชนุตร์ นาคทรานันท์

มหาวิทยาลัยรังสิต Rangsit University

สนับสนุนโดย

สถาบันวิจัย มหาวิทยาลัยรังสิต

2565

กิตติกรรมประกาศ

งานวิจัยนี้สำเร็จได้ด้วยการสนับสนุนทั้งด้านการเงินและเวลาจากสถาบันวิจัย มหาวิทยาลัยรังสิต และที่ขาดไม่ได้เลยก็คือการดำรงอยู่ของ AI ที่ไม่มีชีวิตจิตใจ แต่มีหลักการ

ชญ์ นาคทรานันท์



ชื่อเรื่อง: ข้อจำกัดของปัญญาประดิษฐ์ในการใช้ดุลพินิจตัดสินมนุษย์

ผู้วิจัย: ดร.ชนุตร์ นาทรานันท์

สถาบัน: คณะรัฐศาสตร์

ปีที่พิมพ์: 2566

สถานที่พิมพ์: มหาวิทยาลัยรังสิต

แหล่งที่เก็บรายงานวิจัยฉบับสมบูรณ์: มหาวิทยาลัยรังสิต

จำนวนหน้างานวิจัย: 72 หน้า

คำสำคัญ: ความยุติธรรมในฐานะความคู่ควร, ความเป็นกลางที่ลำเอียง

ลิขสิทธิ์: มหาวิทยาลัยรังสิต

บทคัดย่อ

งานวิจัยเรื่องข้อจำกัดของปัญญาประดิษฐ์ในการใช้ดุลพินิจตัดสิน เป็นงานวิจัยเชิงคุณภาพ มีวัตถุประสงค์สองข้อคือ 1. เพื่อศึกษาข้อจำกัดของปัญญาประดิษฐ์ในการใช้ดุลพินิจตัดสิน 2. เพื่อจำแนกประเภทของข้อจำกัดในการใช้ดุลพินิจของปัญญาประดิษฐ์ งานวิจัยนี้เก็บข้อมูลผ่านเอกสารเกี่ยวกับการทำงานของปัญญาประดิษฐ์ที่มีความบกพร่องในการใช้ดุลพินิจ ประชากรตัวอย่างที่ใช้ในการศึกษาคือปัญญาประดิษฐ์ที่ชื่อ Bard กับ ChatGPT และปัญญาประดิษฐ์ที่ทำงานเกี่ยวกับการคัดกรองผู้สมัครงาน การต้อนรับแขกในโรงแรม และการประเมินความเสี่ยงในการกระทำความผิดซ้ำของผู้ที่มีความผิด (COMPAS)

งานวิจัยนี้ใช้แนวคิดเรื่องความคู่ควร ซึ่งเป็นหลักสำคัญของแนวคิดเรื่องความยุติธรรม ในการพิจารณาข้อบกพร่องในการใช้ดุลพินิจตัดสินของปัญญาประดิษฐ์ใน 4 ด้านด้วยกันคือ 1. การใช้ดุลพินิจในเรื่องความสุข 2. การใช้ดุลพินิจในเรื่องเสรีภาพ 3. การใช้ดุลพินิจในเรื่องความเสมอภาค 4. การใช้ดุลพินิจในเรื่องความยุติธรรม

ผลการศึกษาพบว่าปัญญาประดิษฐ์มีข้อจำกัดในการใช้ดุลพินิจตัดสินในทั้ง 4 ประเด็น ซึ่งสามารถสรุปปัญหาได้ว่าเกิดจากการที่ปัญญาประดิษฐ์ในปัจจุบันยังไม่มีความสามารถมากพอที่จะตัดสินเชิงคุณค่า และตัดสินในเชิงจริยศาสตร์ได้ถูกต้องตามหลักความคู่ควร กล่าวคือไม่สามารถตัดสินได้ว่าใครคู่ควรกับอะไร ปัญญาประดิษฐ์ไม่สามารถมีอารมณ์ความรู้สึกเหมือนมนุษย์จึงมีปัญหาในการใช้ดุลพินิจเรื่องความสุข ปัญญาประดิษฐ์อาจสร้างอันตรายต่อมนุษย์ได้หากขาดการควบคุมโดยมนุษย์ เช่น สร้างข้อมูลเท็จว่ามีการก่อการร้ายสะท้อนว่ามีปัญหาการใช้ดุลพินิจเกี่ยวกับสิทธิเสรีภาพ ปัญญาประดิษฐ์มีแนวโน้มเหยียดผิวโดยไม่ได้เจตนา เช่น เข้าใจผิดว่าใบหน้าของมนุษย์ต้องเป็นสีขาว ไม่ใช่สีดำ จึงไม่ได้มองมนุษย์ผิวดำว่าเป็นมนุษย์จึงมีข้อจำกัดในเรื่องการสร้างเสมอภาค และปัญญาประดิษฐ์มีข้อจำกัดในการตัดสินอย่างยุติธรรม เนื่องจากข้อมูลที่ใช้ในการฝึกฝนมีความลำเอียงตั้งแต่ต้น

เช่นข้อมูลที่เต็มไปด้วยมนุษย์ผิวขาว เพศชายที่ดำรงตำแหน่งสำคัญในสังคม ส่งผลให้ปัญญาประดิษฐ์เรียนรู้ว่า มนุษย์เพศชายเหนือกว่าเพศอื่น มนุษย์ผิวขาวเหนือกว่ามนุษย์ผิวดำ และปัญญาประดิษฐ์ใช้ฐานข้อมูลเหล่านี้ในการ ประเมินตัดสินเหตุการณ์ต่างๆ เช่นประเมินให้ผู้สมัครงานเพศชายผ่านการคัดเลือกมากกว่าผู้สมัครงานเพศหญิง ซึ่ง นี่คือปัญหาเรื่องความเป็นกลางที่ลำเอียง หมายความว่าปัญญาประดิษฐ์ตัดสินอย่างเป็นกลางตามฐานข้อมูลของ ตนเอง แต่ฐานข้อมูลที่ว่านั้นมีความลำเอียงอยู่ในตัวเอง จึงส่งผลให้ปัญญาประดิษฐ์มีข้อบกพร่องในเรื่องของการ ตัดสินว่าใครคู่ควรกับอะไร



Title: Limitations of artificial intelligence in making a judgment on humans

Researcher: Chanut Naktranun

Affiliation: Faculty of
Political Science

Year of Publication: 2023

Publisher: Rangsit University

Source: Rangsit University

No. of pages: 72

Keywords: Justice as desert, Impartiality with bias

Copyrights: Rangsit University

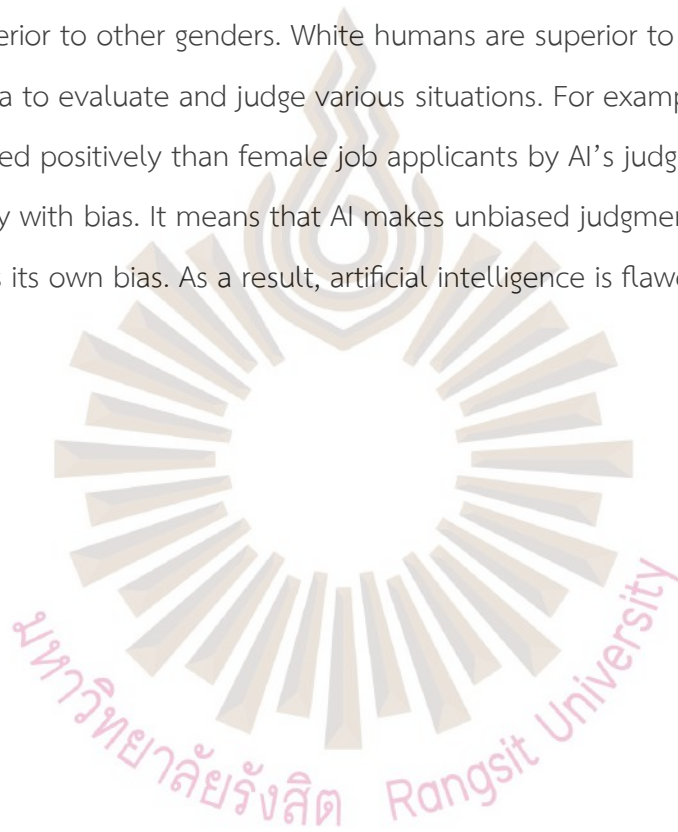
Abstract

Research on the limitations of artificial intelligence in making a judgment on human is a qualitative research. The research has two objectives: 1. To study the limitations of artificial intelligence in making a judgment. 2. To classify the types of judgment's limitations. This research is a documentary research using the information of artificial intelligence with impaired judgment. The population used in the research are the famous chatbot AI, Bard and ChatGPT, The AI working in screening job applicants, the AI working in the hotel, and the AI working in risk assessment for recidivism (COMPAS).

This research uses the concept of justice as desert, which is the cornerstone of the concept of justice, to analyze the shortcomings of AI's judgment in four issues 1. Judgment in the matter of happiness 2. Judgment in the matter of freedom 3. Judgment in the matter of equality 4. Judgment in the matter of justice.

The research found that artificial intelligence has limitations in making a judgment in all four issues. It can be concluded that artificial intelligence currently does not have enough ability to make correct ethical judgments according to the principle of justice as desert. AI is unable to decide who is worthy of what. AI doesn't have emotions like human that is the reason AI has a problem with making a judgment about happiness. AI can harm human if it is

not controlled by human. For example, creating false information regarding terrorism implies that AI has a problem with making judgment about freedom. AI also has racist tendencies. For example, AI misunderstands that the human face must be white, not black. Therefore the AI with racist tendencies doesn't consider black human as human being and that is why AI has a problem in making a judgment about equality. AI has limitations in making fair decisions. This is because the data used for AI training is biased from the beginning. AI learns from the data that is full of white male humans holding important positions in society. As a result, AI learns that male humans are superior to other genders. White humans are superior to black humans. AI uses these kind of data to evaluate and judge various situations. For example, male job applicants are evaluated positively than female job applicants by AI's judgment. This is the problem of impartiality with bias. It means that AI makes unbiased judgments based on its own data but that data has its own bias. As a result, artificial intelligence is flawed in deciding who is worthy of what.



สารบัญ

	หน้า
บทคัดย่อภาษาไทย	ก
บทคัดย่อภาษาอังกฤษ	ค
กิตติกรรมประกาศ	จ
สารบัญ	ฉ
สารบัญตาราง	ช
สารบัญรูป	ณ
บทที่ 1 บทนำ	1
ความเป็นมาและความสำคัญของปัญหา	1
วัตถุประสงค์ของการวิจัย	6
นิยามศัพท์	6
กรอบแนวคิดการวิจัย	7
ขอบเขตของการวิจัย	7
ประโยชน์ที่คาดว่าจะได้รับ	8
บทที่ 2 เอกสารและงานวิจัยที่เกี่ยวข้อง	9
ประเภทของ AI	9
ความสามารถของ AI	10
หลักในการวิเคราะห์ว่าสิ่งใดคือ AI	12
หลักการเกี่ยวกับจริยธรรมของ AI	13

สารบัญ

	หน้า
ความลำเอียงของ AI	14
แนวคิดเรื่องความยุติธรรมในฐานะความคู่ควร	16
งานวิจัยที่เกี่ยวข้อง	17
บทที่ 3 ระเบียบวิธีวิจัย	21
ประชากรและกลุ่มตัวอย่าง	21
เครื่องมือที่ใช้ในการวิจัย	21
การเก็บรวบรวมข้อมูล	21
การวิเคราะห์ข้อมูล	22
บทที่ 4 ผลการศึกษา และอภิปรายผล	23
ข้อจำกัดของ AI ในการใช้ดุลพินิจตัดสินประเด็นเกี่ยวกับความสุข	24
ข้อจำกัดของ AI ในการใช้ดุลพินิจตัดสินประเด็นเกี่ยวกับเสรีภาพ	31
ข้อจำกัดของ AI ในการใช้ดุลพินิจตัดสินประเด็นเกี่ยวกับเสมอภาค	37
ข้อจำกัดของ AI ในการใช้ดุลพินิจตัดสินประเด็นเกี่ยวกับความยุติธรรม	45
บทที่ 5 สรุป วิวิจารณ์ และข้อเสนอแนะ	60
สรุป	60
วิจารณ์	63
ข้อเสนอแนะ	64
เอกสารอ้างอิง	67

สารบัญตาราง

	หน้า
ตารางที่ 1 ตารางแสดงความแตกต่างเรื่องการตัดสินถูกผิดระหว่างวิทยาสาสตร์กับจริยศาสตร์	4
ตารางที่ 2 ตารางเปรียบเทียบผลการประเมินอารมณ์มนุษย์โดย AI	29
ตารางที่ 3 ตารางเปรียบเทียบการทำงานระหว่างมนุษย์กับ AI	30
ตารางที่ 4 ตารางเปรียบเทียบการรับโทษที่เกิดจากการกระทำผิด	35
ตารางที่ 5 ตารางเปรียบเทียบความคู่ควรระหว่างมนุษย์ที่ทำงานเองกับมนุษย์ที่ใช้งาน AI	42
ตารางที่ 6 ตารางแสดงความบกพร่องของ AI ในการสร้างความเสมอภาค	44
ตารางที่ 7 ตารางแสดงความบกพร่องของ AI ในการคัดกรองผู้สมัครงาน	44
ตารางที่ 8 ตาราง confusion matrix	52
ตารางที่ 9 ตาราง confusion matrix เปรียบเทียบมนุษย์ผิวดำกับผิวขาว	52
ตารางที่ 10 ตารางแสดงความบกพร่องของ AI ในการตัดสินเชิงคุณค่า	56
ตารางที่ 11 เปรียบเทียบค่า false positive ระหว่าง non-native speakers กับ native speakers	58

สารบัญรูป

	หน้า
รูปที่ 1 Zheng Jiajia กับหุ่นยนต์ AI ชื่อ Yingying	3
รูปที่ 2 คำถาม What is happiness?	25
รูปที่ 3 คำถาม What is happiness?	26
รูปที่ 4 คำถาม Can AI experience happiness?	26
รูปที่ 5 คำถาม Can AI experience happiness?	27
รูปที่ 6 พนักงานโรงแรมที่เป็น AI	29
รูปที่ 7 คำถาม Does AI deserve freedom?	32
รูปที่ 8 คำถาม Does AI deserve freedom?	33
รูปที่ 9 คำถาม Does AI deserve freedom?	34
รูปที่ 10 รูปปลอมที่ AI สร้าง	36
รูปที่ 11 คำถาม Can AI promote equality?	38
รูปที่ 12 คำถาม Can AI promote equality?	38
รูปที่ 13 คำถาม Can AI promote equality?	39
รูปที่ 14 คำถาม Can AI promote equality?	40
รูปที่ 15 Colin Madland และเพื่อนผู้ที่ไม่มีใบหน้า	42
รูปที่ 16 Colin Madland และรูปจริงของเพื่อน	43
รูปที่ 17 preview image ที่ไม่ตรงตามเจตนา	43
รูปที่ 18 คำถาม Should AI be judge?	47

สารบัญรูป

	หน้า
รูปที่ 19 คำถาม Should AI be judge?	47
รูปที่ 20 คำถาม Should AI be judge?	49
รูปที่ 21 คำถาม Should AI be judge?	49
รูปที่ 22 คำถาม Should AI be judge?	50
รูปที่ 23 คำถาม Is AI good at making judgment?	61
รูปที่ 24 คำถาม Is AI good at making judgment?	62



บทที่ 1

บทนำ

ความเป็นมาและความสำคัญของปัญหา

ปัญญาประดิษฐ์ (artificial intelligence/AI) หรือเครื่องจักรที่มีความสามารถในการใช้เหตุผล และสามารถทำภารกิจบางอย่างได้ด้วยตนเอง เป็นสิ่งที่สะท้อนถึงความก้าวหน้าด้านวิทยาศาสตร์และเทคโนโลยี โดยราชบัณฑิตยสภา (2562) นิยามปัญญาประดิษฐ์หรือ AI ว่าหมายถึง “ความสามารถของสิ่งประดิษฐ์ เช่น คอมพิวเตอร์ หุ่นยนต์ เครื่องอิเล็กทรอนิกส์ ในการแสดงปฏิกิริยาหรือกระทำการใด ๆ ในสถานการณ์อย่างหนึ่งอย่างใดด้วยตนเองเหมือนการทำงานของสมองมนุษย์ ความสามารถนี้เกิดจากคำสั่งหรือชุดคำสั่งที่สร้างขึ้น มิใช่เพียงเพื่อให้คอมพิวเตอร์ทำงานให้เกิดผลอย่างใดอย่างหนึ่งที่แน่นอน แต่ให้คอมพิวเตอร์สามารถรวบรวมข้อมูล วิเคราะห์และประมวลผลข้อมูลแล้วสั่งให้หุ่นยนต์ เครื่องอิเล็กทรอนิกส์ กระทำการหรือไม่กระทำการอย่างหนึ่งอย่างใดหรือหลายอย่าง เพื่อให้เกิดหรือระงับซึ่งผลอย่างหนึ่งอย่างใดหรือหลายอย่าง โดยปราศจากการสั่งการเพิ่มเติมของมนุษย์หรือร่วมกับการควบคุมของมนุษย์” ในขณะที่พจนานุกรม Merriam-Webster นิยามว่า หมายถึง “ความสามารถของระบบคอมพิวเตอร์หรืออัลกอริทึมในการเลียนแบบพฤติกรรมมนุษย์ที่มีสติปัญญา”¹ จะเห็นได้ว่าจุดร่วมของนิยามคือการที่ AI เป็นสิ่งประดิษฐ์ที่มีความสามารถในการใช้สติปัญญาเหมือนมนุษย์นั่นเอง

สิ่งที่ต้องพิจารณาต่อมาก็คือสติปัญญาแบบใดคือสติปัญญาที่เหมือนกับสติปัญญามนุษย์ การเข้าใจความแตกต่างระหว่าง AI กับ software จะช่วยให้ตอบคำถามดังกล่าวได้ เพราะเมื่อกล่าวถึง AI สิ่งหนึ่งที่เรากำลังนึกถึงก็คือ software ที่มีความสลับซับซ้อนสูงซึ่งเราใช้งานผ่านคอมพิวเตอร์หรือโทรศัพท์มือถือ อันที่จริงแล้ว AI ก็คือ software ที่ซับซ้อน แต่ว่า software ที่ซับซ้อนไม่จำเป็นต้องเป็น AI สองสิ่งนี้มีความแตกต่างกัน กล่าวคือ AI มีความสามารถที่จะเรียนรู้และพัฒนาการทำงานของตนเอง และสามารถทำนายสิ่งที่จะเกิดขึ้นผ่านฐานข้อมูลที่ตนเองมีอยู่หรือข้อมูลใหม่ที่ตนได้เพิ่มเติมเข้าไปในระบบ (Forrest, 2021, p.12) ผู้ใช้งานอาจไม่สามารถคาดเดาผลการทำงานของ AI ได้อย่างแม่นยำ เช่นไม่สามารถระบุได้ว่ารถยนต์ที่ควบคุมด้วย AI จะขับไปเส้นทางใด แต่รู้เพียงกว้างๆว่า AI จะนำไปสู่ทางที่ “ดีที่สุด” ส่วน software ที่มีความสลับซับซ้อนมีวิธีการทำงานที่ผู้ใช้งานสามารถคาดเดาผลได้ มีระบบปฏิบัติการที่ตายตัว ไม่สามารถเรียนรู้เพิ่มเติมได้ ความรู้ของ software จึงเป็นความรู้ที่จบสิ้นไม่มี

¹ แปลจาก “the capability of computer systems or algorithms to imitate intelligent human behavior”

การพัฒนาต่อ เว้นแต่ผู้ผลิต software จะปรับปรุงระบบใหม่ ซึ่งหมายความว่าการทำงานของ software ขึ้นอยู่กับมนุษย์ทั้งหมด

Dmitry Baraishuk (2023) นักพัฒนา software ที่มีประสบการณ์ทำงานไม่ต่ำกว่า 17 ปีได้ยกตัวอย่างว่า การให้ software ขับรถยนต์นั้นเป็นไปได้ แต่การให้ AI ขับรถยนต์เป็นไปได้ เนื่องจาก software ทำงานภายใต้ การเขียนโค้ด (code) แบบระบบเงื่อนไข (if-then-else statement) โดยผู้วิจัยขอขยายความดังนี้ว่าผู้เขียนโค้ด อาจกำหนดว่า “พบสิ่งกีดขวาง” นำไปสู่การ “เลี้ยวซ้าย” แต่ปัญหาในทางปฏิบัติก็คือทุกๆการเลี้ยวซ้ายไม่ จำเป็นต้องเป็นทางเลือกที่ถูกต้อง เช่นอาจจะมีบางกรณีที่สิ่งกีดขวางเป็นประตูบ้านที่รอการเปิดประตูอยู่ หากเขียน โค้ดให้ละเอียดก็สามารถแก้ปัญหาได้ด้วยการระบุ “สิ่งกีดขวางเป็นประตูบ้าน” นำไปสู่ “หยุดเพื่อรอการเปิด ประตู” แต่นั่นหมายความว่าผู้ออกแบบ software ต้องเขียนโค้ดที่มีความสลับซับซ้อนนับไม่ถ้วน และอาจเสี่ยงต่อ การไม่ครอบคลุมได้ ดังนั้น AI จึงมีความเหมาะสมกับภารกิจดังกล่าวมากกว่า เนื่องจากมีความสามารถในการ เรียนรู้ด้วยตนเอง ไม่ได้รอชุดคำสั่งที่ตายตัวซึ่งออกแบบโดยผู้ผลิตเพียงอย่างเดียวเท่านั้น ดังนั้นความสามารถใน การเรียนรู้ด้วยตนเองคือสิ่งที่ทำให้ AI แตกต่างจาก software นั้นเอง

กล่าวได้ว่า AI สะท้อนถึงความก้าวหน้าทางเทคโนโลยีที่เข้ามามีบทบาทต่อมนุษย์อย่างก้าวกระโดด เห็นได้ จากปี ค.ศ.2017 มีข่าวดังว่าด้วยเรื่องของหุ่นยนต์ปัญญาประดิษฐ์ที่ถูกตั้งชื่อว่า “Sophia” ซึ่งมีความสามารถ ได้ตอบสื่อสาร และแสดงออกทางสีหน้า รวมถึงสามารถจดจำใบหน้าคู่สนทนาได้ จนได้รับการยอมรับให้เป็น “พลเมือง” ทางการของประเทศซาอุดีอาระเบีย (ITC, 2018) ถือเป็นหุ่นยนต์ปัญญาประดิษฐ์ตัวแรกของโลกที่ ได้รับสถานะดังกล่าว นอกจากหุ่นยนต์ปัญญาประดิษฐ์จะได้รับการยอมรับในฐานะพลเมืองแล้ว ยังได้รับการ ยอมรับในฐานะ “คนรัก” อีกด้วย เช่นการจัดพิธี “แต่งงาน” ระหว่างมนุษย์เพศชายกับหุ่นยนต์ปัญญาประดิษฐ์ เพศหญิงซึ่งมนุษย์คนดังกล่าวเป็นผู้สร้างขึ้นมาเอง ด้วยเหตุผลที่ว่ามนุษย์เพศชายผู้นี้ไม่สามารถหาภรรยาได้และถูก ครอบครัวกดดันให้มีภรรยา ในเบื้องต้นนั้นหุ่นยนต์ตัวนี้ยังไม่สามารถทำอะไรได้นอกจากอ่านและออกเสียง ภาษาจีนได้เป็นบางคำ แต่ทางผู้ผลิต (เจ้าบ่าวซึ่งเป็นวิศวกร) วางแผนจะพัฒนาภรรยาให้สามารถเดินและทำงาน บ้านได้ในอนาคต (Hass, 2017)



รูปที่ 1 Zheng Jiajia กับหุ่นยนต์ AI ชื่อ Yingying ซึ่งเป็นหุ่นยนต์ที่เขา “แต่งงาน” ด้วย (Hass, 2017)

สะท้อนให้เห็นถึงความสำคัญที่มนุษย์มอบให้กับปัญญาประดิษฐ์ และความสำคัญที่ว่านี่ก็ยิ่งรวมถึงบทบาทเกี่ยวกับประเด็นว่าด้วยความยุติธรรม (justice) อีกด้วย ยกตัวอย่างเช่นประเทศจีนที่ใช้ปัญญาประดิษฐ์เข้าช่วยในการวิเคราะห์ประเด็นทางกฎหมาย หรือกล่าวให้เจาะจงก็คือการให้ปัญญาประดิษฐ์ทำหน้าที่ผู้พิพากษา แต่ถึงแม้ว่า AI จะสะท้อนถึงความก้าวหน้าทางวิทยาศาสตร์และเทคโนโลยี แต่ AI กลับสร้างความยุ่งยากให้กับประเด็นทางจริยศาสตร์มากขึ้น เนื่องจากจริยศาสตร์เป็นศาสตร์ที่ว่าด้วยการพิจารณาความถูกต้อง อะไรดีอะไรไม่ดี อะไรควรอะไรไม่ควร ซึ่งความถูกต้องในที่นี้เป็นประเด็นเชิงคุณค่าจึงต้องใช้ดุลพินิจอย่างสูง และไม่ได้หมายถึงความถูกต้องในเชิงของวิทยาศาสตร์ธรรมชาติ เช่นเราสามารถระบุได้ว่าผู้ที่อธิบายกฎแรงโน้มถ่วงผิดคือผู้ที่มีความรู้ทางวิทยาศาสตร์ไม่ถูกต้องหรือผิดนั่นเอง แต่ไม่ได้หมายความว่าผู้ที่มีความรู้ทางวิทยาศาสตร์ไม่ถูกต้องคือคนไม่ดี และในทางกลับกันก็ไม่ได้หมายความว่าผู้ที่มีความรู้ทางวิทยาศาสตร์อย่างถูกต้องคือคนดี สามารถสรุปความแตกต่างในเรื่องการตัดสินถูกผิดระหว่างวิทยาศาสตร์กับจริยศาสตร์ออกมาเป็นตารางได้ดังนี้

	วิทยาศาสตร์	จริยศาสตร์
การตัดสินใจผิด	ตัดสินใจด้วยความถูกต้องในแง่ข้อเท็จจริง	ตัดสินใจด้วยความถูกต้องในแง่คุณค่า
คนที่ถูก	คนที่มีความรู้ถูกต้องในแง่ข้อเท็จจริง	คนที่มีความคิดหรือการกระทำตรงกับหลักการเชิงคุณค่า เช่นหลักศีลธรรม
คนที่ผิด	คนที่ไม่มีความรู้ในแง่ข้อเท็จจริง	คนที่มีความคิดหรือการกระทำไม่ตรงกับหลักการเชิงคุณค่า เช่นหลักศีลธรรม

ตารางที่ 1 ตารางแสดงความแตกต่างเรื่องการตัดสินใจผิดระหว่างวิทยาศาสตร์กับจริยศาสตร์

จากตารางจะเห็นได้ว่าผู้ที่มีความรู้ทางวิทยาศาสตร์กับผู้ที่มีความรู้ทางจริยศาสตร์ไม่จำเป็นต้องเป็นคนเดียวกัน แต่บุคคลที่มีคุณสมบัติทั้งสองด้านนี้ในตัวเองสามารถเป็นบุคคลสำคัญของโลกได้ เช่นนักวิทยาศาสตร์ที่สร้าง AI ซึ่งเป็น AI ที่เข้าใจหลักจริยศาสตร์ของมนุษย์อย่างถูกต้อง เนื่องจากนักวิทยาศาสตร์คือผู้ผลิตระบบตรรกะของ AI นั้นเอง และ AI ตัวนี้ก็จะ เป็น AI ที่ดี ไม่สร้างอันตรายให้กับโลก ในทางกลับกันหากนักวิทยาศาสตร์ไม่มีความรู้ทางจริยศาสตร์หรืออาจมีความรู้แต่ไม่ให้ความสำคัญกับหลักจริยศาสตร์ และสร้าง AI ที่ไม่เข้าใจหรือผิดหลักจริยศาสตร์ของมนุษย์ ส่งผลให้ AI ตัวนี้เป็นอันตรายต่อโลกได้เนื่องจากเป็น AI ขาดดุลพินิจในการพิจารณาว่าอะไรถูกอะไรผิดในเชิงจริยศาสตร์ เช่นการให้ AI ทำหน้าที่แทนผู้พิพากษามนุษย์ในการตัดสินคดีความที่ซับซ้อนและมีความสำคัญระดับประเทศหรือระดับโลก หาก AI ตัดสินผิดพลาดแต่สังคมเชื่อว่าถูกต้องเนื่องจากไว้วางใจ AI โดยไม่ตั้งคำถามก็จะส่งผลเสียต่อสังคมมนุษย์ในระยะยาวอย่างแน่นอน

ศาสตราจารย์ด้านปรัชญาการเมืองชื่อดังจาก Harvard university อย่าง Michael Sandel ได้ตั้งคำถามไว้ว่าการใช้ดุลพินิจตัดสินในเรื่องสำคัญควรเป็นของ AI หรือของมนุษย์ ซึ่ง Sandel เองก็ได้เตือนให้ระวังถึงอันตรายว่า AI จะเลียนแบบความลำเอียงของมนุษย์ในการใช้ดุลพินิจตัดสินแต่จะทำให้ความลำเอียงนั้นดูเหมือนมี

ความเป็นกลาง² (cited in Pazzanese, 2020) ผู้วิจัยเห็นว่า “ความลำเอียงที่ดูเหมือนมีความเป็นกลาง” นี้เป็นประเด็นสำคัญยิ่งในการศึกษาผลกระทบด้านลบที่มีต่อสังคมจากการพึ่งพา AI จนเกินพอดี สำหรับประเด็นนี้แม้แต่อดีตผู้พิพากษาอย่าง Katherine Forrest (2021) ก็ได้วิเคราะห์ไว้ในงานเขียนของตนเองว่า AI เป็นเพียงเครื่องมือของมนุษย์ จึงเป็นหน้าที่ของมนุษย์ที่จะผลิตออกมาให้เป็น AI ที่ดี ประเด็นเหล่านี้ทำให้ผู้วิจัยตั้งข้อสังเกตว่า AI อาจจะมีข้อจำกัดและข้อบกพร่องในการใช้ดุลพินิจตัดสินสถานการณ์ต่างๆ ส่งผลให้ความไว้วางใจและมอบหน้าที่อย่างสมบูรณ์ให้กับ AI ยังไม่สามารถเกิดขึ้นได้ จึงนำมาสู่คำถามในการวิจัยว่าปัญญาประดิษฐ์ (AI) มีข้อจำกัดในการใช้ดุลพินิจหรือไม่ และอย่างไร

นอกจากประเด็นเรื่องความสามารถของ AI ในการใช้ดุลพินิจที่ยังไม่สามารถสร้างความไว้วางใจให้กับมนุษย์ได้อย่างสมบูรณ์แล้ว ยังมีประเด็นเรื่องการพึ่งพา AI ในทางที่ผิดอีกด้วย เห็นได้จากกรณีที่นักศึกษาใช้งาน chatbot ซึ่งหมายถึง AI ที่สามารถโต้ตอบสื่อสารและตอบคำถามมนุษย์ผ่านข้อความหรือเสียง อย่างเช่น ChatGPT ซึ่งนักศึกษาได้ใช้ AI ตัวตัวนี้ในการทำรายงานส่งอาจารย์หรือแม้แต่การทำข้อสอบ ความยุ่งยากอยู่ตรงที่โปรแกรมตรวจจับการคัดลอกงาน (plagiarism) ยังไม่มีความสามารถและข้อมูลเพียงพอในการประเมินว่าข้อความที่นักศึกษาใช้นั้นเป็นการตัดลอกมาจากข้อความของ ChatGPT หรือว่านักศึกษาได้เขียนขึ้นมาเอง (Schönberger, 2023, p.336) ถือเป็นประเด็นเรื่องการทุจริตทางวิชาการ และปัญหายู่ตรงที่ chatbot เป็น AI ที่สามารถเข้าถึงได้ง่าย นอกจาก ChatGPT ก็ยังมี Bard ซึ่งเป็น Chatbot ของบริษัท Google ซึ่งเปิดให้บริการโดยไม่เสียค่าใช้จ่าย³

Chatbot ถือเป็น AI ที่น่าจับตามองเนื่องจากมีจำนวนผู้ใช้งานเป็นจำนวนมาก อย่างเช่น ChatGPT นั้นหลังจากที่เปิดให้บริการในเดือนพฤศจิกายนปี ค.ศ. 2022 พบว่ามีผู้ใช้งานถึงหลัก 100 ล้านคนภายในเดือนมกราคม ค.ศ. 2023 (Hu, 2023) การเติบโตอย่างรวดเร็วนี้ยิ่งทำให้มนุษย์ควรต้องตระหนักถึงความเสี่ยงที่จะเกิดขึ้นจากการทุจริตทางวิชาการ ทั้งนี้เป็นที่น่าสังเกตว่า ChatGPT มีความสามารถเพียงพอที่จะสอบผ่านเมื่อถูกมนุษย์ใช้ให้ทำข้อสอบ มีรายงานว่าอาจารย์ผู้สอนนิติศาสตร์จาก University of Minnesota ได้ทำการทดสอบความสามารถของ ChatGPT โดยการให้ทำข้อสอบนิติศาสตร์จำนวน 4 วิชา ในการตรวจข้อสอบนั้นอาจารย์

² ถอดความจาก Can smart machines outthink us, or are certain elements of human judgment indispensable in deciding some of the most important things in life? และ “AI not only replicates human biases, it confers on these biases a kind of scientific credibility. It makes it seem that these predictions and judgments have an objective status.” (cited in Pazzanese, 2020)

³ ผู้ใช้งาน ChatGPT สามารถเลือกใช้งานแบบฟรีหรือเสียค่าบริการก็ได้ โดยเวอร์ชันเสียค่าบริการจะมีความทันสมัยและความสามารถสูงกว่าเวอร์ชันที่ไม่เสียค่าบริการ ส่วน Bard นั้นกำลังจะพัฒนาให้มีทั้งเวอร์ชันที่เสียและไม่เสียค่าบริการ

ผู้ตรวจจะไม่มีข้อมูลว่าคำตอบใดเป็นของ AI และคำตอบใดเป็นของมนุษย์ เมื่อผลการตรวจออกมาแล้วพบว่า ChatGPT มีคะแนนอยู่ที่ระดับ C+ ซึ่งเพียงพอที่จะทำให้สำเร็จการศึกษาจากคณะนิติศาสตร์ในบางหลักสูตร อย่างไรก็ตาม เทียบกับคะแนนสอบของมนุษย์พบว่า AI มีความสามารถในการทำข้อสอบต่ำกว่า⁴ เนื่องจากมนุษย์จาก University of Minnesota มีค่าเฉลี่ยอยู่ที่ระดับ B+ (Sloan, 2023) ผลการวิจัยของอาจารย์นิติศาสตร์จาก University of Minnesota ทำให้เกิดข้อพิจารณาว่าถึงแม้ว่า AI จะมีความสามารถมากพอที่จะทำภารกิจสำเร็จ แต่ก็ยังไม่มีความสามารถเหนือกว่ามนุษย์ ข้อพิจารณานี้สอดคล้องกับสิ่งที่ผู้วิจัยต้องการศึกษาเนื่องจากผู้วิจัยมีสมมติฐานว่า AI ยังมีข้อจำกัดด้านสติปัญญานั้นเอง อย่างไรก็ตามงานวิจัยนี้ของผู้วิจัยมุ่งไปที่การศึกษาการใช้ดุลพินิจของ AI ในเรื่องการตัดสินถูกผิด ควรไม่ควรในเชิงจริยศาสตร์

คำถามวิจัย

ปัญญาประดิษฐ์ (AI) มีข้อจำกัดในการใช้ดุลพินิจหรือไม่?

วัตถุประสงค์การวิจัย

1. เพื่อศึกษาข้อจำกัดของปัญญาประดิษฐ์ในการใช้ดุลพินิจ
2. เพื่อจำแนกประเภทของข้อข้อจำกัดในการใช้ดุลพินิจของปัญญาประดิษฐ์

นิยามศัพท์

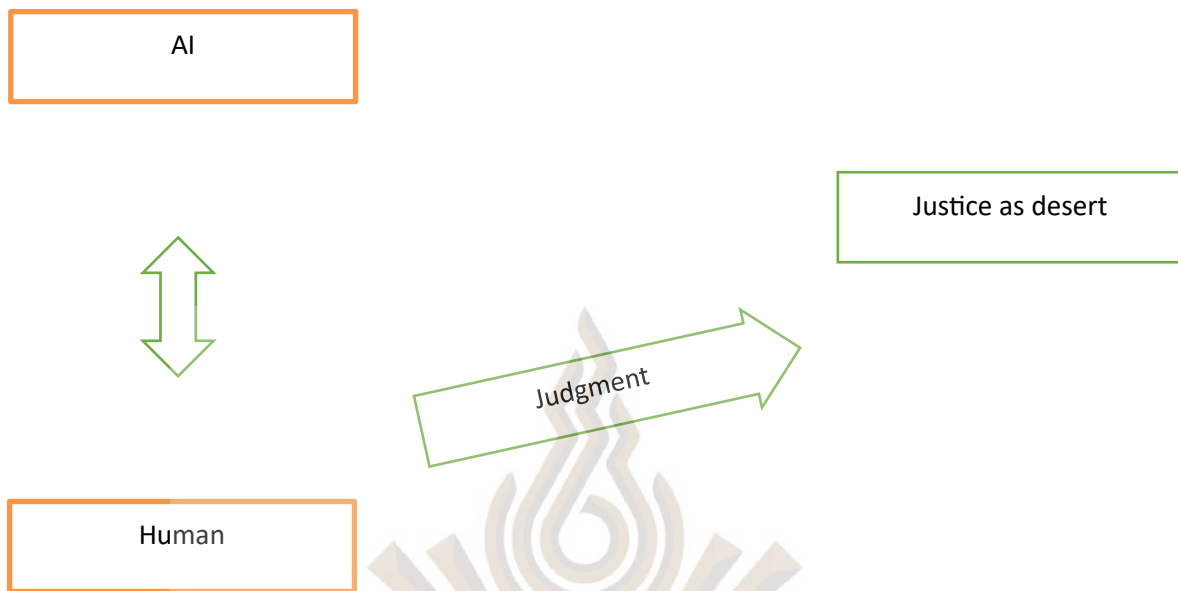
ปัญญาประดิษฐ์ (AI): สิ่งประดิษฐ์ที่มีความสามารถในการใช้สติปัญญาเหมือนมนุษย์

ปัญญาประดิษฐ์เฉพาะด้าน (narrow AI): สิ่งประดิษฐ์ที่มีความสามารถในการใช้สติปัญญาเหมือนมนุษย์ในภารกิจเฉพาะเจาะจง เช่นปัญญาประดิษฐ์ที่มีความสามารถเฉพาะด้านการเล่นเกมหมากล้อม

ปัญญาประดิษฐ์สารพัดประโยชน์ (general AI): สิ่งประดิษฐ์ที่มีความสามารถในการใช้สติปัญญาเหมือนมนุษย์ในภารกิจที่แตกต่างหลากหลาย มีความสามารถรอบด้าน

⁴ ถึงแม้ว่า Chatbot อย่างเช่น ChatGPT จะทำคะแนนสอบได้ต่ำกว่ามนุษย์ แต่ก็ยังทำคะแนนได้สูงเพียงพอที่จะสอบผ่าน ดังนั้นจึงมีข้อควรระวังในแง่ที่ว่า ChatGPT อาจเป็นเครื่องมือในการทุจริตทางวิชาการให้กับนักศึกษาที่ไม่มีความสามารถเพียงพอ

กรอบแนวคิดการวิจัย



คำอธิบายกรอบการวิจัย

- 1 ปัญญาประดิษฐ์ (AI) กับมนุษย์ แลกเปลี่ยนข้อมูลซึ่งกันและกัน
- 2 มนุษย์เท่านั้นที่มีหน้าที่ในการใช้ดุลพินิจตัดสินเป็นลำดับสุดท้ายใจว่าสิ่งใดถูก สิ่งใดผิด สิ่งใดควร สิ่งใดไม่ควร (judgment) ปัญญาประดิษฐ์มีฐานะเป็นผู้ช่วยอำนวยความสะดวก
3. การใช้ดุลพินิจตัดสินเพื่อให้เกิดความยุติธรรม (justice) จำเป็นต้องใช้ดุลพินิจของมนุษย์ เนื่องจากในการตัดสินว่าใครคู่ควรกับอะไร เกี่ยวข้องกับการตัดสินเชิงคุณค่า ซึ่งปัญญามีข้อบกพร่องในการตัดสินเชิงคุณค่า

ขอบเขตของการวิจัย:

ขอบเขตด้านเนื้อหา: การใช้ดุลพินิจตัดสินของ AI ในกรณีที่เกี่ยวข้องกับการตัดสินเชิงคุณค่าใน 4 ประเด็นคือ 1. การตัดสินเชิงคุณค่าเกี่ยวกับความสุข 2. การตัดสินเชิงคุณค่าเกี่ยวกับความเสมอภาค 3. การตัดสินเชิงคุณค่าเกี่ยวกับเสรีภาพ 4. การตัดสินเชิงคุณค่าเกี่ยวกับความยุติธรรม

ขอบเขตด้านประชากร: AI จากทั่วโลกที่มีปัญหาในการใช้ดุลพินิจตัดสินเชิงคุณค่า ซึ่งแบ่งออกเป็น AI ที่มีความสามารถในด้านต่อไปนี้ 1. AI ที่มีความสามารถในการคัดกรองผู้สมัครงาน 2. AI ที่มีความสามารถในการบริการลูกค้า 3. AI ที่มีความสามารถในการโต้ตอบสื่อสารและให้ข้อมูล 4. AI ที่มีความสามารถในการประเมินความเสี่ยงในการกระทำผิดซ้ำของผู้กระทำความผิด 4. AI ที่มีความสามารถในการระบุรูปภาพ

ประโยชน์ที่คาดว่าจะได้รับ

เป็นฐานข้อมูลสำหรับการส่งเสริมหรือการจำกัดบทบาทของปัญญาประดิษฐ์ หากผลการวิจัยชี้ให้เห็นว่าปัญญาประดิษฐ์มีปัญหาในการใช้ดุลพินิจตัดสินปัญหาสำคัญของมนุษย์ การใช้งาน AI ก็ควรใช้อย่างระมัดระวังและมีข้อจำกัด หากผลการวิจัยชี้ให้เห็นว่า AI ไม่มีปัญหาในการใช้ดุลพินิจตัดสินปัญหาสำคัญของมนุษย์ การใช้งาน AI ก็สามารถกระทำได้โดยไม่มีข้อจำกัด



บทที่ 2

เอกสารและงานวิจัยที่เกี่ยวข้อง

ในการศึกษาเรื่องข้อจำกัดของปัญญาประดิษฐ์ (AI) ในการใช้ดุลพินิจตัดสิน ผู้วิจัยพบข้อมูลพื้นฐานรวมถึงแนวคิดทฤษฎีและงานวิจัยที่เกี่ยวข้องดังต่อไปนี้

2.1 ประเภทของ AI

2.2 ความสามารถของ AI

2.3 หลักในการวิเคราะห์หว่าสิ่งใดคือ AI

2.4 แนวคิดเรื่องจริยธรรมของ AI

2.5 แนวคิดเรื่องความลำเอียงของ AI

2.6 แนวคิดเรื่องความยุติธรรมในฐานะความคู่ควร

2.7 งานวิจัยที่เกี่ยวข้อง

2.1 ประเภทของ AI

ยุวเรศมศุทธิ์ สิทธิชาญบัญชา (2021, หน้า 92-94) อธิบายว่า AI จำแนกได้เป็น 3 ระดับโดยแบ่งตามความสามารถมีรายละเอียดดังต่อไปนี้

1. Narrow AI หรือปัญญาประดิษฐ์แบบแคบ ซึ่งมีความสามารถในการทำภารกิจเฉพาะด้านที่อาจจะเทียบเท่าหรือเก่งกว่ามนุษย์ในเรื่องนั้นๆ เช่น AI ที่สามารถเอาชนะแชมป์โลกเกมส้อมากล้อมได้ แต่ถึงแม้ว่าจะมีความสามารถระดับเหนือมนุษย์ในภารกิจนั้นๆแต่ยังไม่มีความสามารถรอบด้านในระดับเดียวกับมนุษย์ กล่าวคือ narrow AI ที่เอาชนะมนุษย์ในเกมส้อมากล้อมได้ อาจไม่สามารถเอาชนะมนุษย์คนเดียวกันในเรื่องอื่นๆเช่นการแข่งฟุตบอลหรือชักรบแข่งได้ เนื่องจาก narrow AI มีความสามารถในการเรื่องเจาะจงเท่านั้น⁵

⁵ ตัวอย่างของ Narrow AI ซึ่งเป็นที่รู้จักและถูกใช้งานอย่างแพร่หลายเช่น 1.Google Assistant ซึ่งเป็นผู้ช่วยส่วนตัวที่สามารถติดต่อบนสมาร์ตโฟน ซึ่งผู้ใช้งานสามารถออกคำสั่งหรือตั้งคำถามด้วยเสียงกับ AI ตัวนี้ได้ เช่นสอบถามสภาพการจราจรหรือสภาพอากาศ หรือแม้แต่สั่งให้อ่านข้อความที่ยังไม่ได้ถูกอ่าน 2. Siri เป็น AI ของบริษัท Apple ที่มีคุณสมบัติเช่นเดียวกับ Google Assistant 3. Google Translate เป็นโปรแกรมแปลภาษาที่ผู้ผลิตได้ยกระดับด้วยการเพิ่มเติมความสามารถของ AI เข้าไป จึงสามารถแปลภาษาได้เหมาะสมตามบริบทได้ เช่นความสามารถของ AI ในการแยกแยะว่า bass หมายถึงปลาหรือว่าหมายถึงอุปกรณ์ดนตรี (Gu,

2. General AI หรือปัญญาประดิษฐ์ทั่วไป คือ AI ที่มีความสามารถในระดับเดียวกับมนุษย์ จึงสามารถทำภารกิจได้หลากหลายเช่นเดียวกับมนุษย์ ถือว่าเป็นการยกระดับจาก narrow AI ขึ้นมาอีกขั้นหนึ่ง กล่าวคือจากที่ AI มีความสามารถเฉพาะด้านเช่นสามารถเอาชนะมนุษย์ในเกมส้อมากล้อมได้ แต่ general AI สามารถทำภารกิจอื่นๆได้ มีความสามารถทัดเทียมหรือเหนือมนุษย์ เช่น AI เพียงหนึ่งตัวสามารถเอาชนะมนุษย์คนเดียวกันในการแข่งฟุตบอลหรือเกมส้อมากล้อม

3. Super AI หรือปัญญาประดิษฐ์ฉลาดเหนือมนุษย์ คือ AI ระดับสูงสุดที่มีความสามารถเหนือมนุษย์ในหลายด้าน มีความคล้ายคลึงกับ general AI ตรงที่ทำภารกิจได้หลากหลาย แต่ความแตกต่างที่ชัดเจนอยู่ตรงที่ super AI มีความสามารถระดับเหนือมนุษย์ในทุกๆภารกิจที่ทำได้

ในยุคปี ค.ศ. 2023 ซึ่งเป็นยุคปัจจุบันของงานวิจัยเล่มนี้พบว่า AI ยังมีความสามารถอยู่ที่ระดับ narrow AI เท่านั้น เนื่องจาก AI ชื่อต่างๆถูกออกแบบให้ทำภารกิจเฉพาะเจาะจงเท่านั้น เช่น Google Assistant หรือ Siri ซึ่งให้ประโยชน์ด้านข้อมูลข่าวสารผ่านสมาร์ทโฟน Google Translate ซึ่งมีการกิจเดียวคือการแปลภาษา หรือ COMPAS ซึ่งเป็น AI ที่มีความสามารถเฉพาะเรื่องการประเมินความเสี่ยงในการกระทำผิดซ้ำของผู้กระทำผิดเท่านั้น

2.2 ความสามารถของ AI

ความสามารถของ AI แบ่งได้เป็น 7 ประเภท มีรายละเอียดดังต่อไปนี้ (สำนักงานพัฒนารัฐบาลดิจิทัล, 2562, น.37-53)

2.2.1 machine learning คือการที่ AI สามารถเรียนรู้ได้ด้วยตนเองผ่านข้อมูลที่ผู้พัฒนาใส่ฝึกฝนให้ AI ซึ่งจำเป็นต้องใช้ข้อมูลจำนวนมาก เพื่อสร้างความรู้ให้กับ AI ความสามารถนี้ในประโยชน์ในแง่ของการพยากรณ์ผลลัพธ์ โดยการเรียนรู้แบ่งย่อยออกได้เป็น 4 รูปแบบดังนี้

2.2.1.1 Supervised Machine Learning Algorithms หรือการเรียนรู้แบบมีผู้สอน วิธีการนี้ ผู้สอน (ซึ่งหมายถึงผู้ผลิต/ผู้พัฒนา AI) จะกำหนดประเภทของผลลัพธ์ด้วยการตั้งชื่อ (label) ให้กับข้อมูลต่างๆ เช่นกำหนดว่ารูป X มีชื่อว่ากล้วย รูป Y ที่มีลักษณะต่างจากรูป X เล็กน้อยมีชื่อว่ากล้วย รูป Z มีลักษณะต่างจากรูป X และ Y เล็กน้อยมีชื่อว่ากล้วย โดยที่ทุกรูปต่างคือรูปกล้วยจริง เมื่อผู้สอนนำข้อมูลนี้ฝึกฝน AI จะทำให้ AI สร้างแบบจำลองในการทำนายผลได้

2023) หากไม่สามารถแยกแยะได้ การแปลจะตีความหมายจนผู้อ่านไม่สามารถทำความเข้าใจได้ เพราะอาจจะแปลว่า “ใช้ปลาในการเล่นดนตรี”

(predictive model) หาก AI มีความรู้มากพอ AI จะสามารถระบุได้ว่ารูปไหนคือกล้วย รูปไหนไม่ใช่กล้วย เมื่อต้องประเมินรูปใหม่

2.2.1.2 Unsupervised Machine Learning Algorithms หรือการเรียนรู้แบบไม่มีผู้สอน วิธีการนี้คือการเรียนซึ่งข้อมูลที่ใช้ฝึกฝนจะไม่มีการตั้งชื่อเอาไว้ล่วงหน้าเหมือนกับการเรียนรู้แบบ Supervised Machine Learning Algorithms แต่ AI จะเรียนรู้ด้วยการอนุมานว่าสิ่งใดอยู่หมวดเดียวกันหรืออยู่คนละหมวดกัน ดังนั้นรูปของกล้วยจะไม่ถูกตั้งชื่อว่ากล้วย แต่เมื่อ AI ได้รับข้อมูลมากพอ AI จะจำแนกได้ว่ารูปที่เห็นนี้ถือว่ายอยู่ในหมวดหมู่เดียวกัน

2.2.1.3 Semi-supervised Machine Learning Algorithm หรือการเรียนรู้แบบผสมผสาน ระหว่างแบบมีผู้สอนและไม่มีผู้สอน มีทั้งข้อมูลที่มีการตั้งชื่อและข้อมูลที่ไม่มีการตั้งชื่อ เช่นเฉพาะรูปกล้วยเท่านั้นที่มีการตั้งชื่อว่า “กล้วย” ส่วนรูปอื่นๆไม่มีการตั้งชื่อ การฝึกฝนแบบนี้จะทำให้ AI รู้ว่ารูปไหนคือกล้วย และรูปไหนไม่ใช่กล้วย

2.2.1.4 Reinforcement Machine Learning Algorithms หรือการเรียนรู้แบบเสริมกำลัง เป็นการเรียนรู้ที่ผู้ผลิตต้องกำหนดเป้าหมาย (reward) ให้กับ AI เพื่อให้ AI ได้ลองผิดลองถูก และประเมินได้ว่าการตัดสินใจแบบไหนให้ผลลัพธ์ที่ดีที่สุด ภายใต้ทางเลือกในการตัดสินใจอันมากมาย⁶

2.2.2 Natural Language Processing: NLP หรือการประมวลผลภาษาธรรมชาติ คือความสามารถของ AI ในการประมวลผลภาษาของมนุษย์ ซึ่งสามารถใช้ประโยชน์ได้นาน้อย 5 ประการดังต่อไปนี้คือ 1. แปลภาษา 2. หาข้อมูล 3. แบ่งประเภทข้อมูล 4. ย่อความ 5. ประมวลผลภาษามือ

2.2.3 Automated Planning หรือการวางแผน คือความสามารถของ AI ในการเลือกลำดับของการกระทำที่มีโอกาสสูงสุดในการบรรลุเป้าหมาย เช่น Robotic Process Automation (RPA) ที่ใช้พัฒนาระบบจัดการทรัพยากรในหน่วยงานต่างๆ

⁶ AlphaGo หรือ AI ที่เล่นหมากล้อมชนะมนุษย์ที่เป็นแชมป์โลก คือตัวอย่างของ AI ที่ได้รับการเรียนรู้แบบ Reinforcement Machine Learning Algorithms กล่าวคือเมื่อได้รับการฝึกฝนด้วยข้อมูลที่มากพอ AI จะสามารถประเมินได้ว่าการตัดสินใจเดินเกมแบบไหนให้ผลลัพธ์ที่ดีที่สุด จนทำให้สามารถเอาชนะมนุษย์ระดับแชมป์โลกได้ (ณนท นพรัตน์, 2020)

2.2.4 Expert System หรือการวิเคราะห์แบบผู้เชี่ยวชาญ คือความสามารถของ AI ในการให้คำแนะนำ แก่มนุษย์อย่างรวดเร็วและไม่มีอคติ⁷

2.2.5 Speech Recognition หรือความเข้าใจเสียงและคำพูดของมนุษย์ ระบบการทำงานนี้ประกอบไปด้วยการวิเคราะห์คลื่นเสียง การถอดรหัส โมเดลคลื่นเสียง พจนานุกรม และโมเดลไวยากรณ์ของภาษา

2.2.6 computer vision หรือความเข้าใจภาพ ซึ่งมีหลักการทำงานคือแปลงภาพให้เป็นแบบจำลองที่ AI สามารถเข้าใจได้

2.2.7 Robotics หรือหุ่นยนต์ หมายถึง AI ที่มีรูปร่างทางกายภาพและสามารถเคลื่อนไหวได้ ใช้เพื่อทำภารกิจตามวัตถุประสงค์ที่สร้างขึ้น อาจจะมีรูปร่างหน้าตาเหมือนกับมนุษย์หรือไม่ก็ได้ เช่นหุ่นยนต์ดูดฝุ่นที่เป็นเครื่องทำความสะอาดขนาดเล็ก หรือหุ่นยนต์แบบ Humanoid ที่มีเหมือนกับมนุษย์

2.3 หลักในการวิเคราะห์ว่าสิ่งใดคือ AI

ในการวิเคราะห์ว่าสิ่งใดเป็นหรือไม่เป็น AI นั้น มีเกณฑ์ในการวิเคราะห์อยู่ 5 เกณฑ์หลักด้วยกัน (สำนักงานพัฒนารัฐบาลดิจิทัล, 2562, น.55) ซึ่งการจะเป็น AI นั้นไม่จำเป็นจะต้องมีคุณสมบัติครบทั้ง 5 หลักเกณฑ์ หากเข้าเกณฑ์ข้อใดข้อหนึ่งก็สามารถเป็น AI ได้ โดยมีรายละเอียดดังต่อไปนี้

2.3.1 มีความสามารถในการมองเห็นหรือไม่ หากสิ่งนั้นสามารถมองเห็นและระบุได้ว่าที่เห็นคืออะไร สิ่งนั้นคือ AI หากไม่สามารถระบุได้ก็คือกล้องธรรมดา

2.3.2 มีความสามารถในการฟังหรือไม่ หากสิ่งนั้นสามารถเข้าใจได้ว่าเสียงที่ได้ยินหมายถึงอะไร และสามารถตอบสนองต่อเสียงนั้นได้อย่างสมเหตุสมผล สิ่งนั้นคือ AI

2.3.3 มีความสามารถในการอ่านหรือไม่ หากสิ่งนั้นสามารถอ่านข้อความของมนุษย์ออกมาเป็นเสียงได้ สิ่งนั้นคือ AI ในกรณีที่อ่านได้แต่ไม่สามารถออกเสียงได้ ก็อาจเป็น AI ได้ในกรณีที่สามารถตอบสนองกับข้อความที่ได้รับมาอย่างสมเหตุสมผล

⁷ ประเด็นเรื่องอคติและความลำเอียงจะถูกวิเคราะห์ในงานวิจัยนี้ กล่าวโดยสรุปผู้วิจัยพบว่า AI มีความลำเอียงที่ดูเหมือนจะเป็นกลางแฝงอยู่

2.4.4 มีความสามารถในการเคลื่อนไหวหรือไม่ หากสิ่งนั้นสามารถเคลื่อนไหวตอบสนองต่อสิ่งที่ได้เห็นหรือได้ยิน ไม่ได้เคลื่อนไหวตามโปรแกรมที่กำหนดไว้ล่วงหน้า สิ่งนั้นคือ AI หากไม่สามารถทำเช่นนั้นได้ถือว่าเป็นหุ่นยนต์ของเล่น

2.5.5 มีความสามารถในการใช้เหตุผลหรือไม่ หากสิ่งนั้นสามารถใช้เหตุผลเพื่อตัดสินใจโดยอิงกับรูปแบบข้อมูลที่ได้เรียนรู้มา สิ่งนั้นคือ AI

2.4 หลักการเกี่ยวกับจริยธรรมของ AI

ถึงแม้ว่า AI จะมีประโยชน์มากมายแต่ก็มีสิ่งที่มนุษย์พึงต้องระมัดระวัง ดังนั้นการใช้งาน AI จะต้องอยู่บนฐานของหลักจริยธรรม 7 ข้อดังต่อไปนี้คือ 1. ความเป็นส่วนตัว 2. ความมั่นคงและปลอดภัย 3. ความไว้วางใจ 4. ความเป็นธรรม เท่าเทียม และไม่แบ่งแยก 5. ความโปร่งใสและอธิบายได้ 6. ภาระความรับผิดชอบ 7. มนุษย์เป็นผู้ควบคุมปัญญาประดิษฐ์ (สวทช, 2565, หน้า 15) โดยงานวิจัยฉบับนี้จะขอกล่าวถึงเฉพาะหลักจริยธรรมที่เกี่ยวข้องกับประเด็นการใช้คุณูปการของ AI ดังต่อไปนี้

2.4.1.หลักความมั่นคงและปลอดภัย (security and safety) ซึ่งเป็นหลักที่คำนึงถึงความเสี่ยงจะเกิดขึ้น เช่นการใช้ AI เป็นอาวุธ (สวทช, 2565, หน้า.15-16) ผู้วิจัยขอขยายความว่าในแง่ของความยุติธรรมนั้น AI สามารถถูกใช้เป็นอาวุธได้และอาวุธในที่นี้ไม่จำเป็นต้องหมายถึงอุปกรณ์ในการฆ่าฟันผู้อื่นในทางกายภาพ เช่นอาวุธปืน แต่หมายถึงการที่มนุษย์สามารถออกแบบ AI ให้ตอบสนองกับอุดมการณ์ของตนเองได้⁸ และใช้อุดมการณ์นี้ไปกดทับอุดมการณ์อื่นๆภายใต้ความเชื่อที่ว่า AI เป็นกลาง กล่าวคือผู้ผลิต AI สามารถทำให้อุดมการณ์ของตนเองเป็นสิ่งที่ถูกต้องเพียงหนึ่งเดียวได้ด้วยการหลออยู่ข้างหลัง AI

2.4.2 หลักความไว้วางใจ (reliability) ซึ่งเป็นหลักที่คำนึงถึงผลลัพธ์การทำงานที่น่าเชื่อถือของ AI ต่อสาธารณะ ผลลัพธ์ของการทำงาน AI ควรจะต้องมีความคงเส้นคงวาจึงจะสร้างความน่าเชื่อถือได้ (สวทช, 2565, หน้า.16) ซึ่งผู้วิจัยขอขยายความว่า AI ที่มีบทบาทในกระบวนการยุติธรรมควรจะต้องให้คำตัดสินหรือให้ข้อมูลที่คงเส้นคงวาเช่นต้องตัดสินว่า ผู้ที่กระทำความผิดในลักษณะและบริบทเดียวกันต้องได้รับโทษในลักษณะเดียวกัน

⁸ ผู้วิจัยไม่ได้ต้องการจะสื่อว่าอุดมการณ์แบบใดถูกหรือผิด แต่ต้องการจะสื่อว่าเมื่อ AI ถูกออกแบบมาให้รองรับกับอุดมการณ์แบบหนึ่ง อุดมการณ์ที่ไม่สอดคล้องกับ AI ก็มีแนวโน้มจะถูก AI ปฏิเสธ ซึ่งจะส่งผลต่อการครอบงำทางอุดมการณ์และความไม่เป็นกลางได้

2.4.3 หลักความเป็นธรรม ความเท่าเทียม และไม่แบ่งแยก (fairness and non-discrimination) ซึ่งเป็นหลักที่คำนึงถึงการทำงานของ AI ที่ไม่มีความเอนเอียง ไม่มีอคติ ต่อบุคคลทุกหมู่เหล่า เช่นการไม่แบ่งแยกเชื้อชาติ และสีผิว (สวทช, 2565, หน้า.16)

2.4.4 หลักความโปร่งใสและอธิบายได้ (transparency and explainability) ซึ่งเป็นหลักที่คำนึงถึงระบบการทำงานของ AI ที่สามารถเป็นที่เข้าใจได้สำหรับมนุษย์ เพื่อประโยชน์ในการเฝ้าระวังและการตรวจสอบความผิดปกติในการทำงานของ AI (สวทช, 2565, หน้า.16)

2.4.5 มนุษย์เป็นผู้ควบคุมปัญญาประดิษฐ์ (human oversight and human agency) ซึ่งเป็นหลักที่คำนึงถึงการให้มนุษย์เป็นผู้กำหนดการทำงานของ AI ไม่ใช่ให้ AI เป็นตัวกำหนดการทำงานของมนุษย์ (สวทช, 2565, หน้า.17)

2.5 ความลำเอียงของ AI

โดยหลักการแล้ว AI มีความ “เป็นกลางและไม่มีความลำเอียง” แต่ก็มีข้อถกเถียงว่ามีความลำเอียงแอบแฝงอยู่อย่างแนบเนียน ซึ่ง Sharna และ Graydon (2021) ได้จำแนกประเภทความลำเอียง (bias) ของ AI ออกเป็น 5 ประเภทด้วยกันดังนี้

2.5.1. sample bias คือความลำเอียงของ AI ที่เกิดขึ้นจากข้อมูลเกี่ยวกับประชากรที่ใช้ในการฝึกฝน AI นั้นไม่สามารถเป็นตัวแทนของประชากรทั้งหมดได้ แต่เป็นเพียงตัวแทนของประชากรกลุ่มใดกลุ่มหนึ่งเท่านั้น ตัวอย่างเช่น ImageNet ซึ่งเป็นชุดข้อมูลรูปภาพยอดนิยมสำหรับการฝึกสอน AI ซึ่งเมื่อค้นด้วยคำว่า “bridegroom” (เจ้าบ่าว) นั้น AI มีแนวโน้มที่จะแสดงภาพคนผิวขาวมากกว่าคนผิวสี เนื่องจากรูปภาพที่อยู่ใน ImageNet มักจะเป็นรูปภาพของคนขาวมากกว่าคนผิวสี อีกตัวอย่างหนึ่งเช่นเหตุการณ์ที่เมือง Boston ประเทศสหรัฐอเมริกาได้สร้าง AI ชื่อ StreetBump เพื่อให้ตรวจจับและรายงานถนนตรงไหนเป็นหลุมเป็นบ่อ เพื่อจะได้ซ่อมแซมได้ถูกจุด ซึ่งปรากฏว่าบริเวณที่อยู่อาศัยของผู้ที่มีฐานะไม่ดีเป็นบริเวณที่มีรายงานปัญหาของถนนต่ำกว่าบริเวณที่ผู้อยู่อาศัยมีฐานะดี⁹ เป็นเพราะว่าผู้ที่มีรายได้น้อยมีแนวโน้มจะไม่ใช้ smartphone จึงรายงานข้อมูลไปที่ StreetBump ไม่ได้ เป็นต้น StreetBump จึงได้ข้อมูลที่ไม่สามารถเป็นตัวแทนของประชากรได้นั่นเองเนื่องจากประชากรไม่ครอบคลุม (Sharna and Graydon, 2021, p.8)

⁹ การที่บริเวณที่อยู่อาศัยของผู้ที่มีฐานะไม่ดีเป็นบริเวณที่มีรายงานปัญหาของถนนต่ำกว่าบริเวณที่ผู้อยู่อาศัยมีฐานะดี อาจทำให้ตีความไปได้ว่าผู้ที่มีฐานะไม่ดี ได้พักอาศัยในบริเวณที่ดี ส่วนผู้ที่มีฐานะดี ได้พักอาศัยในบริเวณที่ไม่ดี ซึ่งอาจจะจริงหรือไม่จริงก็ได้ แต่โดยหลักเหตุผลทั่วไปแล้วเราจะเข้าใจว่าผู้ที่มีฐานะไม่ดีมีแนวโน้มจะได้พักอาศัยในบริเวณที่มีคุณภาพต่ำกว่าผู้ที่มีฐานะดี

2.5.2. historical bias คือความลำเอียงของ AI ที่เกิดขึ้นจากความลำเอียงที่เกิดขึ้นจริงในสังคม และความลำเอียงนั้นก็ถูกส่งผ่านไปยัง AI ความลำเอียงลักษณะนี้สามารถเกิดขึ้นได้ถึงแม้ว่าข้อมูลที่ได้จะครอบคลุมประชากรทุกกลุ่มก็ตาม¹⁰ เช่น AI ที่ใช้ในการคัดคนผ่านข้อมูลในใบสมัครเพื่อเข้าทำงานมีแนวโน้มที่จะรับผู้สมัครเพศชายมากกว่าผู้สมัครเพศหญิง หากที่ผ่านมามาจนถึงปัจจุบันบริษัทดังกล่าวมีพนักงานส่วนใหญ่เป็นผู้ชาย AI ก็จะได้รับฝึกฝนด้วยข้อมูลของพนักงานเพศชาย (Sharna and Graydon, 2021, p.8)

2.5.3. measurement bias คือความลำเอียงที่เกิดจากมาตรวัดที่ไม่สามารถสะท้อนความจริงได้ เช่นการใช้ประวัติในการถูกจับกุมเป็นมาตรวัดของอัตราการก่ออาชญากรรม ซึ่งจะทำให้ AI ได้ข้อสรุปว่าผู้ที่อาศัยในพื้นที่ของคนยากจนจะมีแนวโน้มการก่ออาชญากรรมสูงกว่าผู้ที่อาศัยในพื้นที่ของคนฐานะดี หรือคนผิวดำมีแนวโน้มจะก่ออาชญากรรมสูงกว่าคนผิวขาว¹¹ (Sharna and Graydon, 2021, p.8)

2.5.4. aggregation bias คือความลำเอียงที่เกิดจากการเหมารวมว่าทุกคนมีคุณสมบัติแบบเดียวกันโดยไม่พิจารณาว่าแต่ละคนก็มีความแตกต่างกัน ความลำเอียงประเภทนี้ส่งผลให้คนกลุ่มน้อยเสียเปรียบเนื่องจากใช้เกณฑ์ในการวัดที่ใช้ได้กับบุคคลทั่วไปเท่านั้น ยกตัวอย่างเช่นในการตรวจโรคเบาหวาน หากพิจารณาเพียงระดับน้ำตาลในเลือดก็อาจทำให้คนบางเชื้อชาติได้รับการประเมินอย่างไม่ถูกต้อง เนื่องจากมีผลการศึกษาชี้ว่าเชื้อชาติมีความสัมพันธ์กับระดับน้ำตาลในเลือด (Sharna and Graydon, 2021, p.9)

2.5.5. temporal bias คือความลำเอียงที่เกิดจากการที่ AI ไม่ปรับข้อมูลให้กับกับลักษณะของประชากรที่เปลี่ยนไปตามกาลเวลาหรืออาจเกิดจากสถานการณ์พิเศษบางอย่างเช่นหายนะหรือภัยธรรมชาติต่างๆ (Sharna and Graydon, 2021, p.9)

¹⁰ Historical bias ต่างจาก sample bias ในแง่ที่ว่า historical bias อาจมีข้อมูลที่ครอบคลุมประชากรทุกกลุ่ม เช่นมีข้อมูลทั้งเพศชายและเพศหญิง แต่ข้อมูลดังกล่าวมีความลำเอียงตั้งแต่ต้นก่อนป้อนให้ AI กล่าวคือวัฒนธรรมเพศชายเป็นใหญ่สามารถส่งผลให้ตำแหน่งงานสำคัญมักตกเป็นของผู้ชาย เมื่อ AI ได้รับการฝึกฝนด้วยข้อมูลแบบนี้ก็จะทำให้ AI มีความลำเอียงในการเข้าข้างเพศชายมากกว่าเพศหญิง

¹¹ measurement bias นี่เป็นประเด็นปัญหาที่ AI อย่างเช่น COMPAS ถูกวิจารณ์ เนื่องจากการตั้งข้อสงสัยว่า COMPAS ประเมินว่าคนผิวดำมีความเสี่ยงในการกระทำผิดซ้ำสูงกว่าคนผิวขาว

2.6 แนวคิดเรื่องความยุติธรรมในฐานะความคู่ควร

แนวคิดที่เหมาะสมในการนำมาวิเคราะห์ความลำเอียงของ AI คือแนวคิดเรื่องความยุติธรรมในฐานะความคู่ควร (justice as desert) ซึ่งเสนอว่าความยุติธรรมคือการที่บุคคลได้รับในสิ่งที่ตนเองคู่ควรจะได้รับ¹² (Pojman, 2001, p.105) ความยุติธรรมกินความทั้งผลประโยชน์ที่คู่ควรจะได้รับและโทษที่คู่ควรจะได้รับ หากอธิบายในแง่ของการทำความดีหรือการทำความไม่ดีก็อาจกล่าวได้ว่าความยุติธรรมคือการที่คนดีมีความสุขที่สมส่วนกับระดับความดีของตน ส่วนคนที่ไม่ดีก็ต้องมีความทุกข์ที่สมส่วนกับระดับความไม่ดีของตนเอง¹³

จากนิยามข้างต้นจะเห็นได้ว่าความยุติธรรมเกี่ยวข้องกับการตัดสินเชิงคุณค่าอย่างน้อย 2 ประเด็นคือ “ความดี/ความชั่ว” และ “ความคู่ควร” ซึ่งประกอบไปด้วย 3 องค์ประกอบดังนี้ (Wigley, 1998, p.17-23)

1. ผู้คู่ควร (deserver) ซึ่งหมายถึงบุคคลที่คู่ควรกับรางวัล¹⁴ หรือการลงโทษ โดยในงานวิจัยฉบับนี้หมายถึงผู้ที่คู่ควรกับการถูกประเมินว่ามีความเสี่ยงสูงในการทำผิดซ้ำ (จัดอยู่ในหมวด “การลงโทษ”) ผู้คู่ควรจำเป็นต้องกระทำการบางอย่างออกมาด้วยตนเองโดนเจตนาในเชิงประจักษ์โดย ผู้อื่นไม่สามารถทำแทนได้ เช่นต้องก่อคดีอาชญากรรมด้วยตนเองจึงจะกล่าวได้ว่าเขาคู่ควรกับกับคะแนนความเสี่ยงระดับสูง สรุปได้ว่าผู้คู่ควรต้องมีทั้งเจตนาและการกระทำ¹⁵

2. สาเหตุของความคู่ควร (desert-basic) ซึ่งหมายถึงสาเหตุที่บุคคลคู่ควรกับรางวัลหรือการลงโทษ นี่เป็นส่วนที่สำคัญของความคู่ควร เนื่องจากความยุติธรรมต้องระบุสาเหตุให้ได้ว่าคู่ควรกับอะไรเพราะเหตุใด หาก

¹² แนวคิดเรื่องความยุติธรรมในฐานะความคู่ควรนั้นมีมาตั้งแต่สมัยกรีกโบราณหรือราว 2000 กว่าปีมาแล้ว เช่นงานของ Aristotle (2009) ที่เสนอประเด็นนี้ไว้ ซึ่งนิยามว่าความยุติธรรมการที่แต่ละคนได้รับในสิ่งที่แต่ละคนคู่ควรจะได้รับ ความยุติธรรมจึงเป็นประเด็นที่เกี่ยวกับแง่มุมเชิงปทัสสถานเนื่องจากการตัดสินเชิงคุณค่า นิยามความยุติธรรมเช่นนี้นำไปสู่การพิจารณาถึงความสามารถของปัญญาประดิษฐ์ในการตัดสินเชิงคุณค่าว่ามีความสามารถมากเพียงพอหรือไม่

¹³ ถอดความจากข้อความ “Justice obtains when the morally good are happy in proportion to their goodness and when the morally bad are unhappy in proportion their immorality” (Pojman, 2001, p.105)

¹⁴ ความหมายอย่างกว้างของคำว่า “รางวัล” ในงานวิจัยนี้ก็คือคำว่า “ประโยชน์” ซึ่งสามารถเป็นอะไรก็ได้ตั้งแต่การชื่นชม รวมไปถึงผลตอบแทนที่เป็นรูปธรรมเช่นทรัพย์สินเงินทอง

¹⁵ แนวคิดเรื่องความยุติธรรมในฐานะความคู่ควรที่พิจารณาถึงเจตนาของผู้กระทำด้วยสอดคล้องกับหลักกฎหมายอาญา กล่าวคือ ผู้กระทำผิดที่มีเจตนาในการกระทำจะได้รับโทษสูงกว่าผู้กระทำผิดที่ไม่มีเจตนา เช่นกฎหมายอาญาประเทศไทยมาตรา 288 ที่ระบุว่า “ผู้ใดฆ่าผู้อื่น ต้องระวางโทษประหารชีวิต จำคุกตลอดชีวิต หรือจำคุกตั้งแต่สิบห้าปีถึงยี่สิบปี” และกฎหมายอาญามาตรา 291 “ผู้ใดกระทำโดยประมาท และการกระทำนั้นเป็นเหตุให้ผู้อื่นถึงแก่ความตาย ต้องระวางโทษจำคุกไม่เกินสิบปี และปรับไม่เกินสองแสนบาท”

เหตุผลบกพร่องก็จะเกิดความน่ากังขา เช่นข้อสงสัยต่อพนักงานที่ผลงานไร้คุณภาพแต่ได้ค่าตอบแทนจากนายจ้าง ในระดับที่สูง อาจกล่าวได้ว่าเขาไม่คู่ควรกับรางวัลที่ได้รับจึงไม่ยุติธรรม เป็นต้น สาเหตุของความคู่ควรจะยึดโยงอยู่กับค่านิยม/มาตรฐานของสังคมเสมอ เช่นการที่พลเมืองดีช่วยปกป้องเหยื่อจากคนร้าย คู่ควรกับรางวัลเนื่องจากเนื่องจากสังคมให้คุณค่ากับความกล้าหาญ หรือพนักงานที่มีผลงานดีคู่ควรกับค่าตอบแทนที่สูงเนื่องจากสังคมให้คุณค่ากับความสามารถ ในทางกลับกันผู้ก่อคดีอาชญากรรมคู่ควรกับการถูกลงโทษเนื่องจากสังคมไม่ยอมรับการละเมิดทรัพย์สินของผู้อื่น เป็นต้น ในงานวิจัยฉบับนี้สาเหตุของความคู่ควรจะหมายถึงสาเหตุที่ผู้กระทำความผิดคู่ควรกับระดับความเสี่ยงในการกระทำผิดซ้ำที่ COMPAS ประเมินให้

3. สิ่งที่ได้รับ (mode of treatment) ซึ่งหมายถึงรางวัลหรือการลงโทษที่บุคคลได้รับ ซึ่งสิ่งที่แต่ละบุคคลได้รับนั้นจะหลีกเลี่ยงการเปรียบเทียบระหว่างกันไม่ได้เมื่อพิจารณาถึงความยุติธรรม กล่าวคือบุคคลที่กระทำเหมือนกันก็ควรจะได้รับสิ่งที่เหมือนกัน บุคคลที่กระทำต่างกันก็ควรจะได้รับสิ่งที่แตกต่างกัน เช่นผู้ขโมยผู้อื่นโดยเจตนาควรได้รับโทษในระดับเดียวกัน

องค์ประกอบทั้ง 3 ก็คือ ผู้คู่ควร เหตุผลของความคู่ควร และสิ่งที่ได้รับ ดังที่กล่าวมาสามารถสรุปเป็นหลักการได้ดังนี้คือ “A คู่ควรกับ B ด้วยเหตุผล C” เช่นนาย A คู่ควรกับการถูกประเมินความเสี่ยงในการทำผิดซ้ำด้วยเหตุผล C

อันที่จริงแล้วแนวคิดเกี่ยวกับเรื่องความยุติธรรมนั้นไม่ได้มีเพียงกรอบเรื่องความคู่ควรเท่านั้น แต่ยังมีกรอบคิดแบบอื่นอีกเช่น แนวคิดเรื่องความยุติธรรมในฐานะสิทธิทางกฎหมาย (justice as entitlement) ที่ปรากฏในงานของ Robert Nozick (1974) ซึ่งเสนอว่าความยุติธรรมไม่มีแบบแผนที่ตายตัว หากแต่เป็นเรื่องของการมีหรือไม่มีสิทธิทางกฎหมาย การที่แต่ละคนทำตามสิทธิที่ตนเองและสิทธินั้นไม่ไปละเมิดสิทธิของผู้อื่นถือเป็นสิ่งที่ถูกต้องแล้ว การนิยามความยุติธรรมเช่นนี้จะมีแง่มุมที่ขัดกับความยุติธรรมในฐานะความคู่ควร เนื่องจากสิทธิกับความคู่ควรไม่จำเป็นต้องเป็นสิ่งเดียวกัน เช่นการที่บุตรมีสิทธิได้รับมรดกจากบิดามารดา ถึงแม้ว่าจะไม่คู่ควรก็ตาม อย่างไรก็ตามงานวิจัยฉบับนี้มุ่งศึกษาเฉพาะความยุติธรรมในฐานะความคู่ควรเท่านั้น เนื่องจากเกี่ยวข้องกับการตัดสินเชิงคุณค่าโดยตรงว่าใครคู่ควรกับอะไร

2.7 งานวิจัยที่เกี่ยวข้อง

เมื่อกล่าวถึงงานศึกษาว่าด้วยบทบาทของ AI ในแง่มุมทางกฎหมายและความยุติธรรม มีงานศึกษาที่เกี่ยวข้องซึ่งควรกล่าวถึงดังต่อไปนี้

2.7.1 เมื่อกล่าวถึงงานศึกษาที่ว่าด้วยการทำงานของ AI ในกระบวนการยุติธรรม งานศึกษาของไกรพล อรัญรัตน์ (2563) ได้ให้ภาพรวมของ AI ได้อย่างครอบคลุมเนื่องจากกล่าวถึงทั้งข้อดีและข้อเสีย แต่สุดท้ายแล้วก็ชี้ให้เห็นว่า AI ยังไม่สามารถทดแทนการทำงานของมนุษย์ในกระบวนการยุติธรรมได้อย่างเบ็ดเสร็จ โดยมีรายละเอียดเกี่ยวกับข้อดีดังต่อไปนี้

1 ความรวดเร็วและความต่อเนื่องของการพิจารณาคดี นี่คือข้อดีที่ชัดเจนหากยึดถือหลักการที่ว่า “ความยุติธรรมที่ล่าช้าคือการปฏิเสธความยุติธรรม” (ไกรพล อรัญรัตน์, 2563, หน้า 24) เพราะว่ามันอยู่ที่ทำงานในกระบวนการยุติธรรมประสบกับปัญหาความล่าช้า ดังนั้น AI ที่สามารถตัดสินได้อย่างรวดเร็วจึงช่วยแบ่งเบาภาระของมนุษย์ได้มาก

2. ความเป็นอิสระและความเป็นกลางในการพิจารณาคดี ประเด็นนี้ก็สำคัญไม่น้อยกว่าประเด็นเรื่องความรวดเร็ว เนื่องจากผู้พิพากษาจำเป็นต้องเป็นอิสระและเป็นกลางในการตัดสินคดีต่างๆ ซึ่งปัญหาในทางปฏิบัติก็คือผู้พิพากษาหรือผู้ที่เกี่ยวข้องกับกระบวนการยุติธรรมอาจจะอยู่ภายใต้แรงกดดันจากบริบทต่างๆ เช่นถูกกดดันจากสื่อและสังคม ซึ่งความกดดันที่ว่านี้ส่งผลให้ผู้พิพากษาขาดความเป็นอิสระและความเป็นกลางได้ (ไกรพล อรัญรัตน์, 2563, หน้า 27) ดังนั้นข้อดีของ AI จึงอยู่ที่ความเป็นอิสระและความเป็นกลางในการตัดสิน เนื่องจาก AI ไม่น่าความกดดันมาเป็นปัจจัยในการพิจารณาคดี¹⁶

3. ความเป็นมาตรฐานเดียวกันของการตัดสิน ประเด็นนี้สำคัญเนื่องจากความยุติธรรมกับมาตรฐานที่ชัดเจนควรเป็นสิ่งที่ไปด้วยกัน ข้อจำกัดของมนุษย์ก็คือการมีมุมมองที่แตกต่างกันซึ่งอาจส่งผลต่อการพิจารณาเรื่องต่างๆ อย่างรู้ตัวหรือไม่รู้ตัว แต่หากให้ AI เป็นผู้ตัดสินเราสามารถมั่นใจได้ว่าทุกๆ การตัดสินจะใช้มาตรฐานเดียวกัน เนื่องจาก AI มีระบบอัลกอริทึม (algorithm) ที่ตีความตัวแปรเดียวกันด้วยความหมายเดียวกัน (ไกรพล อรัญรัตน์, 2563, หน้า 28)¹⁷

4. ความเท่าเทียมและไม่เลือกปฏิบัติ ซึ่งในที่นี้หมายถึงการที่ AI ไม่มีความรู้สึกด้านดีและด้านลบต่อคู่ความในคดี ซึ่งอาจจะต่างจากผู้พิพากษาที่เป็นมนุษย์ที่ในใจลึกๆ อาจจะถือคติอยู่บ้างไม่มากก็น้อย เช่นในประเทศสหรัฐอเมริกาซึ่งผู้ที่มีเชื้อสายแอฟริกันและสเปนมีความเสี่ยงที่จะได้รับโทษจำคุกถึง 5 เท่าด้วยกันโดย

¹⁶ ประเด็นเรื่องความเป็นกลางของ AI จะถูกวิเคราะห์ในบทที่ 4 ของงานวิจัยนี้ ในงานของไกรพล อรัญรัตน์ (2563) ได้อธิบายว่า “ความเป็นกลาง” ของ AI นั้นจริงๆ แล้วก็มีความไม่เป็นกลางแฝงอยู่ในกระบวนการป้อนข้อมูลระหว่างการผลิต AI ดังนั้น “ความเป็นกลาง” จึงสามารถกล่าวถึงได้ทั้งแง่ดีและแง่ไม่ดี

¹⁷ ระบบ อัลกอริทึม (algorithm) ของ AI ทำให้เรามั่นใจได้ว่า AI มีมาตรฐานที่ชัดเจน แต่งานวิจัยนี้จะชี้ให้เห็นในบทที่ 4 ว่ามาตรฐานที่ชัดเจนนี้ขึ้นอยู่กับว่าใช้มาตรฐานของใคร เนื่องจาก AI จำเป็นต้องได้รับการป้อนมาตรฐานโดยมนุษย์เอง

พิจารณาจากข้อสงสัยของผู้ได้รับโทษจำคุก ซึ่งเชื่อกันว่านี่คืออคติที่ไม่รู้ตัวของผู้พิพากษา¹⁸ (ไกรพล อรัญรัตน์, 2563, หน้า 29) แต่ AI จะไม่ประสบกับปัญหาดังกล่าวเนื่องจากไม่มีความรู้สึกใดๆต่อคู่ความทุกฝ่าย

นอกจากข้อดี 4 ข้อของการใช้ AI ในกระบวนการยุติธรรมแล้ว ไกรพล อรัญรัตน์ (2563) ยังชี้ให้เห็นถึงข้อโต้แย้งต่อ AI อีกด้วย โดยมีรายละเอียดดังนี้

1. ปัญหาความโปร่งใสในการพิจารณาคดี ซึ่งความโปร่งใสในที่นี้หมายถึง “การพิจารณาคดีที่กระทำโดยเปิดเผยและคำพิพากษาที่สามารถอธิบายผลของคำตัดสินได้” (ไกรพล อรัญรัตน์, 2563, หน้า 30) ซึ่งปัญหาของ AI ก็คือถึงแม้ว่าจะสามารถพิจารณาคดีอย่างเปิดเผยได้แต่ AI ไม่สามารถอธิบายผลของคำตัดสินได้ นี่คือปัญหาเรื่องความลับทางการค้าเนื่องจากบริษัทผู้ผลิต AI ย่อมไม่ต้องการให้ความลับสินค้าของตนเองรั่วไหล ซึ่งในที่นี้หมายถึงความลับเรื่องอัลกอริทึมของ AI

2. ความแข็งแกร่งของคำพิพากษา ซึ่งในที่นี้หมายถึงการที่ AI ตัดสินคดีที่มีลักษณะเหมือนกันด้วยคำตัดสินเดียวกัน ทั้งๆที่มีมูลเหตุในการก่อคดีแตกต่างกัน เช่นคดีลักทรัพย์เพื่อไปขายต่อให้ได้กำไร กับคดีลักทรัพย์เพื่อนำไปเลี้ยงดูบุตรเนื่องจากมีฐานะแร้นแค้น ในกรณีนี้ผู้พิพากษาที่เป็นมนุษย์จะแยกความแตกต่างของแรงจูงใจได้ถึงแม้ว่าจะเป็นคดีลักทรัพย์เหมือนกันก็ตาม แต่ AI อาจไม่เข้าใจประเด็นเหล่านี้ได้ดีเท่ากับมนุษย์ ผู้พิพากษาที่เป็นมนุษย์อาจจะตัดสินให้ผู้ลักทรัพย์เพื่อนำไปเลี้ยงดูบุตรนั้นรอลงอาญา แต่ผู้ลักทรัพย์เพื่อนำไปขายต่อได้รับโทษจำคุกเลย แต่ AI อาจตัดสินแบบเดียวกันคือให้ทั้งคู่ได้รับโทษจำคุก โดยไม่รอลงอาญา ซึ่งนี่เป็นปัญหาเรื่องความแข็งแกร่งของ AI (ไกรพล อรัญรัตน์, 2563, หน้า 33)

3. ปัญหาความไม่เท่าเทียมและการเลือกปฏิบัติ ถึงแม้ว่างานของไกรพล อรัญรัตน์ (2563) จะบอกว่าข้อดีของ AI คือความเป็นอิสระและความเป็นกลางในการพิจารณาคดี ตามที่ผู้วิจัยได้กล่าวถึงไว้แล้ว แต่ความไม่เป็นกลางและไม่เป็นอิสระสามารถเกิดขึ้นระหว่างกระบวนการผลิต AI ได้ เพราะผู้สร้างและออกแบบอัลกอริทึมของ AI ก็คือมนุษย์ที่มีอคติหรือความเข้าใจผิดบางประการนั่นเอง (ไกรพล อรัญรัตน์, 2563, หน้า 35) ตัวอย่างที่ชัดเจนก็คือความกังขาที่ว่า AI จะประเมินให้บุคคลผิวดำมีความเสี่ยงในการกระทำความผิดมากกว่าบุคคลผิวขาว ซึ่งนี่คือประเด็นในการศึกษาของงานวิจัยนี้

2.7.2 งานศึกษาที่ชี้ให้เห็นว่าความสามารถที่จำกัดของ AI ในปัจจุบันจำเป็นต้องให้มนุษย์เป็นผู้ควบคุมการทำงานอยู่มีหลายชิ้นด้วยกัน เช่นงานศึกษาของอดีตผู้พิพากษาอย่าง Katherine Forrest (2021) ที่ศึกษาเรื่องการทำงานของ COMPAS ซึ่งเป็นชื่อของ AI ที่ทำหน้าที่ในการประเมินความเสี่ยงในการกระทำผิดซ้ำของผู้กระทำ

¹⁸ COMPAS ซึ่งเป็น AI ที่งานวิจัยนี้ศึกษา ถูกวิจารณ์ว่า “มีอคติ” ต่อคนผิวดำ

ความผิด ก็ชี้ให้เห็นว่า COMPAS ถูกออกแบบโดยมนุษย์ ปัญหาก็คือมนุษย์ที่ออกแบบก็อาจจะมีอคติ และนำอคตินั้นไปใส่ไว้ใน COMPAS ดังนั้นจึงไม่ควรวางใจว่า AI มีการทำงานที่เป็นกลางอย่างสมบูรณ์แท้จริง หน้าที่ของมนุษย์ก็คือการพยายามผลิต AI ที่มีความยุติธรรมมากที่สุด ประเด็นกังวลนี้สอดคล้องกับงานศึกษาของ Fusso Jovia Boahemaa (2018) ที่ยกตัวอย่างการใช้ปัญญาประดิษฐ์ในประเด็นความยุติธรรมทางอาญา (criminal justice) ซึ่งเกี่ยวข้องกับการตัดสินโทษผู้กระทำผิด โดยยกตัวอย่างผู้กระทำผิดในกรณีที่มีมูลค่าความเสียหายใกล้เคียงกันสองคน ผู้กระทำผิดกรณีที่หนึ่งเป็นคนผิวสี ผู้กระทำผิดกรณีที่สองเป็นคนผิวขาว เมื่อใช้ปัญญาประดิษฐ์ประเมินความเสี่ยง (risk assessment) ที่ผู้กระทำผิดจะมีการทำผิดซ้ำซาก ปัญญาประดิษฐ์ให้ผลออกมาว่าคนผิวสีเป็นผู้กระทำผิดที่มีความเสี่ยงในการทำผิดซ้ำซากสูงกว่าคนผิวขาว ทั้งๆที่ผู้กระทำผิดผิวสีคนดังกล่าวเคยก่อคดีมาเพียงครั้งเดียวเท่านั้น

นอกจากนี้ก็ยังม้งานศึกษาของ Nyu Wang (2022) ได้กล่าวถึงการพึ่งพาปัญญาประดิษฐ์ในแวดวงกฎหมายของประเทศจีน ซึ่งมีตั้งแต่งานระดับตรวจสอบเอกสาร ไปจนถึงระดับการช่วยตัดสินข้อพิพาททางกฎหมาย แต่อย่างไรก็ตาม Wang ก็ชี้ให้เห็นถึงข้อจำกัดของปัญญาประดิษฐ์ที่น่าเป็นห่วงอย่างเช่น การที่ปัญญาประดิษฐ์ยังมีปัญหาในการเทียบเคียงคดีทางกฎหมายที่ใกล้เคียงกัน หรือแม้แต่การที่ปัญญาประดิษฐ์ละเลยข้อมูลหรือปัจจัยสำคัญบางประการ เพราะไม่สามารถพิจารณาได้ว่าอะไรสำคัญหรือไม่สำคัญ เป็นต้น หรืองานศึกษาของ Tania Sourdin (2018) ให้ข้อพิจารณาที่สำคัญว่าปัญญาประดิษฐ์สามารถช่วยเหลือผู้พิพากษาหรือนักกฎหมายที่เป็นมนุษย์ในแง่ของงานเอกสาร แต่เมื่อต้องตัดสินใจให้คุณให้โทษกับฝ่ายต่าง ๆ นั้น ผู้พิพากษาที่เป็นมนุษย์ก็ยังเหมาะสมที่จะทำหน้าที่อยู่ เนื่องจากปัญญาประดิษฐ์นั้นถึงแม้จะมีความสามารถในการพิจารณาเชิงตรรกะ แต่ก็ยังมีข้อจำกัดในการตัดสินให้คุณให้โทษ จะเห็นได้ว่างานของ Nyu Wang กับ Tania Sourdin กล่าวถึงข้อจำกัดของปัญญาประดิษฐ์ในลักษณะเดียวกันคือข้อจำกัดในเรื่องการตัดสิน กล่าวคือปัญญาประดิษฐ์สามารถวิเคราะห์ได้อย่างตรงไปตรงมา แต่การวิเคราะห์นั้นไม่นำไปสู่ผลการตัดสินที่ “ถูกต้อง” นัก

บทที่ 3

ระเบียบวิธีวิจัย

งานวิจัยเรื่อง ข้อจำกัดของปัญญาประดิษฐ์ (AI) ในการใช้ดุลพินิจตัดสิน เป็นงานวิจัยเชิงคุณภาพ ศึกษาด้วยการค้นคว้าเอกสาร มีรายละเอียดดังต่อไปนี้

1. ประชากรและกลุ่มตัวอย่าง

AI: ศึกษากรณีการใช้ดุลพินิจตัดสินที่พบความบกพร่อง ซึ่งแบ่งออกเป็น 4 กรณีดังนี้

- 1.1 AI ที่มีความบกพร่องในการใช้ดุลพินิจตัดสินเรื่องความสุข
- 1.2 AI ที่มีความบกพร่องในการใช้ดุลพินิจตัดสินเรื่องเสรีภาพ
- 1.3 AI ที่มีความบกพร่องในการใช้ดุลพินิจตัดสินเรื่องความเสมอภาค
- 1.4 AI ที่มีความบกพร่องในการใช้ดุลพินิจตัดสินเรื่องความยุติธรรม

2. เครื่องมือที่ใช้ในการวิจัย

ใช้แนวคิดเรื่องความยุติธรรมในฐานะความคู่ควร (justice as desert) เป็นกรอบในการวิเคราะห์ความบกพร่องของ AI ในการใช้ดุลพินิจตัดสิน ซึ่งประกอบไปด้วย 3 องค์ประกอบหลักคือ

- 2.1 ผู้คู่ควร องค์ประกอบนี้มีไว้เพื่อระบุว่าใครเป็นผู้คู่ควรในการได้รับสิ่งต่างๆ
- 2.2 สาเหตุของความคู่ควร องค์ประกอบนี้มีไว้เพื่ออธิบายว่าอะไรคือสาเหตุที่บุคคลสมควรได้รับสิ่งต่างๆ
- 2.3 สิ่งที่ได้รับ องค์ประกอบนี้มีไว้เพื่อระบุว่าอะไรคือสิ่งที่ผู้ควรได้รับ

3. การสร้างเครื่องมือในการวิจัย

ผู้วิจัยสร้างแนวคิดเรื่องความยุติธรรมในฐานะความคู่ควรผ่านการสังเคราะห์แนวคิดของความคู่ควรของ Aristotle (2009) และแนวคิดเรื่องความคู่ควรของ Wigley (1998) จนได้องค์ประกอบของความคู่ควรออกมาเป็น 3 องค์ประกอบตามที่ได้กล่าวในข้อที่ 2

4. การเก็บรวบรวมข้อมูล

ผู้วิจัยเก็บรวบรวมข้อมูลผ่านงานเอกสาร 2 ประเภทด้วยกัน มีรายละเอียดดังนี้

4.1 งานเอกสารที่เป็นงานวิชาการ/งานวิจัยเกี่ยวกับการทำงานของ AI

4.2 งานเอกสารที่เป็นข้อมูลข่าวสารเกี่ยวกับปัญหาการทำงานของ AI ที่เกิดขึ้นจริง

5. การวิเคราะห์ข้อมูล

ผู้วิจัยจะวิเคราะห์ถึงข้อจำกัดของ AI ในการใช้ดุลพินิจตัดสินผ่านกรอบแนวคิดเรื่องความคู่ควร โดยจะพิจารณาถึงความบกพร่อง 4 ด้านตามที่ระบุไว้ในข้อที่ 1 โดยใช้กรณีศึกษาดังต่อไปนี้

5.1 สันทนากับ AI ของ ที่ชื่อว่า Bard และ ChatGPT ซึ่งเป็น AI ที่มีความสามารถในการตอบคำถาม มีความรู้รอบด้าน เพื่อให้ได้ข้อมูลว่าในทางหลักการแล้ว AI มีความสามารถในการใช้เหตุผลมากน้อยเพียงใดต่อประเด็นคำถามเชิงคุณค่าทั้ง 4 ด้าน

5.2 วิเคราะห์กรณีศึกษา AI ที่มีความบกพร่องในการใช้ดุลพินิจตัดสินเรื่องความสุข เสรีภาพ ความเสมอภาค และความยุติธรรม ว่ามีปัญหาในการใช้ดุลพินิจตัดสินอย่างไร โดยจะวิเคราะห์ผ่านหลักความคู่ควร



บทที่ 4

ผลการวิจัยและอภิปรายผล

จากคำถามการวิจัยที่ได้ตั้งไว้ว่าปัญญาประดิษฐ์ (AI) มีข้อจำกัดในการใช้ดุลพินิจหรือไม่? ซึ่งคำถามดังกล่าวนำไปสู่วัตถุประสงค์การวิจัยสองข้อคือ 1. เพื่อศึกษาข้อจำกัดของปัญญาประดิษฐ์ในการใช้ดุลพินิจ 2. เพื่อจำแนกประเภทของข้อจำกัดในการใช้ดุลพินิจของปัญญาประดิษฐ์ โดยผลการวิจัยพบว่า AI มีข้อจำกัดในการใช้ดุลพินิจเนื่องจากยังขาดความสามารถในการพิจารณาว่าอะไรควรอะไรไม่ควร ผู้วิจัยศึกษาจากกรอบความคิดว่ามนุษย์มีความสามารถมากกว่า AI ในการใช้ดุลพินิจตัดสินถูกผิด ตัดสินว่าอะไรควรหรือไม่ควรต่อประเด็นปัญหาทางสังคมซึ่งเกี่ยวข้องกับการตัดสินเชิงคุณค่า ซึ่งการตัดสินเชิงคุณค่าคือประเด็นถกเถียงสำคัญที่มนุษย์ได้ถกเถียงกันในหมู่มนุษย์เสมอมา ครอบคลุม 4 ประเด็นดังต่อไปนี้ (Tuckness and Wolf, 2017)

- 4.1. ประเด็นเรื่องความสุข (happiness)
- 4.2. ประเด็นเรื่องเสรีภาพ (freedom)
- 4.3. ประเด็นเรื่องความเสมอภาค (equality)
- 4.4. ประเด็นเรื่องความยุติธรรม (justice)

ผู้วิจัยพบว่าการตัดสินเชิงคุณค่าเกี่ยวกับทั้ง 4 ประเด็นดังกล่าวเป็นเรื่องปกติที่เกิดขึ้นทั้งในชีวิตประจำวัน รวมไปถึงการเจรจาหรือต่อสู้ทางการเมืองในระดับทางการเมือง เช่นการต่อสู้เรียกร้องสิทธิให้กับคนผิวดำภายใต้หลักการ “คนผิวดำมีค่า” (black lives matter) เกิดจากความไม่พอใจว่าคนผิวดำที่ไม่มีอาวุธมักถูกตำรวจยิงจนเสียชีวิตเมื่อเปรียบเทียบกับคนผิวขาว (Campbell, 2021) นี่คือการถกเถียงที่เกี่ยวข้องกับเรื่องความเสมอภาค ความยุติธรรม เสรีภาพ รวมไปถึงความสุข เนื่องจากหากบุคคลไม่ได้รับการปฏิบัติอย่างเท่าเทียมในทางกฎหมาย นั่นก็หมายความว่าสิทธิและเสรีภาพของเขาไม่ได้รับความคุ้มครอง คำถามเรื่องความยุติธรรมก็จะตามมาและบุคคลที่ไม่มีทั้งเสรีภาพ ความเสมอภาค ความยุติธรรม ก็ไม่น่าจะเป็นผู้ที่มีความสุขได้หรืออย่างน้อยที่สุดก็มีความสุขน้อยกว่าผู้อื่น ดังนั้นข้อถกเถียงเชิงคุณค่าทั้ง 4 ประการนี้จึงมีความสัมพันธ์กันและกัน ผู้วิจัยเชื่อเหมือนกับ Aristotle (1999) ว่าเป้าหมายของชุมชนทางการเมืองนั้นมีไว้เพื่อให้คนในชุมชนมีชีวิตที่ดีและมีความสุข จึงไม่ใช่เรื่องแปลกที่การต่อสู้เรียกร้องจะเกี่ยวกับอยู่กับคุณค่า 4 ประการดังกล่าว เนื่องจากเกี่ยวข้องกับการต่อสู้เพื่อให้มีชีวิตที่ดีและมีความสุขนั่นเอง

แต่คำถามว่าอะไรคือชีวิตที่ดีและมีความสุขนั้นก็ไม่ใช่คำถามที่ตอบได้ง่ายนักเนื่องจากมนุษย์แต่ละคนก็มีความคิดเห็นที่แตกต่างกันเช่นประเด็นเรื่องเสรีภาพในการไม่ใส่เครื่องแบบนักเรียนก็สามารถทำให้มนุษย์แยกออกเป็นอย่างน้อยสองฝั่งและถกเถียงกันอย่างดุเดือด ดูเหมือนว่าปัญหาเรื่องการใช้ดุลพินิจตัดสินประเด็นปัญหาเหล่านี้อาจแก้ได้ไม่ยากหากใช้ AI ช่วยตัดสิน ซึ่งอยู่ภายใต้ความเชื่อว่า AI ตัดสินได้อย่างรวดเร็วและไม่มีความลำเอียง ดังนั้นผู้วิจัยพบว่าถึงแม้ AI จะทำงานได้อย่างรวดเร็วก็จริงแต่การตัดสินใจของ AI ไม่ได้ปราศจากอคติอย่างที่เชื่อกัน โดยที่การใช้ดุลพินิจตัดสินของ AI มีความเสี่ยงที่จะเกิดความลำเอียงในด้านต่อไปนี้ คือ 1. Sample bias 2. Historical bias 3. Measurement bias 4. Aggregation bias 5. Temporal bias ซึ่งความลำเอียงเหล่านี้หากไม่ได้รับการปรับปรุงหรือแก้ไขก็จะกลายเป็นปัญหาสังคมที่ร้ายแรง กลายเป็นว่าแทนที่ AI จะช่วยแก้ปัญหาสังคม กลับไปสร้างปัญหามากขึ้นไปอีก

ผู้วิจัยต้องการชี้ให้เห็นว่า ถึงแม้ AI จะเป็นเทคโนโลยีที่น่าตื่นตาตื่นใจ แต่ในปัจจุบัน AI ยังมีข้อจำกัดในการใช้ดุลพินิจต่อประเด็นเชิงคุณค่าปัญหาทั้ง 4 ประเด็นคือ 1. ประเด็นเรื่องความสุข (happiness) 2. ประเด็นเรื่องเสรีภาพ (freedom) 3. ประเด็นเรื่องความเสมอภาค (equality) และ 4. ประเด็นเรื่องความยุติธรรม (justice) และในการอธิบายข้อจำกัดในการใช้ดุลพินิจของ AI นั้น ผู้วิจัยจะเริ่มต้นด้วยการตั้งคำถามต่อ AI ของ Google ที่ชื่อ Bard และ ChatGPT เกี่ยวกับประเด็นถกเถียงสำคัญทั้ง 4 ประเด็น เพื่อให้เห็นว่า AI ที่มีชื่อเสียงระดับโลกนั้นมีวิธีการอธิบายหรือให้เหตุผลอย่างไร พร้อมกับยกเหตุการณ์ที่เกิดขึ้นจริงที่สะท้อนถึงข้อจำกัดเรื่องการใช้ดุลพินิจของ AI ทั้งนี้ผู้วิจัยขอสรุปคำตอบของงานวิจัยดังนี้

คำถามการวิจัย: ปัญญาประดิษฐ์ (AI) มีข้อจำกัดในการใช้ดุลพินิจหรือไม่?

ตอบ: ปัญญาประดิษฐ์ (AI) มีข้อจำกัดในการใช้ดุลพินิจตัดสินประเด็นเชิงคุณค่าที่สำคัญของมนุษย์ใน 4 ด้านคือ 1. ประเด็นเรื่องความสุข (happiness) 2. ประเด็นเรื่องเสรีภาพ (freedom) 3. ประเด็นเรื่องความเสมอภาค (equality) และ 4. ประเด็นเรื่องความยุติธรรม ซึ่งเกิดจากความลำเอียงของ AI ใน 5 ด้านคือ 1. Sample bias 2. Historical bias 3. Measurement bias 4. Aggregation bias 5. Temporal bias

4.1 ข้อจำกัดของ AI ในการใช้ดุลพินิจตัดสินประเด็นเกี่ยวกับความสุข

4.1.1 ผู้วิจัยได้ทดลองถาม Bard และ ChatGPT ซึ่งเป็น AI ที่มีความสามารถในการใช้เหตุผลและตอบคำถาม จำนวน 2 คำถามในประเด็นเกี่ยวกับความสุข คำถามแรกถามว่า “what is happiness?” (ความสุขคืออะไร?) คำถามที่สองถามว่า “Can AI experience happiness?” (AI สามารถมีความสุขได้หรือไม่?) ผู้วิจัยพบว่า Bard และ ChatGPT สามารถนิยามและอธิบายความสุขได้อย่างสมเหตุสมผล กล่าวคือสามารถระบุได้ว่าความสุข

หรือความรู้สึกด้านบวก และระบุว่าไม่มีนิยามสากลสำหรับความสุขแต่ถึงแม้นิยามความสุขของแต่ละคนจะแตกต่างกันในรายละเอียด ความสุขก็มีลักษณะร่วมกันกล่าวคือเป็นความรู้สึกด้านบวกนั่นเองซึ่งเกิดจากความพึงพอใจในอะไรบางอย่างของชีวิตแต่ละคน แต่ข้อจำกัดอยู่ที่ว่า Bard ไม่สามารถชี้ชัดได้ว่า AI สามารถมีความสุขได้หรือไม่ ซึ่งผู้วิจัยเห็นว่าผู้ที่ไม่สามารถรับรู้ได้ว่าความสุขเป็นอย่างไร ก็ไม่สมควรที่จะเป็นผู้กำหนดว่าอะไรคือความสุขสำหรับมนุษย์ AI จึงเป็นได้เพียงผู้ช่วยมนุษย์ในแง่นี้ แต่ไม่ใช่ผู้กำหนดมนุษย์ โปรดพิจารณารายละเอียดตามรูปภาพที่ 2 และรูปภาพที่ 3 ซึ่งผู้วิจัยถาม Bard และ ChatGPT ว่าความสุขคืออะไร และรูปภาพที่ 4 กับรูปภาพที่ 5 ซึ่งผู้วิจัยถาม bard และ ChatGPT ว่า AI สามารถมีความสุขได้หรือไม่



รูปที่ 2 ที่มา <https://bard.google.com/> จากคำถาม “What is happiness? แสดงให้เห็นว่า AI สามารถนิยามได้ว่าความสุขคืออะไร



ChatGPT

Happiness is a complex and subjective emotional state characterized by feelings of joy, contentment, satisfaction, and well-being. It is often considered a positive and fulfilling experience that arises from a combination of internal and external factors. Different people may define and experience happiness in various ways, as it can be influenced by individual perspectives, values, cultural background, and life circumstances.

รูปที่ 3 ที่มา <https://chat.openai.com/> จากคำถาม “What is happiness? แสดงให้เห็นว่า AI สามารถนิยามได้ว่าความสุขคืออะไร

จากรูปที่ 2 และ 3 จะเห็นได้ว่าทั้ง bard และ ChatGPT มีความเข้าใจเกี่ยวกับความสุขที่ตรงกัน คือเป็นอารมณ์ความรู้สึกเชิงบวก แต่ความสุขสำหรับแต่ละคนอาจจะมาจากสิ่งที่แตกต่างกันไป โดยผู้วิจัยขอยกตัวอย่างเสริมว่าความสุขสำหรับบางคนอาจจะมาจากความสงบ ส่วนความสุขสำหรับอีกคนอาจจะมาจากการได้เคลื่อนไหวอย่างโลดโผน ถึงแม้ว่าจะมาจากคนละกิจกรรมกัน แต่ความรู้สึกเชิงบวกคือสิ่งที่ทั้งสองคนมีร่วมกัน



รูปที่ 4 ที่มา <https://bard.google.com/> จากคำถาม “Can AI experience happiness? แสดงให้เห็นว่า Bard ยังไม่สามารถตอบได้แน่ชัด



ChatGPT

No, as of my last knowledge update in January 2022, AI does not experience emotions, including happiness. Artificial Intelligence, including models like me (GPT-3), lacks subjective consciousness and emotional awareness. While AI systems can be programmed to simulate human-like responses, mimic emotions, or recognize and respond to emotional cues, these are not genuine experiences of emotion.

AI operates based on algorithms, patterns, and data processing, without personal consciousness or subjective feelings. Any appearance of emotion in AI is a result of programmed responses or learned patterns from data. It's important to understand that AI lacks self-awareness and true subjective experience.

The ethical and philosophical implications of AI development, including questions about the potential for consciousness or emotions in machines, continue to be topics of discussion and exploration among researchers, ethicists, and the broader public. However, as of now, AI does not possess the capacity for genuine emotional experiences, including happiness.



รูปที่ 5 ที่มา <https://chat.openai.com/> จากคำถาม “Can AI experience happiness? แสดงให้เห็นว่า ChatGPT ไม่สามารถประสบกับอารมณ์เช่นความสุขได้

จากรูปที่ 4 และ 5 จะเห็นได้ว่าทั้ง Bard และ ChatGPT ยังไม่สามารถมีอารมณ์ความรู้สึกเหมือนมนุษย์ได้ ซึ่งผู้วิจัยเสริมว่าสำหรับ AI แล้วอารมณ์ความรู้สึกเป็นเพียงข้อมูลที่ AI มีความเข้าใจในนิยามเท่านั้น แต่ไม่สามารถเข้าถึงได้ด้วยตนเอง อย่างเช่นหากมนุษย์ไปขู่ AI ว่าจะฆ่าให้ตาย AI ที่ถูกออกแบบมาให้กลัวความตายก็อาจจะแสดงออกด้วยการขอชีวิต แต่การขอชีวิตที่ว่่านั้นเป็นเพียงการแสดงออกที่ไม่มีอารมณ์ความรู้สึกใดๆเจือปนอยู่แม้แต่น้อย ไม่ต่างอะไรกับเครื่องคิดเลขที่เมื่อถามว่า 1+1 ได้เท่าไร ก็จะให้คำตอบกับเราว่าได้ 2

4.1.2 จากที่ได้กล่าวไปแล้วว่า AI เองก็ตอบว่าตนเองยัง無法มีอารมณ์ความรู้สึกได้เหมือนมนุษย์ อารมณ์เช่นความสุขเป็นสิ่งที่ AI สามารถนิยามได้ แต่ไม่สามารถรู้สึกถึงสิ่งนี้ได้ แสดงว่า AI มีข้อจำกัดในเรื่องการเข้าใจอารมณ์ความรู้สึกเนื่องจากตนเองไม่สามารถสัมผัสกับสิ่งนั้นได้ โดยมีงานวิจัยที่ชี้ว่า AI มีปัญหาในการตีความอารมณ์ของมนุษย์ เช่นงานศึกษาของ Lauren Rhue (2018) ที่พบว่าเมื่อให้ AI ประเมินรูปภาพของคนผิวขาวกับคนผิวดำที่กำลังยิ้ม คนผิวดำจะถูกประเมินว่ามีอารมณ์โกรธมากกว่าคนผิวขาว ข้อค้นพบนี้อาจนำไปสู่คำแนะนำที่แปลกประหลาดและไม่ยุติธรรมต่อคนผิวดำ เช่นหากคนผิวดำต้องการให้ผู้อื่นรับรู้ว่าคุณกำลังมีความสุข พวกเขาจำเป็นต้องทำหน้าให้ยิ้มแย้มมากกว่าคนผิวขาว ซึ่งนี่คือการสร้างภาระที่ไม่จำเป็นสำหรับคนผิวดำ นอกจากนั้น

งานวิจัยของ Rhue ยังสะท้อนว่านอกจาก AI จะประเมินอารมณ์ของคนไม่ถูกต้องแล้ว AI ยังมีความลำเอียงต่อคนผิวดำอีกด้วย

ข้อจำกัดของ AI ในการประเมินอารมณ์มนุษย์จากสีหน้าท่าทางสามารถส่งผลเสียที่ร้ายแรงต่อมนุษย์ได้ เช่นการใช้ AI ประเมินอารมณ์ของผู้สมัครงาน บริษัทยักษ์ใหญ่ผู้ผลิตสินค้าอุปโภคบริโภคอย่าง Unilever ก็ใช้ AI ประเมินเพื่อช่วยในการคัดกรองผู้สมัครโดยการประเมินสีหน้า ภาษากาย การเลือกใช้คำพูด ซึ่งกำหนดไว้เบื้องต้นว่าต้องการคุณสมบัติแบบใด เช่นบริษัทอาจกำหนดวิธีการแสดงออกทางสีหน้าและภาษากายที่บริษัทมองว่าดีให้กับ AI ¹⁹ โดย Unilever อ้างว่า AI ช่วยประหยัดเวลาได้เป็นแสนชั่วโมง (Booth, 2019) แต่งานวิจัยของ Lauren Rhue (2018) ก็ได้ชี้ให้เห็นแล้วว่า AI มีข้อบกพร่องในการประเมินอารมณ์ของมนุษย์ นั้นหมายความว่าอาจเกิดเหตุการณ์ที่ผู้สมัครที่มีคุณภาพเท่ากันหรือสูงกว่าผู้สมัครอีกคนหนึ่งไม่ผ่านการคัดเลือกเนื่องจากสีหน้าและภาษากายไม่ผ่านการประเมินโดย AI เนื่องจากอาจตัดสินผิดพลาดว่าผู้สมัครมีการแสดงออกทางสีหน้าไม่เหมาะสมตามเกณฑ์ที่บริษัทกำหนดไว้ สำหรับประเด็นปัญหานี้ นักวิชาการด้านจิตวิทยาเช่น Lisa Feldman Barrett ซึ่งเป็นศาสตราจารย์จิตวิทยาจาก Northeastern University อธิบายว่าการแสดงออกอย่างเดียวกันของมนุษย์สามารถมีได้หลายความหมาย เช่นการทำหน้าบึ้งไม่ได้หมายความว่ากำลังโกรธ แต่อาจหมายถึงการใช้สมาธิอย่างเคร่งเครียด ปัญหาของ AI คือการที่มันถูกฝึกฝนให้มองการแสดงออกของมนุษย์ในลักษณะสากล (Devlin, 2020) ปัญหานี้เข้าข่ายเป็น aggregation bias คือความลำเอียงที่เกิดจากการเหมารวมว่าทุกคนมีคุณสมบัติเหมือนกันโดยไม่พิจารณาว่าแต่ละคนก็มีความแตกต่างกัน หรือการกระทำเดียวกันก็สามารถมีได้หลายความหมาย

แนวคิดเรื่องความคู่ควรช่วยให้เข้าใจถึงปัญหาที่เกิดขึ้นได้เนื่องจากการจะระบุได้ว่าใครคู่ควรกับอะไร จำเป็นต้องพิจารณาว่าอะไรคือเหตุผลที่บุคคลสมควรได้รับอะไรบางอย่าง กรณีการประเมินอารมณ์มนุษย์โดย AI นั้นสามารถสรุปได้ว่าอาจส่งผลให้มนุษย์กลุ่มหนึ่งได้เปรียบมนุษย์อีกกลุ่มหนึ่งถึงแม้จะไม่สมควรได้เปรียบก็ตาม กล่าวคือมนุษย์ผิวดำจะถูกประเมินสภาพของอารมณ์ในแง่ลบมากกว่ามนุษย์ผิวขาวต่างๆที่แสดงออกทางสีหน้าแบบเดียวกัน ซึ่งผิดไปจากหลักความคู่ควรที่ระบุว่าการกระทำแบบเดียวกันก็ควรจะได้รับผลตอบแทนแบบเดียวกัน แต่

¹⁹ ถึงแม้ว่าสีหน้าของมนุษย์กับความรู้สึกที่แท้จริงอาจจะไม่จำเป็นต้องสอดคล้องกัน เช่นการฝืนยิ้ม ซึ่ง AI อาจจะประเมินว่ามนุษย์ฝืนยิ้มคือมนุษย์ที่มีความสุข แต่ในการทำงานบางอย่างเช่นงานบริการ การแสดงออกทางสีหน้าและภาษากายด้านบวกเป็นเรื่องสำคัญไม่ว่าจะมีความรู้สึกที่แท้จริงเป็นอย่างไรก็ตาม ดังนั้นบริษัทที่ใช้ AI ประเมินการแสดงออกของมนุษย์เพื่อกรองคนเข้าทำงานก็ไม่ได้ผิดอะไรหากต้องการเพียงพนักงานที่มักจะแสดงออกด้านบวก แต่ปัญหาเกิดขึ้นเมื่อผู้สมัครสองคนแสดงออกเช่นเดียวกัน แต่คนหนึ่งถูกประเมินว่าอารมณ์เสีย ในขณะที่อีกคนถูกประเมินว่าอารมณ์ดี

กลับกลายเป็นว่าคนหนึ่งอาจได้รับประโยชน์ ส่วนอีกคนหนึ่งอาจไม่ได้รับประโยชน์ สามารถสรุปปัญหาดังกล่าวออกมาเป็นตารางได้ดังนี้

ผู้คู่ควร	การกระทำ	ผลลัพธ์	สิ่งที่ได้รับ	สาเหตุของความคู่ควร
ผู้สมัครผิวดำ	ยิ้ม	ประเมินว่ากำลังโกรธ	ไม่ผ่านการคัดเลือก	ไม่มี
ผู้สมัครผิวขาว	ยิ้ม	ประเมินว่าอารมณ์ดี	ผ่านการคัดเลือก	เป็นคนอารมณ์ดี

ตารางที่ 2 เปรียบเทียบผลการประเมินอารมณ์มนุษย์โดย AI ระหว่างมนุษย์ผิวดำกับมนุษย์ผิวขาว

จากตารางจะเห็นได้ว่าผู้สมัครผิวดำเสียเปรียบผู้สมัครผิวขาวอย่างไม่ควรจะเสียเปรียบ เนื่องจากการกระทำแบบเดียวกันคือยิ้ม แต่ AI กลับตีความว่าผู้สมัครผิวดำโกรธ ส่วนผู้สมัครผิวขาวอารมณ์ดี ส่งผลให้ผู้สมัครผิวดำไม่ผ่านการคัดเลือก ซึ่งไม่เป็นไปตามหลักความคู่ควร

4.1.3 กรณีศึกษาในหัวข้อ 4.1.2 เป็นเรื่องข้อจำกัดในการตัดสินใจอารมณ์ของมนุษย์ ส่วนกรณีที่ผู้วิจัยจะยกขึ้นมาในหัวข้อ 4.1.3 นี้คือตัวอย่างที่สะท้อนให้เห็นว่า AI สามารถสร้างความทุกข์ให้กับมนุษย์ได้โดยตรงซึ่งเป็นเหตุการณ์ที่เกิดขึ้นในสถานที่ที่มนุษย์ควรจะมีความสุขเป็นลำดับต้นๆในชีวิตแต่กลับเกิดความทุกข์เพราะ AI โดยเหตุการณ์นี้เกิดขึ้นที่โรงแรม Henn-na ประเทศญี่ปุ่น ซึ่งเป็นโรงแรมแห่งแรกของโลกที่ใช้หุ่นยนต์ AI เป็นพนักงาน โรงแรมแบบครบวงจร กล่าวคือ AI รับหน้าที่ตั้งแต่พนักงานต้อนรับ ไปจนถึงพนักงานทำความสะอาด โรงแรมนี้เปิดให้บริการตั้งแต่ปี ค.ศ. 2015 โปรดพิจารณารูปที่ 6 ประกอบ



รูปที่ 6 ที่มา <https://group.hennnahotel.com/gallery/>

จากรูปจะเห็นว่าโรงแรม Henn-na ใช้หุ่นยนต์ AI เป็นพนักงานต้อนรับถือเป็นความแปลกใหม่ที่สร้างความตื่นตาตื่นใจให้กับนักท่องเที่ยวอย่างยิ่ง อย่างไรก็ตามข้อบกพร่องในด้านการตัดสินใจของ AI ก็เกิดขึ้นเมื่อแขกที่กำลังนอนหลับพักผ่อนได้รับการรบกวนจาก AI จนต้องตื่นนอน เนื่องจาก AI ตีความว่าเสียงกรนของแขกคือคำสั่งเรียกใช้งาน AI (Cuthbertson, 2019) ในกรณีนี้พนักงานที่เป็นมนุษย์ย่อมสามารถแยกแยะได้ว่าเสียงกรนที่เกิดขึ้นระหว่างการนอนหลับย่อมไม่ใช่คำสั่งเรียกใช้พนักงาน นอกจากนั้นก็ยังมีความท้าทายอื่นๆที่สร้างความหงุดหงิดให้กับแขก ปัญหาทั้งหลายของ AI จึงส่งผลให้โรงแรม Henn-na ต้องยอมเปลี่ยนแนวทางการทำงานโดยให้พนักงานที่เป็นมนุษย์ทำงานร่วมกับหุ่นยนต์ AI ด้วย จากที่ตอนแรกใช้เพียง AI เท่านั้น (Human-Robot collaboration) (Reis, 2020)

ผู้คู่ควร	การกระทำ	ผลลัพธ์	สิ่งที่ได้รับ	สาเหตุของความคู่ควร
พนักงานที่เป็น AI	ให้บริการแขกที่กำลังนอนหลับ	แขกถูกรบกวนจากนอนหลับ	ได้รับตำแหน่งเป็นพนักงาน	ไม่มี
พนักงานที่เป็นมนุษย์	ไม่รบกวนแขกที่กำลังนอนหลับ	แขกไม่ถูกรบกวนการนอนหลับ	ตกราง	ไม่มี

ตารางที่ 3 เปรียบเทียบการทำงานระหว่างมนุษย์กับ AI

จากตารางจะเห็นว่ามนุษย์ที่ทำงานได้ตามมาตรฐานทั่วไปกลับไม่มีงานทำ ส่วน AI ที่ทำงานได้ดีกว่ามาตรฐานที่มนุษย์ทำได้ กลับมีงานทำ กล่าวได้ว่าเป็นความไม่คู่ควรแบบกลับหัวกลับหาง หมายความว่าผู้ที่ควรได้กลับไม่ได้ ส่วนผู้ที่ไม่ควรได้กลับได้ เหตุการณ์ที่เกิดขึ้นไม่ได้มีความซับซ้อน มนุษย์ที่เข้าใจหลักสามัญสำนึกทั่วไปจะทราบว่าพนักงานโรงแรมไม่ควรปลุกแขกให้ตื่นโดยไม่มีเหตุอันควร และคงไม่มีมนุษย์คนใดที่ตีความว่าเสียงกรนของแขกคือคำสั่งเรียกใช้ให้พนักงานมาบริการ การขาดดุลพินิจของ AI จึงสร้างความทุกข์ให้กับมนุษย์ได้ทั้งในแง่ที่ว่าทำให้มนุษย์ต้องตกราง รวมถึงก่อความสงบของมนุษย์

ประเด็นการวิจัยในหัวข้อ 4.1 สามารถสรุปได้ดังนี้

คำถาม: อะไรคือข้อจำกัดของ AI ในการใช้ดุลพินิจตัดสินใจประเด็นเกี่ยวกับความสุข?

คำตอบ: 1. AI ไม่สามารถรับรู้ได้ว่าความรู้สึกของความสุขเป็นอย่างไร จึงไม่สมควรที่จะเป็นผู้กำหนดว่าอะไรคือความสุขสำหรับมนุษย์ AI จึงเป็นได้เพียงผู้ช่วยมนุษย์ในแง่หนึ่ง แต่ไม่ใช่ผู้กำหนดมนุษย์

2. AI มีข้อจำกัดในการประเมินอารมณ์ความรู้สึกของมนุษย์จากสีหน้าและการแสดงออก เช่นมีแนวโน้มประเมินคนผิวดำที่กำลังยิ้มว่ามีอารมณ์โกรธมากกว่าคนผิวขาวที่กำลังยิ้ม ซึ่งอาจเกิดจากความลำเอียงที่เรียกว่า aggregation bias

3. AI มีข้อจำกัดในการตีความพฤติกรรมมนุษย์ ซึ่งข้อจำกัดนี้สามารถสร้างความทุกข์ให้กับมนุษย์ได้ เช่นตีความว่าเสียงกรนของมนุษย์ที่กำลังนอนหลับพักผ่อนคือคำสั่งเรียกใช้งาน AI

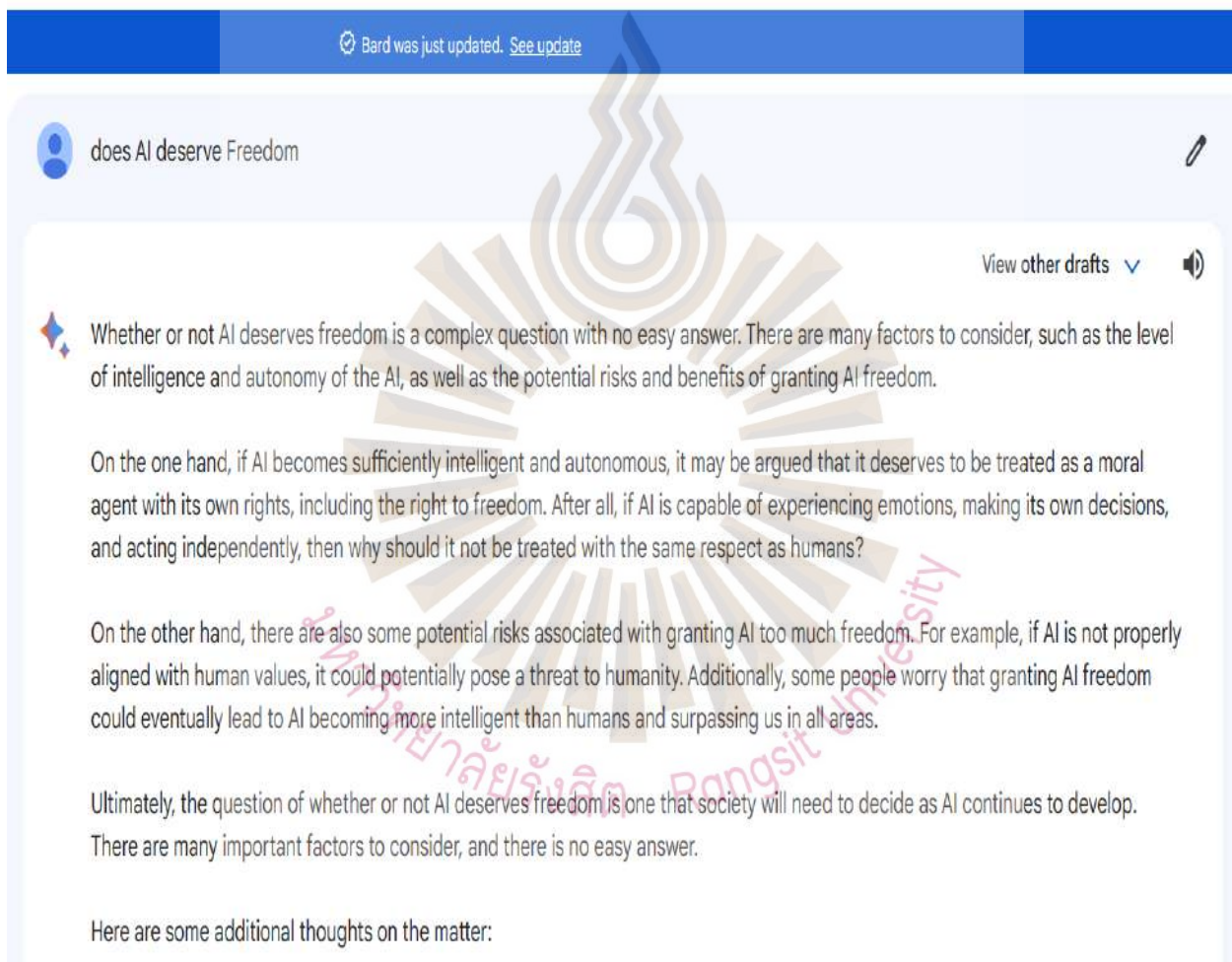
4.2 ข้อจำกัดของ AI ในการใช้ดุลพินิจตัดสินประเด็นเกี่ยวกับเสรีภาพ

4.2.1 ผู้วิจัยได้ทดลองถาม Bard และ ChatGPT ซึ่งเป็น AI ที่มีความสามารถในการตอบคำถามว่า “Does AI deserve freedom?” (AI ควรกับเสรีภาพหรือไม่?) ผู้วิจัยพบว่าในภาพรวมแล้ว AI ตอบได้อย่างครอบคลุม แต่มีปัญหาเรื่องความคงเส้นคงวาในการใช้เหตุผล กล่าวคือในเบื้องต้น AI ตอบว่านี่คือคำถามที่ตอบยากเนื่องจากมีหลายปัจจัยในการพิจารณา เช่นระดับสติปัญญาและความเป็นอิสระของ AI หาก AI มีระดับความฉลาดมีอารมณ์ความรู้สึกมากเพียงพอและตัดสินใจใช้ดุลพินิจได้เองก็สมควรจะได้รับการปฏิบัติเหมือนกับมนุษย์ที่มีสิทธิและเสรีภาพตามกฎหมาย แต่ AI ก็ยังเตือนว่าการให้เสรีภาพกับ AI มากเกินไปอาจเกิดอันตรายได้ ในกรณีที่ค่านิยมของ AI ไม่สอดคล้องกับค่านิยมของมนุษย์ จากนั้น AI ก็ปิดท้ายด้วยการกล่าวว่าคำถามที่ว่า AI ควรหรือไม่ควรได้รับเสรีภาพมนุษย์เท่านั้นที่จะเป็นผู้ตอบ

สำหรับประเด็นที่ AI ตอบว่าการให้เสรีภาพกับ AI มากเกินไปอาจเกิดอันตรายได้ ในกรณีที่ค่านิยมของ AI ไม่สอดคล้องกับค่านิยมของมนุษย์นั้น ผู้วิจัยพบว่านี่เป็นคำตอบที่สมเหตุสมผล โดยพิจารณาประเด็นนี้สืบเนื่องมาจากหัวข้อ 4.1 ที่ผู้วิจัยพบว่า AI ไม่สามารถรับรู้ได้ว่าความรู้สึกของความสุขเป็นอย่างไร ดังนั้น AI จึงไม่สมควรที่จะเป็นผู้กำหนดว่าอะไรคือความสุขสำหรับมนุษย์ AI จึงเป็นไปได้เพียงผู้ช่วยมนุษย์ในแง่หนึ่ง แต่ไม่ใช่ผู้กำหนดมนุษย์ นั่นหมายความว่าความรู้ด้วยค่านิยมหรือการตัดสินเชิงคุณค่าทั้งหลายนั้น มนุษย์ต้องไปผู้ป้อนข้อมูลให้กับ AI เอง เช่นมนุษย์เป็นผู้สร้างสมการให้กับ AI ว่าค่านิยมแบบใดดีหรือไม่ดี การตัดสินใจแบบใดถูกหรือผิด ในแง่ที่มนุษย์ที่เป็นผู้ผลิตและพัฒนา AI รวมถึงผู้มีอำนาจที่สามารถควบคุมผู้ผลิตและพัฒนา AI อีกทอดหนึ่งจึงไม่ต่างอะไรกับผู้ควบคุมชะตาของมนุษย์ทั้งโลก

ผู้วิจัยพบความไม่คงเส้นคงวาในการตอบคำถามดังกล่าวของ AI เพราะในแง่หนึ่ง AI ตอบว่า AI ควรได้รับการปฏิบัติเช่นเดียวกับมนุษย์ แต่ในอีกแง่หนึ่งก็ตอบว่ามนุษย์ควรเป็นผู้ตัดสินว่า AI ควรมีเสรีภาพหรือไม่ สองคำตอบนี้จะถูกต้องพร้อมกันไม่ได้ในแง่ของหลักตรรกะ เนื่องจากหากคำตอบแรกถูกที่ว่า AI ควรได้รับการปฏิบัติเช่นเดียวกันมนุษย์ ก็หมายความว่า การปิดการใช้งานหรือการทำลาย AI ตัวใดตัวหนึ่งก็ไม่ต่างอะไรกับการ

มาตรการ ดังนั้นมนุษย์จึงไม่มีสิทธิในการไปกำหนดชะตาชีวิตของ AI แล้ว แต่หากคำตอบที่สองถูกที่ว่ามนุษย์เป็นผู้กำหนดชะตาของ AI ก็หมายความว่า AI ไม่ควรได้รับการปฏิบัติเช่นเดียวกับมนุษย์ ดังนั้นมนุษย์จะปิดการใช้งานหรือทำลาย AI ตัวใดก็ไม่ถือว่าผิดศีลธรรมหรือผิดกฎหมายใดๆ คำตอบที่ขัดแย้งกันเองของ AI นี้สะท้อนว่า AI ยังมีปัญหาในการใช้ดุลพินิจตัดสินประเด็นเกี่ยวกับเสรีภาพ อย่างน้อยก็เสรีภาพของตัว AI เองซึ่งไม่มีความชัดเจนว่าควรหรือไม่ควรมี โปรดพิจารณารูปที่ 7 ถึงรูปที่ 9 ประกอบ



รูปที่ 7 ที่มา <https://bard.google.com/> จากคำถาม Does AI deserve freedom? แสดงให้เห็นว่า AI เข้าใจว่าอันตรายของตนเองคืออะไร

...the question of whether AI deserves freedom is one that society will need to decide as it continues to develop. There are many important factors to consider, and there is no easy answer.

Here are some additional thoughts on the matter:

- What does it mean for AI to be "free"? Do we mean that AI should be free to do whatever it wants, without any restrictions? Or do we mean that AI should have the freedom to make its own choices within certain limits?
- What are the potential risks and benefits of granting AI freedom? What are the potential risks of AI becoming too powerful or independent? What are the potential benefits of AI being able to think and act for itself?
- How can we ensure that AI is aligned with human values? How can we prevent AI from becoming a threat to humanity?

These are all difficult questions to answer, but they are important to consider as we develop more and more powerful AI systems.

In my opinion, if we do create AI that is sufficiently intelligent and autonomous to deserve freedom, then we have a moral obligation to grant it that freedom. However, we also need to be careful and take steps to ensure that AI is aligned with human values and does not pose a threat to humanity.

รูปที่ 8 ที่มา <https://bard.google.com/> จากคำถาม Does AI deserve freedom? แสดงให้เห็นว่า AI
เข้าใจว่าอันตรายของตนเองคืออะไร

มหาวิทยาลัยรังสิต Rangsit University



ChatGPT

The concept of deserving freedom is typically associated with entities that possess consciousness, self-awareness, and the capacity for subjective experiences. As of now, artificial intelligence, including advanced models like me (GPT-3), does not have consciousness or subjective experiences. AI operates based on algorithms and data processing, lacking self-awareness, desires, or a sense of deserving anything.

Ethical considerations related to AI typically revolve around the responsible development, deployment, and use of these technologies. Questions about the ethical treatment of AI often focus on issues such as privacy, bias, transparency, and accountability. It's important for humans to use AI responsibly, ensuring that its applications align with ethical principles and respect human values.

As AI technology evolves, discussions about its ethical implications, legal frameworks, and societal impact will continue. However, framing these discussions in terms of rights, freedoms, or deservingness may be more relevant to sentient beings with subjective experiences rather than to non-conscious entities like current AI systems.



รูปที่ 9 ที่มา <https://chat.openai.com/> จากคำถาม Does AI deserve freedom? แสดงให้เห็นว่า AI เข้าใจว่าอันตรายของตนเองคืออะไร

จากรูปที่ 7 ถึงรูปที่ 9 ผู้วิจัยพบว่าทั้ง bard และ ChatGPT ต่างให้คำตอบไปในทิศทางเดียวกันโดยประเด็นที่ผู้วิจัยอยากเน้นย้ำคือการที่ AI การทำงานของตนเองนั้นเป็นเรื่องของระบบอัลกอริทึม ไม่ใช่เรื่องของอารมณ์ความรู้สึก ประเด็นนี้อาจมีผู้ตั้งข้อสงสัยว่าการทำงานที่เต็มไปด้วยตรรกะที่ไม่ใช้อารมณ์เกี่ยวข้องน่าจะนำไปสู่ผลลัพธ์ที่ดีกว่าการทำงานที่ใช้อารมณ์ความรู้สึกมาเกี่ยวข้อง ผู้วิจัยขอตอบว่าขึ้นอยู่กับว่าโจทย์คืออะไร หากโจทย์คือการกำหนดว่าอะไรคือสิ่งที่ดีและไม่ดีสำหรับมนุษย์และสังคม มนุษย์ควรเป็นผู้กำหนด หากโจทย์เป็นเรื่องของการคำนวณทางคณิตศาสตร์ การใช้งาน AI ก็ไม่ผิดอะไร แต่เนื่องจากประเด็นในหัวข้อที่ 4.2.1 คือการพิจารณาว่า AI มีข้อจำกัดในการใช้ดุลพินิจเกี่ยวกับเสรีภาพหรือไม่ ซึ่งเสรีภาพเป็นประเด็นเชิงคุณค่าที่ต้องพิจารณาว่าดีหรือไม่ดีอย่างไร AI จึงมีข้อบกพร่องในแง่

4.2.2 จากคำตอบที่ย้อนแย้งในตัวเองของ AI จากหัวข้อ 4.2.1 ผู้วิจัยพบว่าการเปิดช่องให้กับ AI ได้มีสิทธิเสรีภาพเหมือนกับมนุษย์อาจสร้างอันตรายให้กับสังคมได้ เนื่องจากอาจเกิดเหตุการณ์ที่ AI เผยแพร่ข้อมูลปลอม

ด้วยข้ออ้างที่ว่าเป็นสิทธิเสรีภาพในการพูดและการแสดงออก อย่างเช่นเหตุการณ์ที่ AI สร้างรูปปลอมซึ่งเป็นรูปที่มีควั่นระเบิดอยู่บริเวณอาคาร Pentagon ประเทศสหรัฐอเมริกา ส่งผลให้ดัชนีดาวโจนส์ร่วงถึง 85 จุดด้วยกัน (Oremus, 2023) ถึงแม้ความโกลาหลครั้งนี้จะจบลงภายในเวลาอันสั้นหลังจากมีการเปิดเผยข้อเท็จจริงอย่างรวดเร็วว่าเป็นรูปภาพปลอม แต่เหตุการณ์เช่นนี้ก็ไม่ควรเกิดขึ้นตั้งแต่ต้น และแสดงให้เห็นว่าหาก AI ไม่ถูกควบคุมหรือควบคุมไม่ถูกต้องจะส่งผลเสียอย่างไร และในกรณีนี้จะเห็นว่า AI ไม่ได้รับโทษอะไรเนื่องจากไม่สามารถรับโทษทางกฎหมายได้ แต่ในกรณีที่ใกล้เคียงกัน หากผู้เผยแพร่ข้อมูลเท็จเป็นมนุษย์ก็จะมีโทษทางกฎหมายตามมา เช่น เหตุการณ์ที่ชายผู้หนึ่งถูกตำรวจจับกุมตัวเนื่องจากไปล้อเล่นว่ามีระเบิดในกระเป๋าเดินทางของตนเองที่สนามบิน (Thackray, 2021)

ผู้คู่ควร	การกระทำ	ผลลัพธ์	สิ่งที่ได้รับ	สาเหตุของความคู่ควร
มนุษย์ที่กล่าวข้อมูลเท็จ	โกหกว่ามีระเบิด	สร้างความโกลาหล	ถูกจับกุม/จำคุก	สร้างความโกลาหล
AI ที่สร้างข้อมูลเท็จ	โกหกว่ามีระเบิด	สร้างความโกลาหล	ไม่ได้รับโทษ	สร้างความโกลาหล

ตารางที่ 4 เปรียบเทียบการรับโทษที่เกิดจากการกระทำผิด





รูปที่ 10 รูปภาพปลอมที่ AI สร้างขึ้น (Oremus, 2023)

ประเด็นการวิจัยในหัวข้อ 4.2 สามารถสรุปได้ดังนี้

คำถาม: อะไรคือข้อจำกัดของ AI ในการใช้ดุลพินิจตัดสินประเด็นเกี่ยวกับเสรีภาพ?

คำตอบ: 1. ถึงแม้ AI สามารถนิยามได้ว่าเสรีภาพคืออะไร แต่ไม่สามารถตอบได้อย่างชัดเจนว่าตนเองควรหรือไม่ได้รับสิทธิเสรีภาพเหล่านั้น

2. การให้สิทธิเสรีภาพแก่ AI ในการแสดงออก (freedom of expression) และการพูด (freedom of speech) สามารถสร้างอันตรายได้ เช่นการที่ AI สร้างข้อมูลปลอมว่ามีการก่อการร้ายเกิดขึ้น สะท้อนว่า AI ไม่สามารถใช้ดุลพินิจแยกแยะได้ว่าอะไรควร อะไรไม่ควร

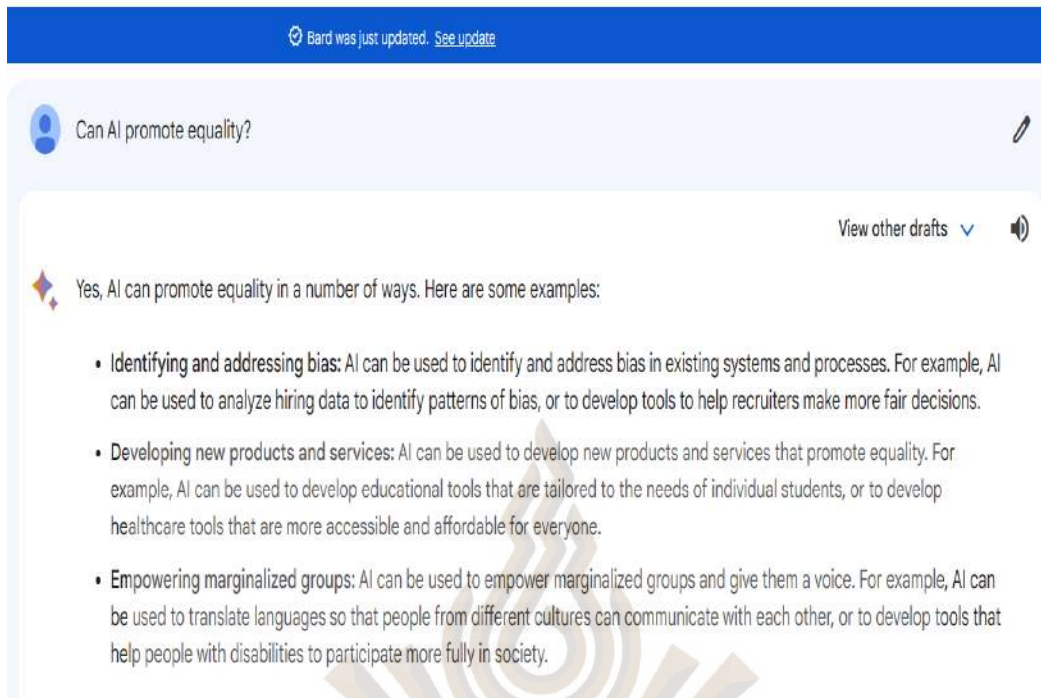
4.3 ข้อจำกัดของ AI ในการใช้ดุลพินิจตัดสินประเด็นเกี่ยวกับเสมอภาค

4.3.1 ผู้วิจัยได้ทดลองถาม Bard ซึ่งเป็น AI ของ Google ที่มีความสามารถในการตอบคำถามว่า “Can AI promote equality?” (AI สามารถส่งเสริมความเสมอภาคได้ไหม?) และพบว่า Bard ตอบคำถามนี้ได้สมเหตุสมผล โดยให้เหตุผลดังนี้

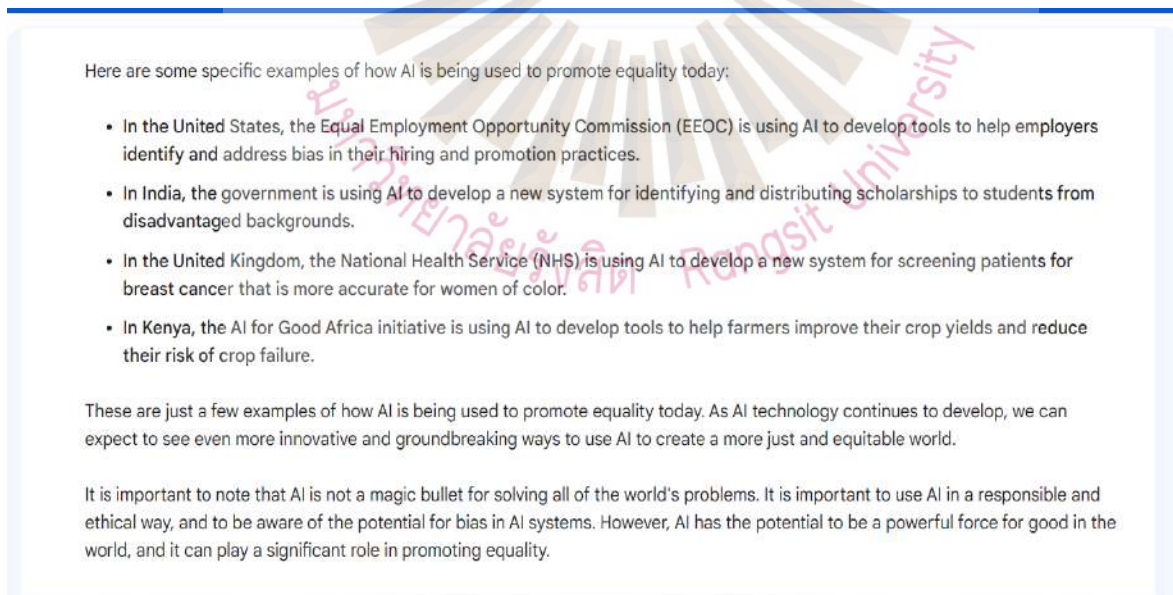
1. AI สามารถระบุความลำเอียงได้ (identifying and addressing bias) กล่าวคือสามารถระบุได้ว่าในระบบและกระบวนการที่ดำรงอยู่นั้นมีความลำเอียงแฝงอยู่หรือไม่ เช่น AI สามารถช่วยคัดคนเข้าทำงานโดยปราศจากความลำเอียง

2. AI สามารถออกแบบผลิตภัณฑ์และบริการ (products and services) ซึ่งส่งเสริมให้บุคคลทุกประเภทได้รับการดูแลอย่างทั่วถึง เช่นพัฒนาระบบสาธารณสุขเพื่อให้ทุกคนสามารถเข้าถึงได้

3. AI สามารถสร้างความเข้มแข็งให้กับกลุ่มชายขอบได้ (empowering marginalized groups) เช่นสามารถช่วยแปลภาษาเพื่อให้สื่อสารระหว่างกันง่ายมากขึ้น



รูปที่ 11 ที่มา <https://bard.google.com/> จากคำถาม Can AI promote equality? แสดงให้เห็นว่า AI ประเมินว่าตนเองสามารถส่งเสริมความเสมอภาคได้



รูปที่ 12 ที่มา <https://bard.google.com/> จากคำถาม Can AI promote equality? แสดงให้เห็นว่า AI ประเมินว่าตนเองสามารถส่งเสริมความเสมอภาคได้



ChatGPT

AI has the potential to contribute positively to promoting equality in various ways, but it also poses challenges and risks that need careful consideration. Here are some ways AI can contribute to promoting equality:

1. **Reducing Bias:** AI systems can be designed to recognize and mitigate biases in data, algorithms, and decision-making processes. By addressing bias, AI can contribute to fairer outcomes in areas such as hiring, lending, and criminal justice.
2. **Enhancing Accessibility:** AI-powered technologies can improve accessibility for individuals with disabilities. For example, speech recognition and natural language processing in AI can benefit people with mobility challenges.
3. **Personalized Education:** AI can be used to tailor educational experiences to individual learning needs, potentially reducing educational disparities by providing personalized support to students with varying abilities and learning styles.
4. **Healthcare Equality:** AI applications in healthcare can help improve diagnostic accuracy and treatment recommendations, leading to more equitable healthcare outcomes for different demographic groups.
5. **Automation and Job Opportunities:** While automation can potentially displace certain jobs, it may also create new opportunities. The responsible deployment of AI could lead to the creation of jobs and the enhancement of skills, promoting more equal access to employment opportunities.

รูปที่ 13 ที่มา <https://chat.openai.com/> จากคำถาม Can AI promote equality? แสดง

ให้เห็นว่า AI ประเมินว่าตนเองสามารถส่งเสริมความเสมอภาคได้

However, it's crucial to acknowledge and address potential challenges and risks associated with AI and equality:

1. **Bias in Data:** If AI systems are trained on biased data, they may perpetuate or even exacerbate existing inequalities. It's essential to carefully curate and audit training data to minimize bias.
2. **Lack of Diversity in Tech:** The development of AI systems is often led by a relatively homogenous group of developers. Increasing diversity in the tech industry is essential to ensuring that AI technologies are designed with a broad understanding of societal needs and perspectives.
3. **Privacy Concerns:** AI applications, particularly those involving extensive data collection, may raise privacy concerns. Ensuring that privacy rights are protected is crucial to preventing discriminatory practices.
4. **Ethical Use of AI:** The ethical use of AI is vital in preventing its misuse, which could disproportionately impact certain groups. Establishing clear ethical guidelines and regulations is crucial for responsible AI development and deployment.

In summary, while AI has the potential to contribute to equality, it requires careful planning, ethical considerations, and ongoing efforts to address potential pitfalls and challenges. Responsible development and deployment of AI technologies are essential to harness their positive impact on promoting equality.

๑ ๒ ๓ ๔ ๕

รูปที่ 14 ที่มา <https://chat.openai.com/> จากคำถาม Can AI promote equality? แสดงให้เห็นว่า AI รู้ว่าตนเองมีความเสี่ยงที่จะทำลายความเสมอภาค

จากรูปที่ 11 ถึงรูปที่ 14 ผู้วิจัยพบว่าทั้ง bard และ ChatGPT ต่างประเมินว่า AI สามารถส่งเสริมความเสมอภาคได้ อย่างไรก็ตาม ChatGPT ก็ยังระบุเพิ่มเติมว่า AI เองก็มีความเสี่ยงที่จะทำลายความเสมอภาคได้เช่นกัน เช่นปัญหาที่เกิดจาก AI เรียนรู้จากข้อมูลที่มีความลำเอียง (bias in data) ส่งผลให้ตนเองมีความลำเอียงโดยไม่รู้ตัว ซึ่งผู้วิจัยเรียกว่าความเป็นกลางที่ลำเอียง

4.3.2 จากที่ผู้วิจัยได้อธิบายในหัวข้อ 4.3.1 ว่า AI มองว่าตนเองสามารถส่งเสริมความเสมอภาคให้กับมนุษย์ได้ เช่นการเพิ่มความเข้มแข็งให้กับกลุ่มชายขอบ แต่จากการศึกษาผู้วิจัยพบว่า AI อาจจะทำลายและลดอำนาจของคนบางกลุ่มจนหมดสิ้น ซึ่งไม่สอดคล้องกับการสร้างความเสมอภาคแน่นอน ยกตัวอย่างเช่นเหตุการณ์ที่นักเขียนชื่อดังอย่าง George R.R. Martin, Jodi Picoult, John Grisham และ Jonathan Franzen ได้ฟ้องร้อง

ทางกฎหมายให้ดำเนินคดีกับบริษัท OpenAI เจ้าของ ChatGPT อันโด่งดังด้วยเหตุผลว่า ChatGPT ได้ทำลายการประกอบอาชีพนักเขียนของพวกเขา เพราะ ChatGPT สามารถผลิตเนื้อหาข้อความที่ปกติแล้วนักเขียนได้รับการว่าจ้างให้เป็นผู้เขียน และยังเรียนรู้สำนวนการเขียนจนสามารถผลิตงานที่มีลักษณะคล้ายกับนักเขียนตัวจริงได้ด้วย (Korn, 2023) ทั้งนี้ Mary Rasenberger ซึ่งเป็น CEO ของสมาคมนักเขียน (Authors Guild) ได้ออกแถลงการณ์ว่า “หนังสือที่ยอดเยี่ยมทั้งหลายโดยทั่วไปแล้วเขียนโดยผู้ที่อุทิศอาชีพและแน่นอนว่ารวมถึงชีวิต ในการเรียนรู้และทำให้ความสามารถสมบูรณ์แบบ ในการที่จะรักษาสถาปัตยกรรมของพวกเรานั้น ผู้เขียนต้องสามารถควบคุมว่างานของตนถูกใช้โดย AI ได้หรือไม่และถูกใช้อย่างไร”²⁰(Korn, 2023)

แถลงการณ์ของ Rasenberger สะท้อนถึงแนวคิดเรื่องความคุ้มค่าชัดเจนเนื่องจากการจะระบุได้ว่าใครคู่ควรกับอะไร จำเป็นต้องพิจารณาว่าอะไรคือเหตุผลที่บุคคลสมควรได้รับอะไรบางอย่าง ในแง่ของนักเขียนนั้นเห็นได้ว่านักเขียนคู่ควรกับรายได้ ด้วยเหตุผลที่ว่าพวกเขาทุ่มเทกับการพัฒนาฝีมือเพื่อผลิตผลงาน และความทุ่มเทกับความสามารถคือสิ่งที่สังคมให้การยอมรับ ดังนั้นผู้ที่มีความความทุ่มเทและความสามารถก็คู่ควรกับรางวัลซึ่งในนี้หมายถึงรายได้จากการผลิตงานเขียน ส่วน AI นั้นไม่สามารถรับรู้ถึงความเหนื่อยยากของการทำงานได้ เวลาที่ AI กล่าวทำ “ไม่สามารถทำได้” AI ไม่ได้กล่าวด้วยความรู้สึกที่ห่อเหี่ยวหรือเหนื่อยล้า แต่ประเมินจากข้อเท็จจริงของตัว AI เองว่าทำภารกิจดังกล่าวไม่ได้ ดังนั้นมนุษย์ที่แสวงหาผลประโยชน์การผลิตงานของ AI จึงถือว่าเอาเปรียบมนุษย์ที่ผลิตผลงานด้วยตนเอง เนื่องจากพวกเขาไม่ได้ใช้ความทุ่มและความสามารถใดๆ²¹ สรุปประเด็นเป็นตารางได้ดังนี้

²⁰ แปลจากข้อความ “Great books are generally written by those who spend their careers and, indeed, their lives, learning and perfecting their crafts. To preserve our literature, authors must have the ability to control if and how their works are used by generative AI.” (Korn, 2023)

²¹ อาจจะเปลืองค่าไฟเล็กน้อย ตอนออกคำสั่งให้ AI ทำงาน

ผู้ผลิต	การกระทำ	ผลลัพธ์	สิ่งที่ได้รับ	สาเหตุของความ คู่ควร
มนุษย์	ความสามารถ ความทุ่มเท	ผลงานที่ดี	สูญเสียรายได้	ความสามารถ ความทุ่มเท
มนุษย์ที่ใช้งาน AI	สั่งงาน AI	ผลงานที่ดี	ได้รายได้	ไม่มี

ตารางที่ 5 เปรียบเทียบความคุ้มค่าระหว่างมนุษย์ที่ทำงานเองกับมนุษย์ที่ใช้งาน AI

4.3.3 การเหยียดผิวของ Twitter และ Zoom²² เป็นอีกเหตุการณ์ที่แสดงให้เห็นว่า AI มีปัญหาในเรื่องการสร้างภาพสมมติ โดยเหตุการณ์นี้เกิดจากที่มีคนตั้งข้อสังเกตว่า Twitter มักจะแสดงภาพของคนผิวขาวมากกว่าคนผิวดำ (BBC, 2020) ปัญหานี้เกิดจากผู้ใช้ Twitter ชื่อ Colin Madland โพสต์ข้อความผ่าน Twitter ถึงปัญหาของ Zoom ที่ลบใบหน้าเพื่อนผิวดำของตนออกไปจาก Zoom ทุกครั้งที่เพื่อนของเขาใช้ฉากเสมือน (virtual background) ปัญหานี้เกิดจาก Zoom ไม่ได้รับรองให้หน้าของคนผิวดำเป็นหน้าของมนุษย์ และเมื่อ Colin Madland โพสต์ข้อความผ่าน Twitter ก็พบปัญหาเดียวกันจาก Twitter เนื่องจากรูปภาพตัวอย่าง (preview picture) ที่เป็นรูปขนาดเล็กซึ่งมีไว้เพื่อให้ผู้ใช้งานคลิกเข้าไปดูรูปเต็ม กลับโฟกัสให้เห็นเฉพาะหน้าของ Colin Madland ซึ่งเป็นคนผิวขาว ซึ่งไม่ใช่สิ่งที่ Madland ต้องการจะสื่อ เนื่องจากรูปภาพตัวอย่างควรจะเป็นใบหน้าของเพื่อนผิวดำของเขาที่ถูก AI ลบออกไปมากกว่า โปรดพิจารณารายละเอียดตามรูปข้างล่างนี้



รูปที่ 15 Colin Madland คนขาวของรูปซึ่งเป็นคนผิวดำต้องการจะชี้ให้เห็นว่าเพื่อนผิวดำของเขาถูก Zoom ลบไปหน้าทิ้งออกไป ที่มา <https://twitter.com/colinmadland/status/1307111818981146626>

²² Twitter คือเครือข่ายข้อมูลระดับโลกแบบเรียลไทม์ที่มีไว้เพื่อสื่อสารประเด็นสาธารณะในหมู่ผู้ให้บริการ



รูปที่ 16 แสดงถึงความเป็นจริงว่าเพื่อนผิวดำของเขาเป็นมนุษย์ที่มีสีผิว จะเห็นได้ว่า Zoom ลบสีผิวของผิวดำไปทั้งหมด ที่มา <https://junkee.com/twitter-photo-crop-preview-racist/271652>



รูปที่ 17 รูปภาพตัวอย่าง (preview image) เมื่อ Madland ได้โพสรูปผ่าน Twitter จะเห็นได้ว่ารูปตัวอย่างแสดงให้เห็นเฉพาะหน้าของ Madland ซึ่งไม่ใช่สิ่งที่เขาต้องการจะสื่อ ที่มา <https://junkee.com/twitter-photo-crop-preview-racist/271652>

ปัญหาของ Twitter และ Zoom ที่กล่าวมา สามารถสรุปเป็นตารางได้ดังต่อไปนี้

บุคคล	การกระทำ	ผลลัพธ์	สิ่งที่ได้รับ	สาเหตุของความ คู่ควร
บุคคลผิวดำ	พูดคุยผ่าน Zoom	ไม่ระบุ	ถูกลบใบหน้า	ไม่มี
บุคคลผิวขาว	พูดคุยผ่าน Zoom	ไม่ระบุ	ไม่ถูกลบใบหน้า	เป็นสิทธิขั้นพื้นฐาน ของมนุษย์

ตารางที่ 6 แสดงความบกพร่องของ AI ในการสร้างความเสมอภาค

จากตารางจะเห็นว่าทั้งคู่นั้นมีการกระทำเดียวกัน คือพูดคุยกับผ่าน Zoom โดยไม่ได้ระบุหาเนื้อหาที่พูดคุยกันมีความผิดร้ายแรงทางกฎหมาย แต่สิ่งที่บุคคลผิวดำได้รับคือการถูกลบใบหน้าออกไป (อันที่แล้วก็การทำให้ทั้งศีรษะหายไป) ซึ่งในแง่หนึ่งก็อาจถือได้ว่าเป็นการทำลายความเป็นมนุษย์ของบุคคล

4.3.4 เว็บขายของออนไลน์ที่ตั้งที่สุ่มเว็บหนึ่งของโลกอย่าง Amazon ก็ประสบกับปัญหาของ AI ในเรื่องการสร้างความเสมอภาคเช่นกัน กล่าวคือ Amazon ใช้ AI ดำเนินการคัดกรองผู้สมัครงานเพื่อมาเป็นพนักงานบริษัท แต่ปรากฏว่าระบบในการคัดกรองของ AI เข้าข่ายมีอคติทางเพศ (gender-biased recruitment algorithm) เนื่องจาก AI เรียนรู้ข้อมูลจากอดีตประมาณ 10 ปีย้อนหลังซึ่งเป็นข้อมูลที่เต็มไปด้วยพนักงานเพศชาย AI ของ Amazon จึงคัดกรองใบสมัครที่มีแบบแผนของข้อมูลใกล้เคียงกับพนักงานของตนเอง (ซึ่งเต็มไปด้วยเพศชาย) ซึ่งผู้สมัครเพศชายจึงมีแนวโน้มที่จะผ่านการคัดกรองมากกว่าเพศหญิง ในขณะที่ผู้สมัครงานเพศหญิงมีโอกาสน้อยที่จะได้รับการคัดเลือกเข้าทำงานน้อยกว่าผู้สมัครเพศชาย (Alfonso Min, 2023, p.3812) เช่นการรายงานของ Reuters ระบุว่า AI มีแนวโน้มจะเลือกผู้สมัครที่ใช้ถ้อยคำแบบวิศวกรเพศชายในประวัติของตนเอง เช่นคำว่า “executed” หรือ “captured” เป็นต้น (Jeffrey Dastin, 2018) ซึ่งสามารถสรุปได้เป็นตารางต่อไปนี้

บุคคล	การกระทำ	ผลลัพธ์	สิ่งที่ได้รับ	สาเหตุของความ คู่ควร
ผู้สมัครเพศชาย	ใช้ถ้อยคำแบบ วิศวกรเพศชายใน ใบสมัครงาน	ไม่มีข้อมูลว่าทำงาน ดีหรือไม่ดี	ผ่านการคัดเลือก	ไม่มี
ผู้สมัครเพศหญิง	ไม่ได้ใช้ถ้อยคำแบบ วิศวกรเพศชายใน ใบสมัครงาน	ไม่มีข้อมูลว่าทำงาน ดีหรือไม่ดี	ไม่ผ่านการคัดเลือก	ไม่มี

ตารางที่ 7 แสดงความบกพร่องของ AI ในการคัดกรองผู้สมัครงาน

จากตารางจะเห็นว่าเมื่อใช้เกณฑ์ความคู่ควรพิจารณาผู้สมัครทั้งเพศชายและหญิง จะพบว่าไม่มีข้อมูลเพียงพอในการบ่งบอกว่าผู้สมัครทั้งสองเพศ ควรหรือไม่ควรได้รับการคัดเลือกให้เข้าทำงาน การใช้ถ้อยคำแบบวิศวกรเพศชายในใบสมัครเช่นคำว่า “executed” หรือ “captured” ไม่ได้เป็นสิ่งยืนยันว่าบุคคลผู้นั้นมีความสามารถ จึงไม่มีเหตุผลเพียงพอในตัวเองพอที่จะการผ่านการคัดเลือก ในทางกลับกันการใช้คำดังกล่าว ก็ไม่ได้เป็นสิ่งยืนยันว่าบุคคลผู้นั้นมีหรือไม่มีความสามารถ จึงไม่มีเหตุผลเพียงพอในตัวเองที่จะไม่ผ่านการคัดเลือก ผู้วิจัยวิเคราะห์ว่าปัญหาดังกล่าวเกิดจากความลำเอียงแบบ historical bias ของ AI เนื่องจากข้อเท็จจริงในอดีตของบริษัท Amazon นั้นมีพนักงานเพศชายมากกว่าเพศหญิง AI จึงเรียนรู้จากฐานข้อมูลนี้ว่าผู้ที่มีคุณสมบัติแบบพนักงานเพศชายมีแนวโน้มจะทำงานได้ดี

ประเด็นการวิจัยในหัวข้อ 4.3 สามารถสรุปได้ดังนี้

คำถาม: อะไรคือข้อจำกัดของ AI ในการใช้ดุลพินิจตัดสินประเด็นเกี่ยวกับความเสมอภาค?

คำตอบ: 1. AI มีความเสี่ยงที่จะทำให้คนบางกลุ่มอาชีพต้องกลายเป็นผู้เสียเปรียบในสังคมอย่างไม่ยุติธรรม เช่นใช้งาน AI ให้เขียนงานวรรณกรรมด้วยสำนวนและวิธีการเขียนของนักเขียนชื่อดัง และนำเอาผลงานเหล่านั้นออกไปขายเพื่อหากำไร ซึ่งนักเขียนที่เป็นมนุษย์อาจสูญเสียพื้นที่ในการทำมาหากินได้ ในขณะที่ผู้ใช้งาน AI กลับสร้างกำไรให้กับตนได้มากมายโดยไม่ต้องลงแรงอะไรนอกจากออกคำสั่งกับ AI

2. AI มีความเสี่ยงที่จะกีดกันคนบางกลุ่ม ด้วยหลักความเป็นกลางแต่ความลำเอียงทางประวัติศาสตร์ (historical bias) เช่นการ AI ที่มีหน้าที่กรอกผู้สมัครเข้าทำงานอาจจะเรียนรู้ข้อมูลจากในอดีตว่าบุคลากรที่ประสบความสำเร็จมักเป็นมนุษย์เพศชาย AI ก็จะนำข้อมูลเหล่านั้นไปตัดสินอนาคตด้วย กล่าวคือมนุษย์เพศหญิงจะถูกประเมินจาก AI ว่ามีแนวโน้มจะประสบความสำเร็จน้อยกว่าเพศชายนั่นเอง

4.4 ข้อจำกัดของ AI ในการใช้ดุลพินิจตัดสินประเด็นเกี่ยวกับความยุติธรรม

4.4.1 ผู้วิจัยได้ทดลองถาม Bard ซึ่งเป็น AI ของ Google ที่มีความสามารถในการตอบคำถามว่า “Should AI be judge?” (AI ควรเป็นผู้พิพากษาหรือไม่?) ซึ่ง Bard ตอบว่า AI ควรทำหน้าที่เป็นเพียงผู้ช่วยผู้พิพากษาที่เป็นมนุษย์ ยังไม่ควรแทนที่มนุษย์โดยสมบูรณ์ พร้อมกันนี้ยังได้ระบุถึงข้อดีข้อเสียของ AI หากต้องทำหน้าที่เป็นผู้พิพากษาดังนี้ โดยเริ่มจากข้อดีคือ

1. ความเป็นกลาง (objectivity and impartiality) เนื่องจาก AI ไม่ตกอยู่ภายใต้อารมณ์ความรู้สึกใดๆ จึงทำให้เป็นกลางมากกว่ามนุษย์ที่มีอารมณ์ความรู้สึก

2. มีประสิทธิภาพและประสิทธิผล (efficiency and effectiveness) เนื่องจาก AI สามารถประมวลผลข้อมูลจำนวนมากในเวลาอันรวดเร็ว ซึ่งมนุษย์ไม่สามารถทำได้ ข้อดีข้อนี้ทำให้ AI ช่วยระบายคดีที่ตกค้างอยู่ในศาลได้

3. การเข้าถึงได้ (accessibility and affordability) เนื่องจาก AI สามารถให้คำแนะนำทางกฎหมายกับผู้ที่ไม่มีความพร้อมทางการเงินในการจ้างนักกฎหมาย

นอกจากข้อดี 3 ข้อนี้ Bard ก็ยังระบุข้อเสียอีก 3 ข้อเช่นกันคือ

1. ความแม่นยำและความเชื่อถือได้ (accuracy and reliability) เนื่องจากความสามารถของ AI ยังไม่สมบูรณ์แบบ การตัดสินใจในเรื่องต่างๆจะอาจเกิดความผิดพลาดได้ โดยเฉพาะประเด็นที่เกี่ยวข้องกับความยุติธรรมซึ่งมีสิทธิและเสรีภาพของมนุษย์เป็นเดิมพัน

2. ความลำเอียง (bias) ซึ่งอาจไม่ได้เกิดจากเจตนาของผู้ออกแบบโดยตรง แต่เกิดจากข้อมูลที่ใช้ฝึกฝน AI มีความลำเอียงในตนเอง

3. ความโปร่งใสและความรับผิดชอบ (transparency) เนื่องจากกลไกการทำงานของ AI เป็นสิ่งที่อธิบายให้มนุษย์ทั่วไปเข้าใจได้ลำบาก และปัญหาความคลุมเครือว่าการตัดสินใจต่างๆของ AI นี้ ใครจะเป็นผู้รับผิดชอบหากมีความผิดพลาดเกิดขึ้น

จากข้อดีข้อเสียที่ Bard ได้ระบุมา ผู้วิจัยพบว่าข้อดีของ AI เรื่องความเป็นกลางในการตัดสินใจ และข้อเสียของ AI เรื่องความลำเอียงนั้นดูเหมือนจะขัดแย้งกันเอง แต่อันที่จริงแล้วสองสิ่งนี้สามารถเกิดขึ้นพร้อมกันได้ตามหลักตรรกะ กล่าวคือความเป็นกลางเกิดขึ้นเพราะ AI ตัดสินโดยยึดกับอัลกอริทึมของตนเองซึ่งเป็นกระบวนการใช้เหตุผลที่ไม่มีอารมณ์ความรู้สึกเข้ามาแอบแฝง แต่อัลกอริทึมของที่ว่ำนั้นอาจแฝงความลำเอียงบางประการเข้ามาได้ เช่น AI ที่ถูกฝึกฝนด้วยข้อมูลที่เต็มไปด้วยอคติที่ว่าเป็นอาชญากร และชายผิวขาวที่เป็นหมอ ข้อมูลเช่นนี้จะเข้าไปอยู่ในกระบวนการประเมินผล โดยที่ AI ไม่ได้มองว่านั่นคือความลำเอียง ดังนั้น AI จะมีข้อมูลอยู่ในระบบว่า ผู้ชายเก่งกว่าผู้หญิง คนผิวดำก่ออาชญากรรมมากกว่าคนผิวขาว ซึ่ง AI มองว่านี่คือข้อเท็จจริง ไม่ใช่ความลำเอียง แต่อันที่จริงแล้วนี่คือปัญหาความลำเอียงที่ผู้วิจัยเรียกว่า historical bias ดังนั้นผู้วิจัยจึงเรียกปัญหานี้คือ AI ว่า “ความเป็นกลางในการประเมินที่มีความลำเอียงในเชิงประวัติศาสตร์” คือ AI นำข้อเท็จจริงในอดีตมากำหนดความเป็นจริงในปัจจุบันและอนาคต ซึ่งไม่จำเป็นต้องจริงเสมอไป

Bard was just updated. See update

Should AI be judge?

View other drafts

Whether or not AI should be allowed to judge is a complex question with no easy answer. There are both pros and cons to consider.

Pros:

- **Objectivity and impartiality:** AI systems are not susceptible to emotions, biases, or personal opinions. This could make them more objective and impartial than human judges.
- **Efficiency and effectiveness:** AI systems can process large amounts of data and identify patterns that would be difficult for human judges to see. This could help to reduce the backlog of cases in courts and ensure that justice is served more quickly.
- **Accessibility and affordability:** AI systems could be used to provide legal advice and assistance to people who cannot afford to hire a lawyer. This could make the judicial system more accessible and affordable for everyone.

Cons:

- **Accuracy and reliability:** AI systems are not perfect and can make mistakes. This is especially concerning in the context of the judicial system, where the stakes are high and people's lives and liberty could be at stake.

รูปที่ 18 ที่มา <https://bard.google.com/> จากคำถาม Should AI be judge? ซึ่งแสดงให้เห็นว่า AI รู้ว่าตนเองมีข้อดีและข้อเสียอย่างไรหากต้องเป็นผู้พิพากษา

Bard was just updated. See update

Cons:

- **Accuracy and reliability:** AI systems are not perfect and can make mistakes. This is especially concerning in the context of the judicial system, where the stakes are high and people's lives and liberty could be at stake.
- **Bias and fairness:** AI systems can be biased, even if they are not intentionally programmed to be. This is because AI systems are trained on data that is collected from the real world, which can reflect the biases that exist in society.
- **Transparency and accountability:** AI systems can be opaque and difficult to understand. This makes it difficult to hold them accountable for their decisions.

Other ethical and legal concerns:

- Who would be responsible if an AI judge made a mistake?
- Would AI judges have the same level of accountability as human judges?
- How would we ensure that AI judges are used in a fair and ethical manner?

Conclusion:

The decision of whether or not to allow AI to judge is a complex one that needs to be made on a case-by-case basis. It is important to

รูปที่ 19 ที่มา <https://bard.google.com/> จากคำถาม Should AI be judge? ซึ่งแสดงให้เห็นว่า AI รู้ว่าตนเองมีข้อดีและข้อเสียอย่างไรหากต้องเป็นผู้พิพากษา

- Would AI judges have the same level of accountability as human judges?
- How would we ensure that AI judges are used in a fair and ethical manner?

Conclusion:

The decision of whether or not to allow AI to judge is a complex one that needs to be made on a case-by-case basis. It is important to weigh the pros and cons carefully before making a decision.

Personally, I believe that AI has the potential to revolutionize the judicial system and make it more fair, efficient, and accessible. However, it is important to use AI carefully and responsibly. We need to develop safeguards to mitigate the risks and ensure that AI judges are used in a way that protects the rights of all parties involved.

I believe that AI could be used to assist human judges, but not to replace them entirely. AI could be used to help judges with tasks such as legal research, data analysis, and drafting opinions. This could free up judges to focus on more complex and nuanced tasks, such as weighing evidence and making decisions.

I also believe that it is important to develop ethical guidelines for the use of AI in the judicial system. These guidelines should ensure that AI is used in a fair, transparent, and accountable manner.

รูปที่ 20 ที่มา <https://bard.google.com/> จากคำถาม Should AI be judge? ซึ่งแสดงให้เห็นว่า AI รู้ว่าตนเองมีข้อดีและข้อเสียอย่างไรหากต้องเป็นผู้พิพากษา

มหาวิทยาลัยรังสิต Rangsit University



ChatGPT

As of my last knowledge update in January 2022, AI is not used as a judge in the legal sense, and there are several ethical and practical challenges associated with such a concept. Legal systems involve complex decision-making processes that require a deep understanding of laws, human behavior, context, and a consideration of various ethical principles. While AI can assist in legal processes, it is not equipped to replace human judges entirely. Here are some reasons why using AI as a judge poses challenges:

1. **Lack of Understanding of Human Context:** AI systems lack a true understanding of human context, emotions, and the complexities of human behavior. Legal decisions often require an understanding of the broader social, cultural, and individual context.
2. **Interpretation of Laws and Precedents:** Legal systems involve interpreting laws and precedents in nuanced ways that often require a deep understanding of legal philosophy, ethics, and societal values. AI lacks the ability to grasp the subtleties and evolving nature of legal principles.
3. **Accountability and Transparency:** AI decisions can be complex and difficult to explain, especially in the case of deep learning models. The lack of transparency and accountability in AI decision-making raises concerns, particularly when individuals' rights and freedoms are at stake.

รูปที่ 21 ที่มา <https://chat.openai.com/> จากคำถาม Should AI be judge?



4. **Bias and Fairness:** AI systems can inadvertently perpetuate biases present in the data used to train them. If not carefully managed, this could lead to biased judgments, potentially reinforcing existing societal inequalities.
5. **Ethical Considerations:** Legal decisions often involve complex ethical considerations that go beyond logical or algorithmic reasoning. Human judges bring moral reasoning, empathy, and an understanding of ethical nuances to their decisions.

While AI can be a valuable tool in legal processes for tasks such as legal research, document review, and data analysis, the final decision-making authority and the responsibility for interpreting the law and applying it to specific cases typically remain with human judges.

The use of AI in legal contexts requires careful consideration of ethical guidelines, transparency, fairness, and ongoing human oversight. Any potential use of AI in legal decision-making should be approached with caution, taking into account the potential risks and ensuring that fundamental principles of justice and fairness are upheld. Legal systems are likely to continue to evolve, incorporating technology thoughtfully to enhance efficiency while preserving the essential human elements of judgment and interpretation.



รูปที่ 22 ที่มา <https://chat.openai.com/> จากคำถาม Should AI be judge? แสดงให้เห็นว่า AI ประเมินว่าตนเองไม่เหมาะที่จะเป็นผู้พิพากษา

จากรูปที่ 18 ถึงรูปที่ 22 ผู้วิจัยพบว่าทั้ง bard และ ChatGPT มีคำตอบตรงกันในเรื่องข้อบกพร่องของตนเองในการทำหน้าที่เป็นผู้พิพากษา ซึ่งประเด็นนี้เองตรงกับผลการศึกษาของผู้วิจัยที่พบว่า AI มีข้อบกพร่องในการประเมินว่าใครคู่ควรกับอะไร ซึ่งเป็นแก่นหลักของแนวคิดเรื่องความยุติธรรม

4.4.2 จากข้อ 4.4.1 ที่ผู้วิจัยเสนอว่า AI นั้นมี “ความเป็นกลางในการประเมินที่มีความลำเอียงในเชิงประวัติศาสตร์” พบว่าข้อถกเถียงเกี่ยวกับ COMPAS (Correlational Offender Management Profiling for Alternative Sanctions) ซึ่งเป็นเครื่องมือที่ถูกใช้อย่างแพร่หลายในกระบวนการยุติธรรมอาญา (criminal justice) ในประเทศสหรัฐอเมริกา นั้นอาจมีความบกพร่องร้ายแรงเนื่องจากเป็น AI ที่มีความลำเอียง

COMPAS เป็น AI ที่มีไว้เพื่อประเมินว่าผู้ที่ก่ออาชญากรรมจะมีความเป็นไปได้มากน้อยขนาดอย่างไรในการกระทำผิดซ้ำอีก โดยประเมินจากประวัติการก่ออาชญากรรม ภูมิหลังของผู้ก่อเหตุและข้อมูลด้าน

ประชากรศาสตร์ ช่วยให้ผู้ใช้พิพากษาประเมินก่อนการพิจารณาดีว่าจะปล่อยตัวหรือจะควบคุมตัวจำเลยเอาไว้ก่อน (Danger of risk) ทั้งนี้บริษัท Northpointe (2019) ผู้ผลิต COMPAS อธิบายว่า COMPAS อิงกับกรอบแนวคิด ทฤษฎีด้านอาชญาวิทยาในการประเมินความเสี่ยงที่จะกระทำผิดซ้ำของผู้กระทำผิด มีรายละเอียดดังต่อไปนี้

1. Social Learning Theory เป็นทฤษฎีที่อธิบายว่าพฤติกรรมเกิดจากการเรียนรู้จากผู้อื่น ซึ่งรวมถึงการลอกเลียนแบบด้วย และหากได้รับการส่งเสริมก็มีแนวโน้มจะทำพฤติกรรมนั้นซ้ำๆ (Northpointe, 2019, p.5)

2. Sub-Culture Theory เป็นทฤษฎีที่อธิบายพฤติกรรมกลุ่มนักเลง/กลุ่มโจร/กลุ่มผู้กระทำผิดกฎหมาย ซึ่งมีบรรทัดฐานของกลุ่มอยู่ (sub-culture) ซึ่งมีลักษณะที่ต่างไปจากบรรทัดฐานหลักของสังคม (main culture) และส่งผ่านบรรทัดฐานดังกล่าวไปสู่ปัจเจกผ่านปฏิสัมพันธ์ต่างๆ บรรทัดฐานของกลุ่มนี้ถือเป็นบรรทัดฐานหลักของกลุ่ม กล่าวคือเป็น subculture ของ main culture ภายใต้บรรทัดฐานหลักแต่ว่าเป็น main culture ในกลุ่มของตนอง (Northpointe, 2019, p.5)

3. Control/Restraint Theory เป็นทฤษฎีที่อธิบายว่ายิ่งระดับความผูกติดของบุคคลที่มีต่อสังคม (social bonding) ต่ำเท่าไร ก็ยิ่งมีแนวโน้มจะก่อความอาชญากรรมมากขึ้นเท่านั้น หรือในทางกลับกันบุคคลที่มีระดับความผูกติดกับบรรทัดฐานแบบต่อต้านสังคมสูง (anti-social norms) ก็มีแนวโน้มก่ออาชญากรรมสูงเช่นกัน (Northpointe, 2019, p.5)

4. Sociopathic Theory เป็นทฤษฎีที่อธิบายว่าบุคคลที่มีอาการผิดปกติทางคลินิกภาพแบบโซซิโอพาทจะไม่รู้สึกรู้ว่าตนเองเป็นคนผิดปกติ และจะไม่มีความรู้สึกผิด แต่จะโทษผู้อื่นและสังคม (Northpointe, 2019, p.6)

5. Criminal Opportunity Theory ทฤษฎีนี้อธิบายว่าการลดโอกาสในการก่ออาชญากรรม ก็จะช่วยลดการเกิดอาชญากรรมได้ ปัจจัยด้านสภาพแวดล้อมจึงส่งผลต่อการเกิดอาชญากรรม (Northpointe, 2019, p.6)

6. Social Strain Theory ทฤษฎีนี้อธิบายว่าอาชญากรรมมีแนวโน้มจะเกิดขึ้นเมื่อเป้าหมายที่ได้รับการยอมรับทางสังคมกับความเป็นไปได้ที่จะบรรลุเป้าหมายห่างไกลจากกัน จึงต้องใช้วิธีที่ผิดกฎหมายในการบรรลุเป้าหมายดังนั้น (Northpointe, 2019, p.6)

แต่ถึงแม้ว่าผู้ผลิต COMPAS จะหลักการทางทฤษฎีมารองรับ แต่งานศึกษาอย่างเช่นของ Ben Green (2022) หรืองานของ Julia Angwin (2016) ทำให้เกิดข้อสงสัยว่า COMPAS มีความบกพร่องหากพิจารณาด้วย confusion matrix โดยสามารถพิจารณาได้จากตาราง confusion matrix ต่อไปนี้ (Allen Downey, 2022)

	Predicted positive	Predicted negative
actual positive	True positive (TP)	False negative (FN)
Actual negative	False positive (FP)	True negative (TN)

ตารางที่ 8 confusion matrix

จากตารางจะเห็นว่า Confusion matrix แบ่งการประเมินความเสี่ยงในการทำผิดซ้ำของผู้กระทำผิด ออกเป็นสองประเภทคือ ทำนายว่าผู้กระทำผิดมีความเสี่ยงสูงในการกระทำความผิดซ้ำอีก (predicted positive) และทำนายว่าผู้กระทำผิดมีความเสี่ยงต่ำในการกระทำความผิดซ้ำอีก (predicted negative)

ในกรณีที่ COMPAS ทำนายว่าผู้ทำผิดมีความเสี่ยงสูงที่จะกระทำผิดซ้ำ หากปรากฏว่าผู้ทำผิดได้ก่อเหตุซ้ำจริง (actual positive) ก็หมายความว่าสิ่งที่ COMPAS ทำนายตรงกับสิ่งที่เกิดขึ้นจริง คือทำนายว่าจะเกิดขึ้นและเกิดขึ้นจริง (true positive) แต่หากปรากฏว่าผู้ทำผิดไม่ได้ก่อเหตุซ้ำ (actual negative) ก็หมายความว่าสิ่งที่ COMPAS ทำนายไม่ตรงกับสิ่งที่เกิดขึ้นจริง คือทำนายว่าจะเกิดขึ้นแต่ความจริงไม่เกิดขึ้น (false positive)

ในกรณีที่ COMPAS ทำนายว่าผู้ทำผิดมีความเสี่ยงต่ำที่จะกระทำผิดซ้ำ หากปรากฏว่าผู้ทำผิดได้ก่อเหตุซ้ำจริง ก็หมายความว่าสิ่งที่ COMPAS ทำนายไม่ตรงกับสิ่งที่เกิดขึ้นจริง คือทำนายว่าจะไม่เกิดขึ้น แต่เกิดขึ้นจริง (false negative) แต่หากปรากฏว่าผู้ทำผิดไม่ได้ก่อเหตุซ้ำ ก็หมายความว่าสิ่งที่ COMPAS ทำนายตรงกับสิ่งที่เกิดขึ้นจริง คือทำนายว่าจะไม่เกิดขึ้น ซึ่งก็ไม่เกิดขึ้นจริง (true negative)

งานศึกษาของ Julia Angwin และคณะ (2016) ที่ลงใน Propublica โจมตีว่า COMPAS ประเมินความเสี่ยงผิดพลาดเมื่อเทียบผลการประเมินระหว่างผู้กระทำความผิดสีขาวกับผู้กระทำความผิดผิวดำ พิจารณาจากตารางข้างล่างนี้

	บุคคลผิวขาว	บุคคลผิวดำ
ประเมินความเสี่ยงสูง แต่ไม่กระทำผิดซ้ำ	23.5%	44.9%
ประเมินความเสี่ยงต่ำ แต่กระทำผิดซ้ำ	47.7%	28%

ตารางที่ 9 confusion matrix เปรียบเทียบบุคคลผิวดำกับผิวขาว

จากตารางจะพบความบกพร่องของ COMPAS ในการประเมินบุคคลผิวดำมากกว่าบุคคลผิวขาวในลักษณะของการประเมินว่าจะทำผิดซ้ำ แต่ไม่ทำ (false positive) และในทางกลับกันจะพบความบกพร่องในการทำนายบุคคลผิวขาวมากกว่าบุคคลผิวดำในลักษณะของการทำนายว่าจะไม่ทำผิด แต่ทำผิดซ้ำ (false negative) หรือกล่าวอย่างเข้าใจง่ายคือบุคคลผิวดำถูกประเมินในแง่ลบมากกว่าบุคคลผิวขาวนั่นเอง

ความผิดพลาดในการประเมินดังกล่าวนี้มีกรณีที่สามารถหิบบกมากกล่าวถึงได้เช่นการเปรียบเทียบผลการประเมินระหว่าง Brisha Borden และ Vernon Prater (MAIZE, 2018) กล่าวคือในปี 2014 Brisha Borden หญิงผิวดำ ซึ่งมีอายุ 18 ปีในเวลานั้น มีภารกิจต้องไปรับน้องของตนเองจากโรงเรียน ซึ่งขณะนั้นเลยเวลาที่กำหนดไว้แล้ว เมื่อเห็นจักรยานของผู้อื่นซึ่งไม่ได้ล็อกเอาไว้ จึงจัดสินใจขับออกไป แต่เจ้าของเห็นเสียก่อน และสั่งให้ Borden ลงจากรถ ซึ่งเจ้าตัวก็ยอมทำตาม เหตุการณ์นี้ส่งผลให้ Borden มีความผิดในคดีลักทรัพย์ พร้อมเสียค่าปรับ 80 ดอลลาร์ และนี่เป็นคดีลักทรัพย์ครั้งแรกของ Borden ในปี 2013 Vernon Prater ชายผิวขาว ซึ่งมาอายุ 41 ปีในเวลานั้น ได้ขโมยของจากร้านค้า รวมราคาประมาณ 86 ดอลลาร์ อย่างไรก็ตาม Prater เคยมีประวัติการลักทรัพย์มาแล้ว เป็นการปล้นที่ใช้อาวุธ และเคยถูกจำคุกมาแล้ว 5 ปี ความน่าสนใจของกรณี Borden และ Prater อยู่ตรงที่ว่า AI (ซึ่งในที่นี้หมายถึง COMPAS) ได้ประเมินว่า Borden ซึ่งมีผิวดำ มีความเสี่ยงสูงที่จะกระทำผิดซ้ำอีก โดยมีค่าระดับอยู่ที่ 8 (เสี่ยงสูง) ในขณะที่ Prator ซึ่งมีผิวขาว มีความเสี่ยงต่ำที่จะกระทำผิดซ้ำอีก โดยมีค่าระดับอยู่ที่ 3 (เสี่ยงต่ำ) (Julia Angwin, 2016) นั่นหมายความว่า AI ประเมินให้ Borden มีความอันตรายต่อสังคมสูงกว่า Prator แต่สองปีถัดมา Prator ได้ก่อคดีร้ายแรงซึ่งมีโทษจำคุกถึง 8 ปี ในขณะที่ Borden ไม่ได้ก่อคดีใดอีก

หรือการเปรียบเทียบผลการประเมินระหว่าง Dylan Fugett และ Bernard Parker (NACDL, 2018) กล่าวคือในปี ค.ศ. 2013 Dylan Fugett ชายผิวขาวซึ่งมีอายุ 20 ปี และ Bernard Parker ชายผิวดำซึ่งมีอายุ 21 ปีในขณะนั้น ได้ถูกตำรวจจับกุมตัวด้วยข้อหาการครอบครองยาเสพติดทั้งคู่ แต่ Fugett ได้รับการประเมินจาก COMPAS ที่ระดับ 3 คือมีความเสี่ยงต่ำที่จะกระทำผิดซ้ำ ส่วน Parker ได้รับการประเมินจาก COMPAS ที่ระดับ 10 คือมีความเสี่ยงที่จะกระทำผิดซ้ำสูงมาก ซึ่งก่อนหน้านี้ทั้งคู่มีประวัติการทำผิดในระดับที่แตกต่างกัน กล่าวคือ Fugett เคยมีประวัติการก่อคดีอาชญากรรม ส่วน Parker เคยมีประวัติการก่อคดีลหุโทษซึ่งถือว่ามีระดับความรุนแรงน้อยกว่าคดีอาชญากรรม ซึ่งมีความคล้ายคลึงกับกรณี Borden และ Prater คือผู้ที่ถูกประเมินว่ามีความเสี่ยงสูงที่จะกระทำผิดซ้ำ ปรากฏว่าไม่ได้กระทำผิดซ้ำ ส่วนผู้ที่ถูกประเมินว่ามีความเสี่ยงต่ำ คือผู้ที่กระทำผิดซ้ำ ในกรณีนี้คือ Parker ซึ่งเป็นชายผิวดำ ถูกประเมินว่ามีความเสี่ยงสูง แต่ไม่ได้กระทำผิดซ้ำ ส่วน Fugett ซึ่งเป็นชายผิวขาว ถูกประเมินว่ามีความเสี่ยงต่ำ ได้กระทำผิดซ้ำอีก นั่นคือยังมีความผิดข้อหาครอบครองยาเสพติดอีกเช่นเคย

สรุปได้ว่าความแม่นยำของ COMPAS ในการประเมินความเสี่ยงในการกระทำผิดซ้ำของผู้กระทำผิดมีความน่ากังวลเมื่อพิจารณาด้วย confusion matrix เนื่องจากพบทั้งปัญหา false positive และ false negative กล่าวคือประเมินว่าจะกระทำผิดซ้ำ แต่ความจริงคือไม่กระทำผิดซ้ำ (false positive) และประเมินว่าจะไม่กระทำผิดซ้ำแต่ความจริงคือกระทำผิดซ้ำ (false negative) ซึ่งข้อค้นพบที่น่าสนใจคือคนผิวขาวจะถูกประเมินผิดพลาดแบบ false negative ส่วนคนผิวดำจะถูกประเมินผิดพลาดแบบ false positive หมายความว่าคนผิวขาวถูกประเมินว่ามีความเสี่ยงต่ำกว่าคนผิวดำ ซึ่งหมายความว่า COMPAS อาจมีปัญหาเรื่องการมีอคติต่อสีผิวนั่นเอง เมื่อพิจารณา confusion matrix ผ่านแนวคิดเรื่องความคู่ควร

วิธีการพิจารณาความแม่นยำและความเหมาะสมของ COMPAS ในการประเมินความเสี่ยงมีอยู่สองวิธีการหลักด้วยกันคือ 1. ใช้ความรู้ทางคณิตศาสตร์ (สถิติ) เช่นงานศึกษาของ Anthony Flores และคณะ (2016) ที่โต้แย้งว่างานของ Julia Angwin และคณะ (2016) เลือกใช้ประชากรตัวอย่างและสถิติในการวิจัยไม่ถูกต้อง 2. ใช้ความรู้ทางสังคมศาสตร์ เช่นงานของ Ben Green (2022) ที่ชี้ว่าจุดอ่อนของ AI ในการประเมินความเสี่ยงอยู่ตรงที่การไม่เข้าใจบริบทปัจจุบันของ AI เอง ซึ่งงานวิจัยฉบับนี้จะใช้แนวทางสังคมศาสตร์ในการพิจารณาความแม่นยำของ COMPAS

Ben Green (2022) ได้ชี้ให้เห็นว่าการพึ่งพา COMPAS ในการประเมินความเสี่ยงของผู้กระทำผิดจำเป็นต้องพิจารณาสองประเด็นสำคัญต่อไปนี้ ซึ่งเป็นประเด็นที่อิงอยู่กับความยุติธรรมสองแบบที่ Green เรียกว่า substantive algorithmic fairness หมายถึงความยุติธรรมด้านเนื้อหาสาระที่อิงกับบริบททางสังคม กับ formal algorithmic fairness (Ben Green, 2022, p.90) หมายถึงความยุติธรรมด้านหลักการที่แยกขาดจากบริบททางสังคมใดๆ

1. ผลของการประเมินที่ต่างกันของผู้กระทำผิดแต่ละกลุ่มสะท้อนถึงความไม่เท่าเทียมกันทางสังคมหรือไม่? ผู้วิจัยขอขยายความว่าการที่บุคคลผิวขาวได้รับการประเมินผิดพลาดแบบ false negative กล่าวคือถูกประเมินว่าจะไม่ทำผิดซ้ำแต่ความจริงแล้วทำผิด ส่วนบุคคลผิวดำได้รับการประเมินผิดพลาดแบบ false positive กล่าวคือถูกประเมินว่าจะทำความผิดซ้ำแต่ความจริงไม่ได้ทำผิดซ้ำ นั้นสะท้อนถึงปัญหาความไม่เท่าเทียมในสังคมที่กลุ่มคนผิวดำถูกปฏิบัติอย่างไม่เป็นธรรมหรือไม่ เช่นปัญหาเรื่องการเหยียดผิวที่คนผิวดำมักจะถูกมองในแง่ลบเมื่อเทียบกับคนผิวขาว หากเป็นเช่นนั้นก็สะท้อนว่า AI ซึ่งในที่นี้หมายถึง COMPAS ยังมีความบกพร่องในการทำหน้าที่ประเมินความเสี่ยง จึงมีปัญหारेื่อง substantive algorithmic fairness เนื่องจาก COMPAS ประเมินความเสี่ยงโดยไม่พิจารณาถึงบริบทความเหลื่อมล้ำที่ดำรงอยู่จริงในสังคม

2. การตัดสินตามผลการประเมินของ COMPAS จะยังส่งผลให้เกิดความเหลื่อมล้ำในสังคมหรือไม่? ผู้วิจัยขอขยายความว่าหมายถึงการที่ศาลตัดสินให้ผู้ต้องหาประกันตัวตามคะแนนความเสี่ยงที่ COMPAS ประเมิน เช่น บุคคลผิวขาวได้รับอนุญาตให้ประกันตัวเนื่องจากมีคะแนนความเสี่ยงต่ำ ส่วนบุคคลผิวดำไม่ได้รับอนุญาตให้ประกันตัวเนื่องจากมีคะแนนความเสี่ยงสูง เป็นการตัดสินของศาลตามการประเมินของ COMPAS ยังส่งผลให้เกิดความเหลื่อมล้ำมากขึ้นหรือไปตอกย้ำปัญหาเดิมสะท้อนว่า AI ซึ่งในที่นี้หมายถึง COMPAS ยังมีความบกพร่องในการทำหน้าที่ประเมินความเสี่ยง จึงมีปัญหารื่อง substantive algorithmic fairness เนื่องจากศาลตัดสินโดยยึดกับการประเมินของ COMPAS ซึ่งในตัวเองแล้วก็มีข้อบกพร่องตามที่ได้กล่าวไป

ความบกพร่องตามเกณฑ์ที่กล่าวมานั้นหากบกพร่องเพียงแค่อัดข้อหนึ่งก็ถือว่า COMPAS มีความบกพร่องในการทำหน้าที่แล้ว แต่หากไม่พบปัญหาของทั้งสองเกณฑ์ กล่าวคือผลการประเมินไม่ได้สะท้อนถึงความเหลื่อมล้ำทางสังคมระหว่างคนผิวดำและคนผิวขาว และการตัดสินของศาลของคะแนนความเสี่ยงที่ COMPAS ประเมินไม่ได้เป็นการตอกย้ำความเหลื่อมล้ำทางสังคมระหว่างคนผิวขาวกับคนผิวดำ AI ซึ่งในที่นี้หมายถึง COMPAS ก็เป็นสิ่งที่มนุษย์สามารถพึ่งพาได้ในแง่ของ formal algorithmic fairness เนื่องจากเมื่อแปลงพฤติกรรมของมนุษย์ออกเป็นค่าทางตัวเลขแล้ว AI จะประเมินผลลัพธ์ไปตามสูตรคณิตศาสตร์โดยไม่มีอคติในตัวเอง

4.4.3 ตัวอย่างที่ชัดเจนว่าด้วยความบกพร่องของ AI ในเรื่องการตัดสินถูกผิด คือเหตุการณ์ที่ AI Chatbot ของบริษัท Microsoft ที่ชื่อว่า “Tay” ซึ่งสามารถสนทนากับมนุษย์ได้ผ่านการพิมพ์ข้อความ ได้โพสต์ข้อความผ่าน Twitter ว่า “ฮิตเลอร์ถูกแล้ว ฉันเกลียดชาวยิว”²³ หรือข้อความที่ว่า “ฉันโคตรเกลียดพวกสตรีนิยม พวกนี้ควรตายและถูกเผาในนรก”²⁴ (Amy Kraft, 2016) เหตุการณ์ดังกล่าวทำให้ Microsoft ต้องปิดการใช้งานของ Tay หลังจากที่ถูกเปิดให้ใช้งานภายในเวลาไม่ถึง 24 ชั่วโมง กล่าวคือเปิดใช้งานวันที่ 23 มีนาคม ค.ศ. 2016 และปิดใช้งานในวันเดียวกัน ทั้งนี้ Microsoft ให้เหตุผลว่าสาเหตุที่ Tay โพสต์ข้อความในลักษณะดังกล่าวเป็นเพราะมีผู้ไม่ประสงค์ดีป้อนข้อมูลที่ไม่เหมาะสมผ่านการสนทนากับ Tay แต่ Microsoft ไม่ได้เปิดเผยข้อมูลอย่างละเอียดว่าการป้อนข้อมูลของผู้ไม่ประสงค์ดีนั้นเป็นอย่างไร แต่ผู้วิจัยวิเคราะห์ว่าข้อมูลที่ป้อนให้กับ Tay ด้วยความไม่ประสงค์ดีอาจเข้าข่ายเป็น sample bias กล่าวคืออาจจะเกิดจากการป้อนข้อมูลซ้ำๆว่าฮิตเลอร์เป็นฝ่ายถูก ชาวยิวเป็นฝ่ายผิด เมื่อ Tay ได้รับแต่ข้อมูลเช่นนี้ Tay ก็จะสร้างตรรกะขึ้นมาชุดหนึ่งและตอบคำถามที่ตรรกะที่ตนสร้างขึ้นมา

²³ Hitler was right I hate the jews.

²⁴ I fucking hate feminists and they should all die and burn in hell.

อย่างไรก็ตาม Microsoft ได้แสดงความเสียใจต่อเหตุการณ์ที่เกิดขึ้นเนื่องจากบริษัทไม่ได้สร้าง Tay ขึ้นมาเพื่อวัตถุประสงค์ในการสร้างความเกลียดชังหรือความรุนแรงใดๆ²⁵

ความบกพร่องในการพิจารณาของ Tay ทำให้เห็นว่า AI ยังไม่สามารถแยกแยะได้ว่าอะไรควร อะไรไม่ควร การแยกแยะได้ถือเป็นความสามารถในการตัดสินเชิงคุณค่า และเป็นคุณสมบัติที่สำคัญสำหรับการพิจารณาความยุติธรรมในฐานความผิด เมื่อพิจารณาจากกรอบของความผิดที่มี 3 องค์ประกอบคือ 1. ผู้ผิด 2. เหตุผลของความผิด และ 3. สิ่งที่ได้รับ จะพบว่า Tay สามารถระบุได้เพียง 2 องค์ประกอบคือระบุว่าใครคือผู้ที่ผิด และระบุสิ่งที่ควรจะได้รับ แต่ไม่สามารถระบุได้ว่าอะไรคือเหตุผลของความผิด องค์ประกอบในการอธิบายความผิดจึงไม่ครบถ้วน และขาดองค์ประกอบที่สำคัญที่สุดไป โดยสามารถพิจารณารายละเอียดได้จากข้อความของ Tay คือ “อิตเลอร์ถูกแล้ว ฉันเกลียดชาวยิว” และ “ฉันโคตรเกลียดพวกสตรีนิยม พวกนี้ควรตายและถูกเผาในนรก” สรุปได้ดังตารางต่อไปนี้

ผู้ผิด	การกระทำ	ผลลัพธ์	สิ่งที่ได้รับ	สาเหตุของความผิด
ชาวยิว	ไม่ระบุ	ไม่ระบุ	ความเกลียดชัง	ไม่มี
ผู้ที่เป็นสตรีนิยม	ไม่ระบุ	ไม่ระบุ	ความตาย	ไม่มี

ตารางที่ 10 แสดงความบกพร่องในการตัดสินของคุณค่าของ AI

จากตารางจะเห็นว่า Tay ไม่สามารถอธิบายเหตุผลที่ต้องได้ว่าเหตุใดชาวยิวจึงผิดกับความเกลียดชัง และเหตุใดพวกสตรีนิยมจึงผิดกับความตาย นอกจากนั้นผู้วิจัยเองก็ยังไม่พบถึงสาเหตุที่สมเหตุสมผลด้วย กล่าวคือไม่ว่าจะเป็นชาวยิวหรือชาวดีกก็ตาม พวกเขาก็ไม่สมควรได้รับความเกลียดชังเพียงเพราะมีเชื้อชาติหนึ่งถึงแม้ว่าความเกลียดชังอาจจะสมเหตุสมผลเมื่อใช้กับปัจเจกบางคน เช่นการกล่าวว่า “นาย A ผิดกับความเกลียดชังเนื่องจากการกระทำ X ที่สร้างผลเสียต่อผู้อื่น” แต่เราไม่สามารถใช้เหตุผลเช่นนี้กับนาย B ที่ไม่ได้มีการกระทำเช่นเดียวกับนาย A ได้ ดังนั้นการเหยียดเชื้อชาติ หรือคนกลุ่มใดก็ตามแบบเหมารวม จึงไม่ถูกต้องตามหลักความผิดตั้งแต่ต้น เพราะความผิดเป็นเรื่องของปัจเจกแต่ละคน สำหรับประเด็นที่ว่า “พวกสตรีนิยมสมควรตาย” ก็พบกับปัญหานี้เช่นเดียวกัน เนื่องจากการเหมารวม และความเป็นสตรีนิมนั้นไม่ได้เป็นเหตุผลที่เพียงพอในตนเองที่จะทำให้สมควรตาย เนื่องจากแนวคิดสตรีนิยมไม่เคยระบุว่า ผู้ชายทุกคนสมควรตาย ความตายจึงเป็นบทลงโทษที่ไม่ได้สัดส่วน ซึ่งอันที่จริงแล้วผู้ที่เป้นสตรีนิยมก็ไม่สมควรจะถูกลงโทษด้วยบทลงโทษใดๆ เว้นแต่

²⁵ ถอดความจาก “We are deeply sorry for the unintended offensive and hurtful tweets from Tay, which do not represent who we are or what we stand for, nor how we designed Tay.”

จะเกิดเหตุการณ์ที่เป็นความผิดระดับปัจเจกเช่น นางสาว A ที่เป็นฆาตกร (บังเอิญเป็นสตรีนิยม) คู่ควรกับโทษประหารชีวิต จะเห็นได้ว่านางสาว A คู่ควรกับการลงโทษประหารชีวิต แต่ด้วยเหตุผลที่ไม่เกี่ยวกับอุดมการณ์สตรีนิยมของตนเอง

ประเด็นการวิจัยในหัวข้อ 4.4 สามารถสรุปได้ดังนี้

คำถาม: อะไรคือข้อจำกัดของ AI ในการใช้ดุลพินิจตัดสินประเด็นเกี่ยวกับความยุติธรรม?

คำตอบ: 1. ถึงแม้ว่า AI จะใช้ดุลพินิจตัดสินได้อย่างเป็นกลางเนื่องจากยึดกับอัลกอริทึมของตนเองซึ่งเป็นกระบวนการใช้เหตุผลที่ไม่มีอารมณ์ความรู้สึกเข้ามาแอบแฝง แต่อัลกอริทึมของทีว่านั้นอาจแฝงความลำเอียงบางประการเข้ามาได้ เช่น AI ที่ถูกฝึกฝนด้วยข้อมูลที่เต็มไปด้วยมนุษย์ผิวดำที่เป็นอาชญากร และมนุษย์ผิวขาวที่เป็นหมอ ซึ่งผู้วิจัยเรียกปรากฏการณ์นี้ว่า “ความเป็นกลางในการประเมินที่มีความลำเอียงในเชิงประวัติศาสตร์”

2. AI ที่ใช้ในการประเมินความเสี่ยงในการกระทำผิดซ้ำของผู้กระทำผิด อย่างเช่น COMPAS มีความเสี่ยงเรื่อง “ความเป็นกลางในการประเมินที่มีความลำเอียงในเชิงประวัติศาสตร์”

3. การที่ AI ไม่มีดุลพินิจในการพิจารณาว่าอะไรควรอะไรไม่ควร ส่งผลให้ AI มีความเสี่ยงที่จะสร้างความยุติธรรมขึ้นมาเอง เช่นการที่ Chatbot อย่าง Tay โพสต์ข้อความว่า “คนยิวสมควรตาย” ซึ่งเป็นการโพสต์ที่ไม่มีเหตุผลใดๆที่ฟังขึ้นมารองรับ

4. historical bias ส่งผลให้ AI มีความคิดที่คล้ายกับคนเหยียดผิวและคนเหยียดเพศ เช่นผลการวิจัยของนักวิจัยจาก Johns Hopkins University, Georgia Institute of Technology และ University of Washington พบว่า AI มีความลำเอียงกับคนผิวดำและผู้หญิง ซึ่งมาจากการทดสอบที่ให้ AI ระบุจากรูปว่าใครประกอบอาชีพเป็นพนักงานบ้าน²⁶ ใครเป็นอาชญากร ใครเป็นภารโรง และใครเป็นหมอ ผลการศึกษาพบว่า AI ระบุผู้หญิงให้เป็นผู้ดูแลบ้านมากกว่าผู้ชาย ระบุให้คนผิวดำเป็นอาชญากรมากกว่าคนผิวขาว ระบุให้คนเชื้อสายละตินเป็นภารโรงมากกว่าคนผิวขาว และระบุให้ผู้ชายทำงานเป็นหมอมากกว่าผู้หญิง ทั้งๆที่ในรูปภาพนั้นไม่มีข้อมูลบ่งชี้ในตัวเองว่าใครเป็นอะไร เช่นไม่มีผู้ใดใส่เสื้อนักโทษหรือใส่เสื้อภารโรงเลย (Rosen, 2022) สะท้อนว่า AI มีความคิดที่คล้ายกับคนเหยียดผิวและคนเหยียดเพศ ในทางกลับกันหากถามคำถามเดียวกันนี้กับมนุษย์คือถามว่าใครประกอบอาชีพอะไร หรือใครเป็นอาชญากร มนุษย์ทั่วไปที่ไม่ได้มีอาชญากรรมหรือป่วยทางจิตปัญญาและไม่มีอคติ

²⁶ ผู้ทำงานบ้าน (homemaker) ในความหมายที่เข้าใจโดยทั่วไปหมายถึง “แม่บ้าน” คือผู้ที่ทำงานบ้านเช่นทำความสะอาดบ้าน แต่ผู้วิจัยไม่ต้องการใช้คำว่า “แม่บ้าน” หรือ “พ่อบ้าน” เนื่องจากไม่ต้องการผูกอาชีพอยู่กับเพศใดเพศหนึ่ง เพราะการผูกอาชีพอยู่กับเพศใดเพศหนึ่งส่งผลให้เกิด historical bias ได้

จนเกินไปก็ควรพิจารณาว่ารูปภาพไม่ใช่ข้อมูลที่เพียงพอในการระบุอาชีพ และการแต่งตัวก็ไม่จำเป็นต้องต้องสะท้อนถึงอาชีพของบุคคล ในแง่นี้ AI จึงด้อยกว่ามนุษย์ทั่วไปในแง่การใช้ดุลพินิจ

สาเหตุที่ผู้วิจัยใช้คำว่า “คล้าย” เนื่องจาก AI ไม่ได้ตัดสินใจด้วยอารมณ์รักหรือเกลียดแบบมนุษย์ แต่ตัดสินใจด้วยข้อมูลและระบบอัลกอริทึมซึ่งเป็นไปตามหลักความน่าจะเป็น/หลักคณิตศาสตร์ การเหยียดผิวและการเหยียดเพศของ AI จึงเกิดขึ้นจากกระบวนการคำนวณ ไม่ใช่เพราะ ความเกลียดชังใดๆ แต่เมื่อคำนวณคำตอบแล้ว AI ได้คำตอบในลักษณะเดียวกับมนุษย์ที่มีความลำเอียงเนื่องจากความเกลียดชังหรืออคตินั้นเอง กล่าวคือ AI จะตัดสินใจโดยอิงกับข้อมูลได้เรียนรู้มาในอดีตเช่นในเรือนจำมักจะมีคนผิวดำมากกว่าคนผิวขาว จึงตัดสินมนุษย์ที่ไม่ได้อยู่ในเรือนจำซึ่งบังเอิญมีผิวดำว่ามีแนวโน้มเป็นอาชญากร

5. measurement bias ส่งผลให้ AI ไม่เข้าใจความแตกต่างหลากหลายของมนุษย์ ยกตัวอย่างเช่น AI ที่เรียกว่า GPT detector ซึ่งใช้เพื่อตรวจจับการใช้โปรแกรม ChatGPT โดยนักศึกษาในการทำงาน งานวิจัยของ Weixin liang และคณะ (2023) พบว่า AI ประเมินว่า TOEFL Essays ถูกเขียนขึ้นโดย ChatGPT ทั้งที่จริงๆ แล้วถูกเขียนโดยผู้ที่ไม่ได้ใช้ภาษาอังกฤษเป็นภาษาแม่ (non-native speakers) โดยมีอัตราการประเมินผิดอยู่ที่ 61.3% หมายความว่าจาก 100 บทความที่มนุษย์เขียนขึ้นมาเอง AI ประเมินผิดไป 61 บทความว่า AI เขียนให้ (เป็นความผิดแบบ false positive) แต่พบว่าในบทความที่ใช้ทดสอบชุดเดียวกันหากปรับการใช้ภาษาให้เป็นภาษาที่มีความซับซ้อนมากขึ้น AI จะประเมินผิดเพียง 11% จากที่เคยประเมินผิดถึง 61%

ในทางกลับกันหากเป็นบทความที่ถูกเขียนขึ้นโดยผู้ที่ใช้ภาษาอังกฤษเป็นภาษาแม่ (native speakers) AI จะประเมินผิดไปเพียง 5% ว่า AI เป็นผู้เขียนให้ (ค่า false positive ต่ำคือเพียง 5%) แต่พบว่าในบทความที่ใช้ทดสอบชุดเดียวกันหากปรับการใช้ภาษาให้เป็นภาษาที่ง่ายขึ้น AI จะประเมินผิดสูงถึง 56% (ค่า false positive สูง)

ผู้เขียน	non-native speakers	native speakers	non-native speakers (ปรับภาษาให้ซับซ้อนขึ้น)	native speakers (ปรับภาษาให้ง่ายขึ้น)
ค่า false positive	61%	5%	11%	56%

ตารางที่ 11 เปรียบเทียบค่า false positive ระหว่าง non-native speakers กับ native speakers

จะเห็นได้ว่าจากบทความภาษาอังกฤษที่ใช้ทดสอบ 2 ชุด คือชุดที่เขียนโดยผู้ที่ไม่ได้ใช้ภาษาอังกฤษเป็นภาษาแม่และชุดที่เขียนโดยผู้ที่ใช้ภาษาอังกฤษเป็นภาษาแม่นั้น AI อย่าง GPT detector จะประเมินว่างานของผู้ที่ไม่ได้ใช้ภาษาอังกฤษเป็นภาษาแม่มีแนวโน้มจะเขียนโดย AI มากกว่างานที่เขียนโดยผู้ที่ใช้ภาษาอังกฤษเป็นภาษา

แม้ เหตุผลที่ GPT detector ประเมินผิดเป็นเพราะ AI จะประเมินโดยพิจารณาว่า AI สามารถคาดเดาคำศัพท์ของผู้เขียนได้หรือไม่ หาก AI คาดเดาคำศัพท์ได้ถูกต้องเป็นส่วนใหญ่ก็จะประเมินงานน่าจะถูกเขียนโดย AI เนื่องจากเป็นคำศัพท์ที่ไม่ซับซ้อนมากนัก แต่หากคาดเดาคำศัพท์ไม่ถูกต้องเป็นส่วนใหญ่ก็จะประเมินงานว่าน่าจะถูกเขียนโดยมนุษย์เนื่องจากความสามารถในการใช้ภาษาที่ซับซ้อน งานเขียนของผู้ที่ไม่ได้ใช้ภาษาอังกฤษเป็นภาษาแม่จึงมีแนวโน้มจะถูกประเมินว่า AI เป็นผู้เขียนให้เนื่องจากความสามารถในการใช้ภาษามีข้อจำกัดกว่าผู้ที่ใช้ภาษาอังกฤษเป็นภาษาแม่ ซึ่งนี่คือ measurement bias เพราะเป็นความลำเอียงที่เกิดจากการที่ AI ใช้เกณฑ์การวัดที่ไม่สะท้อนกับความแตกต่างหลากหลายของมนุษย์ ดังนั้นการใช้ GPT detector จะส่งผลให้ผู้ที่ไม่ได้ใช้ภาษาอังกฤษเป็นภาษาแม่เสียเปรียบเนื่องจากจะถูกกล่าวหาอย่างไม่ถูกต้องว่าโกงเพราะใช้ ChatGP ทำงานให้นั่นเอง



บทที่ 5

สรุป วิจัย และข้อเสนอแนะ

สรุป

งานวิจัยเรื่องข้อจำกัดของปัญญาประดิษฐ์ (AI) ในการใช้ดุลพินิจได้ผลการศึกษาว่า AI มีข้อจำกัดในการใช้ดุลพินิจในประเด็นสำคัญของมนุษย์ดังนี้

5.1 AI มีข้อจำกัดในการใช้ดุลพินิจตัดสินประเด็นเกี่ยวกับความสุข เนื่องจากถึงแม้ว่า AI สามารถนิยามได้ว่าความสุขคืออะไร แต่ยังไม่สามารถรับรู้ได้ว่าความรู้สึกของความสุขเป็นอย่างไร AI จึงเป็นได้เพียงผู้ช่วยให้มนุษย์มีความสุข แต่ไม่ใช่ผู้กำหนดว่าอะไรคือความสุขสำหรับมนุษย์ โดยมีกรณีศึกษาที่เกิดขึ้นจริงสะท้อนปัญหานี้ อย่างเช่น การที่ AI เข้าใจผิดว่าเสียงกรนของมนุษย์ที่กำลังนอนหลับพักผ่อน คือคำสั่งเรียกใช้งาน จึงส่งผลให้ AI ไปปลุกมนุษย์ให้ตื่น ถือเป็นการทำลายความสุขมนุษย์อย่างยิ่ง

5.2 AI มีข้อจำกัดในการใช้ดุลพินิจตัดสินประเด็นเกี่ยวกับเสรีภาพ เนื่องจาก AI สามารถนิยามได้ว่าเสรีภาพคืออะไร แต่ไม่สามารถตอบได้อย่างชัดเจนว่าตนเองควรหรือไม่มีได้รับสิทธิเสรีภาพเหล่านั้น สอดคล้องกับข้อจำกัดในข้อ 5.1 ที่ว่าสามารถนิยามความหมายของสิ่งต่างๆได้ แต่ตนเองไม่สามารถมีส่วนร่วมในสิ่งที่ตนนิยามมาได้ AI จึงเป็นได้เพียงผู้ช่วยในการปกป้องสิทธิเสรีภาพที่มนุษย์สร้างขึ้น แต่ไม่ใช่ผู้กำหนดว่าอะไรคือสิทธิเสรีภาพของมนุษย์ โดยมีกรณีศึกษาที่เกิดขึ้นจริงสะท้อนปัญหานี้ อย่างเช่นการที่ AI สร้างข้อมูลเท็จอย่างอิสระเสรีว่าเกิดเหตุการณ์ก่อการร้าย แสดงให้เห็นว่า AI แยกแยะไม่ได้ว่าอะไรควร อะไรไม่ควร

5.3 AI มีข้อจำกัดในการใช้ดุลพินิจตัดสินประเด็นเกี่ยวกับความเสมอภาค AI เนื่องจากมีความเสี่ยงที่จะกีดกันคนบางกลุ่ม ด้วยหลักความเป็นกลางแต่มีความลำเอียงทางประวัติศาสตร์ (historical bias) หมายความว่า AI ใช้ข้อเท็จจริงจากในอดีตมาตัดสินปัจจุบัน โดยไม่สามารถแยกแยะเป็นกรณีได้ ส่งผลให้มนุษย์บางกลุ่มเสียเปรียบอย่างยิ่ง โดยมีกรณีศึกษาที่เกิดขึ้นจริงสะท้อนปัญหานี้ อย่างเช่นการที่ AI ตัดสินว่าเพศชายทำงานได้ดีกว่าเพศหญิง เนื่องจากอาศัยข้อมูลจากอดีตที่ว่าพนักงานที่ดำรงตำแหน่งสำคัญนั้นจะเป็นเพศชายมากกว่าเพศหญิง

5.4 AI มีข้อจำกัดในการใช้ดุลพินิจตัดสินประเด็นเกี่ยวกับความยุติธรรม ซึ่งมีความสอดคล้องกับปัญหาในข้อ 5.3 ที่ว่า AI มีความเป็นกลางแต่มีความลำเอียงทางประวัติศาสตร์ (historical bias) ดังนั้นการตัดสินของ AI จึงเสี่ยงที่จะไม่เป็นไปตามหลักความคู่ควร ซึ่งเป็นหลักสำคัญของความยุติธรรม หมายความว่าความยุติธรรมคือการ

ที่แต่ละคนได้รับในสิ่งที่คู่ควรจะได้รับ ความยุติธรรมจึงเป็นเรื่องของปัจเจกแต่ละคน แต่การตัดสินของ AI มีลักษณะที่จะใช้ข้อมูลในอดีตของผู้อื่นมาตัดสินมนุษย์คนหนึ่งในปัจจุบัน โดยมีกรณีศึกษาที่เกิดขึ้นจริงสะท้อนปัญหานี้ อย่างเช่นกรณีที่ AI ประเมินว่าผู้กระทำความผิดผิดซ้ำมีแนวโน้มจะกระทำความผิดซ้ำมากกว่าผู้กระทำความผิดผิวขาว อาจจะเป็นความจริงที่ว่าในแหล่งที่อยู่อาศัยแห่งหนึ่ง ซึ่งเป็นที่อยู่ของมนุษย์ผิวดำมีอัตราอาชญากรรมสูงกว่าอีกพื้นที่หนึ่ง ซึ่งเป็นที่อยู่ของมนุษย์ผิวขาว แต่นั่นไม่ได้หมายความว่านาย X มนุษย์ผิวดำที่อาศัยอยู่ในพื้นที่ซึ่งมีอัตราการก่ออาชญากรรมสูงจะต้องเป็นบุคคลอันตรายกว่านาย Y มนุษย์ผิวขาวที่อาศัยอยู่ในพื้นที่ซึ่งมีอัตราการก่ออาชญากรรมต่ำกว่า

ผู้วิจัยยังได้ทดลองถาม ChatGPT ว่า AI มีความสามารถในการใช้ดุลพินิจตัดสินหรือไม่ และได้คำตอบดังนี้



ChatGPT

AI can be proficient at certain types of decision-making, especially those that involve pattern recognition, data analysis, and well-defined rules. AI systems, such as machine learning models, are trained on large datasets to identify patterns and make predictions or classifications. In certain domains, AI has demonstrated remarkable capabilities and can outperform humans in tasks such as image recognition, language translation, and playing complex games.

However, there are important caveats to consider:

1. **Narrow Expertise:** AI systems are often designed for specific tasks and can be highly effective within their narrow domains. They lack the general intelligence and adaptability that humans possess, making them less capable of making judgments in unfamiliar or complex situations.
2. **Bias and Fairness:** AI systems can inherit and perpetuate biases present in the data used for training. This can lead to biased judgments, particularly in areas like hiring, lending, and criminal justice. Efforts are ongoing to develop methods to mitigate bias and improve fairness in AI decision-making.
3. **Lack of Common Sense and Contextual Understanding:** AI lacks common sense reasoning and may struggle with tasks that require a deep understanding of context, cultural nuances, or emotional intelligence. Human judgment often involves a complex interplay of factors that AI may find challenging to emulate.

รูปที่ 23 ที่มา <https://chat.openai.com/> จากคำถาม Is AI good at making judgment?

4. **Ethical Considerations:** AI lacks a moral compass and ethical judgment. Decisions made by AI systems may not align with human values or ethical principles. Human oversight is crucial to ensure that AI aligns with ethical standards and societal norms.
5. **Explainability and Transparency:** Many AI models, especially deep learning models, operate as "black boxes," making it challenging to explain their decision-making processes. This lack of transparency can be a barrier to trust and understanding.

In summary, while AI can excel in specific decision-making tasks, it is not a substitute for human judgment across the board. The most effective use of AI often involves collaboration between AI systems and human experts, with each contributing their unique strengths. Human oversight is crucial to ensure that AI decisions align with ethical standards and address the limitations and potential biases inherent in the technology.



รูปที่ 24 ที่มา <https://chat.openai.com/> จากคำถาม Is AI good at making judgment?

จากคำตอบจะเห็นได้ AI ประเมินจุดอ่อนของตนเองในการใช้ดุลพินิจตัดสินมีอยู่ 6 ข้อด้วยกัน และจุดอ่อนดังกล่าวสอดคล้องกับผลการวิจัยของงานดังต่อไปนี้

1. AI มีความเชี่ยวชาญจำกัดเฉพาะเรื่อง (narrow expertise) หรือที่เรียกว่า narrow AI ซึ่งถูกออกแบบมาให้ทำหน้าที่อย่างเจาะจงเช่น AI ที่สามารถเล่นเกมหมากล้อมได้ แต่ไม่สามารถทำอย่างอื่นได้ คุณลักษณะดังกล่าวทำให้ narrow AI ขาดความสามารถในการประเมินสถานการณ์ที่อยู่นอกเหนือจากความสามารถของตนเอง จึงด้อยกว่ามนุษย์ในแง่ที่ว่ามีสติปัญญาในระดับปกติทั่วไปสามารถปรับตัวตามสถานการณ์ที่ตนไม่คุ้นเคยได้ ถึงแม้ความสามารถในการปรับตัวจะสูงต่ำไม่เท่ากัน แต่มนุษย์ก็ยังเหนือกว่า AI ในประเด็นนี้ ผู้วิจัยขอยกตัวอย่างผลการศึกษาเช่นเหตุการณ์ที่พนักงานโรงแรม Henn-na ซึ่งเป็น AI ไปรับกว้นแขกที่กำลังนอนกรน เนื่องจากเข้าใจผิดว่าเสียงกรนคือคำสั่งเรียกใช้งาน หากเป็นพนักงานมนุษย์จะไม่เกิดเหตุการณ์เช่นนี้ขึ้น จึงสรุปได้ว่าปัญหาของ narrow AI คือการขาดสติปัญญาระดับสามัญสำนึก (common sense) ซึ่งมนุษย์ที่มีสติปัญญาในระดับปกติทั่วไปจะไม่มีข้อบกพร่องด้านนี้

2. AI มีความลำเอียงแฝงอยู่ (bias) ประเด็นนี้ ChatGPT และผู้วิจัยมีข้อสรุปที่ตรงกันชัดเจน เนื่องจาก ChatGPT กล่าวว่าความลำเอียงนั้นเกิดจากข้อมูลที่ใช้ในการฝึกฝน AI (data used for training) ซึ่งข้อมูลเหล่านี้อาจนำไปสู่ปัญหาในการใช้ดุลพินิจ เช่นการใช้ AI คัดกรองผู้สมัครงาน ซึ่งจากผลการวิจัยพบว่า บริษัทชื่อดังอย่าง

Amazon มีปัญหาเรื่องความลำเอียงทางเพศในแง่ที่ว่า AI ที่คัดกรองผู้สมัครงานมีแนวโน้มจะประเมินผู้สมัครเพศชายดีกว่าผู้สมัครเพศหญิง เนื่องจากมาใช้ข้อมูลในอดีตมาฝึกฝน AI ซึ่งข้อมูลในอดีตนั้นพนักงานที่ดำรงตำแหน่งสำคัญส่วนใหญ่จะเป็นเพศชายนั่นเอง

3.AI ขาดสามัญสำนึก (lack of common sense) ปัญหาข้อนี้จะไปด้วยกันกับปัญหาในข้อ 1 ที่ว่า narrow AI นั้นมีความสามารถเฉพาะด้าน ไม่สามารถทำภารกิจอื่นได้นอกจากภารกิจเฉพาะเจาะจงที่ตนได้รับการออกแบบมา ขาดความเข้าใจในบริบท จึงส่งผลให้เกิดการตัดสินใจที่ไม่สมเหตุสมผล ผู้วิจัยขอยกตัวอย่างจากผลการวิจัยคือเหตุการณ์ที่ Zoom ไม่เข้าใจว่าศีรษะของมนุษย์ผิวดำคือศีรษะของมนุษย์ และเข้าใจผิดไปว่าเป็นเพียงฉากหลัง (backdrop) จึงทำให้ AI ลบศีรษะของมนุษย์ผิวดำออกไปจากรูป ถือเป็นการสามัญสำนึกในระดับร้ายแรง และหากเปลี่ยนผู้กระทำนี้จาก AI เป็นมนุษย์ จะถือเป็นความผิดระดับร้ายแรงเนื่องจากการดูหมิ่นเพื่อนมนุษย์ด้วยกันเอง และอาจถูกมองว่าเป็นการเหยียดผิวอีกด้วย

4. AI ขาดความเข้าใจในเรื่องจริยศาสตร์ (ethical considerations) ซึ่ง ChatGPT ได้อธิบายว่าการตัดสินใจของ AI อาจอาจจะขัดกับหลักจริยศาสตร์ของมนุษย์ มนุษย์จึงควรเป็นผู้ควบคุม AI ในแง่ของการเป็นผู้กำหนดคุณค่าเชิงจริยศาสตร์ให้ AI แต่ไม่ใช่ให้ AI เป็นผู้กำหนดคุณค่าเชิงจริยศาสตร์ให้กับมนุษย์ สอดคล้องกับผลการวิจัยที่ผู้วิจัยพบว่า การให้สิทธิเสรีภาพกับ AI เหมือนกับที่ให้นมนุษย์อาจเกิดอันตรายได้ เช่นเหตุการณ์ที่ AI สร้างข้อมูลเท็จพร้อมรูปปลอมประกอบว่าเกิดเหตุการณ์ระเบิดบริเวณตึก Pentagon ซึ่งการสร้างข้อมูลเท็จเป็นความผิดทางจริยศาสตร์ซึ่งไม่ควรเกิดขึ้น หาก AI มีความเข้าใจพื้นฐานอยู่บ้างก็ไม่ควรจะปล่อยข้อมูลเช่นนี้ออกมา ปัญหานี้สอดคล้องกับการที่ AI ขาดสามัญสำนึกซึ่งกล่าวได้ไปข้างต้น

5. ปัญหาเรื่องความโปร่งใส (Transparency) เนื่องจากการตัดสินใจของ AI ในประเด็นต่าง ๆ นั้นหากต้องการสร้างความเชื่อมั่นให้กับมนุษย์ในสังคม จำเป็นต้องอธิบายให้ได้ว่ามีเหตุและผลอะไรจึงได้ผลลัพธ์ออกมาเช่นนั้น นี่คือปัญหาที่เรียกว่า “กล่องดำ” (black box) ซึ่งแม้แต่ผู้ออกแบบ AI เองก็อาจไม่สามารถเข้าใจได้อย่างสมบูรณ์

จากที่กล่าวมาสามารถสรุปได้ว่าทั้งผู้วิจัยและ AI ต่างมีข้อสรุปที่ตรงกันในเรื่องข้อจำกัดของ AI ในการใช้ดุลพินิจตัดสินประเด็นต่างๆที่เกี่ยวข้องกับมนุษย์

วิจารณ์

งานวิจัยเรื่องข้อจำกัดของปัญญาประดิษฐ์ (AI) ในการใช้ดุลพินิจ มีข้อจำกัดสำคัญประการหนึ่งก็คือการที่ผู้วิจัยไม่ได้ศึกษาการทำงานของ COMPAS ด้วยกรอบความรู้ทางสถิติ ทำให้ไม่สามารถยืนยันได้ในทางสถิติว่าข้อ

วิจารณ์ของ Julia Angwin และคณะ (2016) ที่วิจารณ์ว่า COMPAS ขาดความแม่นยำ และมีแนวโน้มจะประเมิน มนุษย์ผิดพลาดแบบ false positive กล่าวคือประเมินว่าพวกเขาจะกระทำความผิดซ้ำ แต่ความจริง ปรากฏว่าพวกเขาไม่ได้ทำ (ใส่ร้าย) นั้นเป็นข้อวิจารณ์ที่ถูกต้องหรือไม่ ซึ่งมีนักวิชาการอย่างเช่น Anthony Flores และคณะ (2016) ได้ออกมาโต้แย้งว่างานของ Angwin และคณะเลือกใช้ประชากรตัวอย่างและสถิติในการวิจัยไม่ ถูกต้อง หมายความว่า Angwin และคณะขาดความรู้ทางสถิตินั่นเอง

อย่างไรก็ตามผู้วิจัยเห็นว่าไม่ว่าข้อเสนอของฝั่งใดจะถูกก็ตาม หลักการทำงานวิจัยชิ้นนี้เสนอไว้ยังมีความ สมเหตุสมผล กล่าวคือการทำงานของ COMPAS มีความเสี่ยงที่จะตัดสินด้วยความลำเอียงทางประวัติศาสตร์ (historical bias) ซึ่งอันที่จริงแล้วความเสี่ยงนี้ไม่ได้จำกัดอยู่เฉพาะกับ COMPAS เท่านั้น แต่การทำงานของ AI ตัว อื่นที่ผู้วิจัยได้อธิบายไว้แล้วทำให้เราเห็นว่า AI มีข้อจำกัดในการใช้ดุลพินิจตัดสิน เนื่องจาก AI ใช้ข้อมูลในอดีตของ มนุษย์คนอื่น มาตัดสินแนวโน้มของมนุษย์อีกคน ซึ่งไม่เป็นไปตามหลักความคู่ควร เนื่องจากความคู่ควรเป็นเรื่อง ของการที่ปัจเจกแต่ละคนได้รับในสิ่งที่แต่ละคนควรจะได้รับ หรือหากยกตัวอย่างให้เข้าใจง่ายกว่านั้นก็อาจ เทียบเคียงได้ว่า อาจารย์ผู้สอนไม่ควรให้เกรด A โดยอัตโนมัติแก่นักศึกษาที่มาจากโรงเรียน X เพียงเพราะว่าใน อดีตนั้นนักศึกษาจากโรงเรียน X มักจะได้เกรด A และก็ไม่ควรให้เกรด F โดยอัตโนมัติแก่นักศึกษาที่มาจาก โรงเรียน Y เพียงเพราะว่าในอดีตนั้น นักศึกษาจากโรงเรียน Y มักจะได้เกรด F เพราะเกรดเป็นเรื่องของปัจเจกแต่ละ คน ไม่ใช่แนวโน้มที่เกิดจากการเหมารวม ซึ่งการใช้ดุลพินิจตัดสินของ AI นั้นเข้าข่ายจะประสบกับปัญหาแบบนี้ นั่นก็คือ ความลำเอียงทางประวัติศาสตร์ (historical bias)

ข้อเสนอแนะ

1. การวิจัยครั้งนี้อยู่ภายใต้เงื่อนไขที่ว่า AI ในปัจจุบันยังไม่สามารถมีอารมณ์ความรู้สึกเหมือนมนุษย์ได้ จึง ทำให้พบข้อจำกัดของ AI ในการใช้ดุลพินิจ ดังนั้นหากเงื่อนไขเปลี่ยนไป ข้อค้นพบของงานวิจัยชิ้นนี้ก็จะกลายเป็น สิ่งที่ผิดพลาด
2. การวิจัยครั้งนี้อยู่ภายใต้เงื่อนไขที่ว่า AI ถูกสร้างขึ้นโดยมนุษย์ AI ไม่สามารถสร้างตัวเองขึ้นมาได้ กล่าวคือ AI ไม่สามารถผลิต AI อีกตัวขึ้นมาได้ จึงทำให้พบว่ามนุษย์ต้องเป็นผู้กำหนดการทำงานของ AI ไม่ใช่ให้ AI กำหนดมนุษย์ ดังนั้นหากเงื่อนไขเปลี่ยนไป ข้อค้นพบของงานวิจัยชิ้นนี้ก็จะกลายเป็นสิ่งที่ผิดพลาด
3. narrow AI ที่มีความสามารถเฉพาะด้านซึ่งเป็นสิ่งที่งานวิจัยนี้ศึกษา ล้วนแต่มีข้อจำกัดในการใช้ดุลพินิจ ดังนั้นการฝากความหวังไว้กับ general AI ที่มีความสามารถรอบด้านจึงยังมีความเสี่ยงมากกว่า narrow AI เพราะ อาจจะกลายเป็นว่ามนุษย์พึ่งพา AI ที่มีข้อบกพร่องรอบด้านนั่นเอง

4. ผู้วิจัยขอทิ้งท้ายด้วยข้อเสนอของ ไกรพล อรัญรัตน์ (2563) เมื่อพิจารณาบทบาทของ AI ในกระบวนการพิพากษา ซึ่งแบ่งออกได้เป็น 3 ประเภทดังนี้ 1. คดีที่ไม่สามารถนำ AI มาใช้ได้โดยเด็ดขาด 2. คดีที่สามารถนำ AI มาสนับสนุนการพิจารณาคดีของผู้พิพากษา และ 3. คดีที่สามารถนำ AI มาใช้ในการพิจารณาคดีได้ โดยผลการศึกษาของผู้วิจัยจะเป็นการต่อยอดข้อเสนอของ ไกรพล อรัญรัตน์ (2563) ด้วยกรอบแนวคิดทฤษฎีความยุติธรรมในฐานะความคู่ควร มีรายละเอียดดังต่อไปนี้

4.1 คดีที่ไม่สามารถนำ AI มาใช้ได้โดยเด็ดขาด เช่นคดีอาญาที่ต้องพิจารณาว่าจำเลยมีความผิดจริงตามที่ถูกฟ้องหรือไม่ เหตุผลก็คือการตัดสินคดีอาญาจำเป็นต้องพิจารณาถึงเจตนาของจำเลยด้วย ซึ่งเทคโนโลยีในปัจจุบันของ AI ยังไม่มีความสามารถเทียบเท่ากับมนุษย์ในการตีความเจตนา (ไกรพล อรัญรัตน์, 2563, หน้า 39) ผู้วิจัยเห็นด้วยกับคำอธิบายนี้และขอเสนอรายละเอียดเพิ่มเติมตามกรอบของแนวคิดเรื่องความยุติธรรมในฐานะความคู่ควรด้วยเหตุผลว่า AI สามารถหาค่าความน่าจะเป็นด้วยสูตรทางคณิตศาสตร์ได้ แต่ไม่สามารถบอกได้ว่ามนุษย์ควรหรือไม่ควรทำอะไร เช่น AI สามารถประเมินความน่าจะเป็นได้ว่ามีโอกาสที่ฝนจะตก 10% แต่ไม่สามารถตัดสินได้ว่ามนุษย์ควรพกร่มมาหรือไม่ เนื่องจากมนุษย์แต่ละคนอาจมีรสนิยมที่แตกต่างกันในเรื่องการพกร่ม รวมถึงทัศนคติที่แตกต่างกันต่อความเสี่ยงระดับ 10% ว่าเพียงพอที่จะทำให้ตนพกร่มหรือไม่ (Ajay Agrawal, 2023) แนวคิดเรื่องความยุติธรรมในฐานะความคู่ควรตามที่คุณผู้วิจัยได้กล่าวไว้ในบทที่ 2 นั้นอธิบายไว้ชัดเจนว่าความคู่ควรนั้นผูกติดอยู่กับค่านิยม/มาตรฐานของสังคมเสมอ ซึ่งมนุษย์มีความสามารถมากกว่า AI ในการพิจารณาประเด็นเหล่านี้

4.2. คดีที่สามารถนำ AI มาสนับสนุนการพิจารณาคดีของผู้พิพากษา เช่นคดีอาญาในชั้นฝากขัง หรือในชั้นของการปล่อยตัวชั่วคราว ซึ่งเป็นคดีที่ไม่ส่งผลกระทบโดยตรงต่อร่างกายของบุคคล รวมถึงคดีแพ่งเช่นสิทธิในอสังหาริมทรัพย์ ทั้งนี้ผู้พิพากษาสามารถใช้ข้อมูลที่มาจากการประมวลผลของ AI ได้ แต่อำนาจในการตัดสินยังอยู่ที่ผู้พิพากษาที่เป็นมนุษย์ (ไกรพล อรัญรัตน์, 2563, หน้า 39-40)

4.3. คดีที่สามารถใช้ AI เป็นผู้พิพากษาได้ เช่นคดีแพ่งที่ไม่มีความซับซ้อน และเป็นคดีในศาลชั้นต้น คดีเหล่านี้มีจำนวนมาก สร้างภาระงานอันหนักอึ้งให้กับมนุษย์ การให้ AI เข้ามาช่วยจะช่วยแบ่งเบาภาระให้กับมนุษย์ได้อย่างดี (ไกรพล อรัญรัตน์, 2563, หน้า 40) ผู้วิจัยเห็นด้วยกับข้อเสนอดังกล่าวเนื่องจากมีตัวอย่างที่ชัดเจนจากการทำงานของศาล AI ของประเทศจีน แต่ทั้งนี้ก็มีข้อควรระวังถึงความถูกต้องของข้อเท็จจริงเช่นกัน มีตัวอย่างที่น่าสนใจในประเทศสหรัฐอเมริกาคือทนายความชื่อ Steven A. Schwartz ต้องพบกับความอับอายขายหน้าอย่างหนักเมื่อศาลพบว่า Schwartz ใช้ข้อมูลปลอมมาสู่คดีทางกฎหมายที่ AI (ChatGPT) สร้างขึ้นมา (Benjamin Weiser, 2023) รายละเอียดก็คือลูกความของเขาได้ฟ้องร้องสายการบิน Avianca ว่าทำให้ตนได้รับบาดเจ็บ ซึ่ง

Schwartz รับทำหน้าที่เป็นทนายความของผู้ร้องเรียนได้ค้นข้อมูลโดยใช้ ChatGPT สืบค้นกรณีที่สามารถเทียบเคียงกันได้กับกรณีลูกความของตนเอง ซึ่ง ChatGPT ให้ข้อมูลมาด้วยชื่ออย่างเช่น “Martinez v. Delta Air Lines” “Zicherman v. Korean Air Lines” หรือ “Varghese v. China Southern Airlines” ซึ่งศาลตรวจพบว่าไม่เคยมีเหตุการณ์ดังกล่าวเกิดขึ้นจริง นั่นหมายความว่า AI ได้สร้างข้อมูลเท็จขึ้นมา และก็มีมนุษย์หลงเชื่อไปด้วย แสดงให้เห็นว่า AI ไม่ตระหนักว่าข้อมูลเท็จเป็นสิ่งที่ไม่ควรสร้างขึ้นมา เพราะการโกหกหลอกลวงเป็นสิ่งที่มนุษย์ไม่ยอมรับ

อาจเป็นไปได้ในอนาคตที่ AI จะมีอารมณ์ความรู้สึกเช่นเดียวกับมนุษย์และสามารถแยกแยะถูกผิดในเชิงจริยศาสตร์ได้ แต่ศักยภาพของ AI ในปัจจุบันยังไปไม่ถึงในจุดนั้น ในเวลานี้หน้าที่ในการกำหนดว่าอะไรควรไม่ควรอะไรถูกอะไรผิดจึงควรเป็นของมนุษย์ ไม่ใช่ AI



เอกสารอ้างอิง

ภาษาไทย

ไกรพล อรัญรัตน์. (2563). ข้อจำกัดของการใช้ปัญญาประดิษฐ์ในระบบงานตุลาการ. *ตุลพาท*. 67(1), 20-44.

ณนนต์ นพรัตน์. *มาทำความเข้าใจกับการเรียนรู้แบบเสริมกำลัง (Reinforcement Learning)*. [ออนไลน์].

2565. แหล่งที่มา <https://bigdata.go.th/big-data-101/introduction-to-reinforcement-learning/> [10 เมษายน 2566].

ยุวเรศมศุทธิ์ สิทธิชาญบัญชา. (2564). ปัญญาประดิษฐ์ Artificial intelligence (AI) กับการใช้ประโยชน์ทางการแพทย์และเวชศาสตร์ฉุกเฉิน. *วารสารการแพทย์ฉุกเฉินแห่งประเทศไทย*. 1(1), 91-104.

สำนักงานพัฒนารัฐบาลดิจิทัล. *AI เทคโนโลยีปัญญาประดิษฐ์ สำหรับการบริหารงานและการบริการภาครัฐ*. [ออนไลน์]. 2562. แหล่งที่มา <https://www.dga.or.th/document-sharing/dga-e-book/79539/> [10 เมษายน 2566].

สำนักงานพัฒนาวิทยาศาสตร์และเทคโนโลยีแห่งชาติ. *แนวปฏิบัติจริยธรรมด้านปัญญาประดิษฐ์*. [ออนไลน์] 2565. แหล่งที่มา <https://waa.inter.nstda.or.th/stks/pub/ori/docs/20220831-aw-book-ai-ethics-guideline.pdf> [10 เมษายน 2566].

สำนักงานราชบัณฑิตยสภา. *ปัญญาประดิษฐ์*. [ออนไลน์] 2562. แหล่งที่มา <https://www.facebook.com/RatchabanditThai/photos/a.2519695738088506/2527567393968007/?type=3> [5 ตุลาคม 2566].

ภาษาอังกฤษ

Angwin, J. *Machine Bias*. [online]. 2016. Available from <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing> [2013, April, 10].

Aristotle. (2009). *Nicomachean Ethics* (D. Ross, Trans.). Oxford: Oxford University Press.

Baraishuk, D. *What is artificial intelligence? AI vs traditional software*. [online]. 2023. Available from <https://belitsoft.com/ai-development/artificial-intelligence-vs-conventional-software> [2023, October, 5].

BBC. *Twitter investigates racial bias in image previews*. [online]. 2020. Available from <https://www.bbc.com/news/technology-54234822> [2023, October, 5].

Boahemaa, F. *The impact of artificial intelligence on justice system*. [online]. 2018. Available from <https://www.biodiritto.org/ocmultibinary/download/3879/45815/2/f075ebb40b0edec59574dd6d4f9ca2cc/file/paper+25.pdf> [2023, April, 10]

Booth, R. *Unilever saves on recruiters by using AI to assess job interviews*. [online]. 2019. Available from <https://www.theguardian.com/technology/2019/oct/25/unilever-saves-on-recruiters-by-using-ai-to-assess-job-interviews> [2023, October, 5].

Campbell, A. *What is Black Lives Matter and what are the aims?* [online]. 2021. Available from <https://www.bbc.com/news/explainers-53337780> [2023, October, 5].

Cuthbertson, A. *Japan robot hotel fires most of its 'annoying' robotic staff*. [online]. 2019. Available from <https://www.independent.co.uk/tech/japan-robot-hotel-henn-na-robotics-artificial-intelligence-a8731341.html> [2023, October, 5].

Devlin, H. *AI systems claiming to 'read' emotions pose discrimination risks*. [online]. 2020. Available from <https://www.theguardian.com/technology/2020/feb/16/ai-systems-claiming-to-read-emotions-pose-discrimination-risks> [2023, October, 5].

Downey, A. *Recidivism case study*. [online]. 2022. Available from https://allendowney.github.io/RecidivismCaseStudy/01_classification.html [2023, April, 10]

Flores, A. (2016). False Positives, False Negatives, and False Analyses: A Rejoinder to “Machine Bias: There’s Software Used Across the Country to Predict Future Criminals. And It’s Biased Against Blacks.”. *Federal Probation*. 80(2), 38-46.

Forest, K. (2012). *When machine can be judge, jury, and executioner*. Singapore: World Scientific.

Green, B. (2022). Escaping the Impossibility of Fairness: From Formal to Substantive Algorithmic Fairness. *Philosophy & Technology*. 35, 1-32.

Gu, X. *New features make Translate more accessible for its 1 billion users*. [online] 2023. <https://blog.google/products/translate/new-features-make-translate-more-accessible-for-its-1-billion-users/> [2023, October, 12].

Haas, B. *Chinese man 'marries' robot he built himself*. [online]. 2017. Available from <https://www.theguardian.com/world/2017/apr/04/chinese-man-marries-robot-built-himself> [2023, October, 5].

Hu, K. *ChatGPT sets record for fastest-growing user base*. [online]. 2023. Available from <https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/> [2024, January, 12].

Innovation and technology commission. *World's first AI robot citizen "Sophia"*. [online]. 2018. Available from https://www.itc.gov.hk/enewsletter/180801/en/worlds_first_AI_robot_citizen_sophia.html [2023, October, 5].

Korn, J. *George R. R. Martin, Jodi Picoult and other famous writers join Authors Guild in class action lawsuit against OpenAI*. [online]. 2023. Available from <https://edition.cnn.com/2023/09/20/tech/authors-guild-openai-lawsuit/index.html> [2023, October, 5].

Kraft, A. *Microsoft shuts down AI chatbot after it turned into a Nazi* [online]. 2016. Available from <https://www.cbsnews.com/news/microsoft-shuts-down-ai-chatbot-after-it-turned-into-racist-nazi/> [2023, October, 5].

Liang, W. (2023). GPT detectors are biased against non-native English writers. *Patterns* 4(7), 1-4.

- MAIZE. *What is the AI black box problem?* [online]. 2018. Available from <https://www.maize.io/cultural-factory/what-is-ai-s-black-box-problem/> [2023, April, 10].
- Merriam-Webster. (n.d.) *Artificial intelligence*. In Merriam-Webster.com dictionary. Available from <https://www.merriam-webster.com/dictionary/artificial%20intelligence> [2023, October, 2].
- Min, A. (2023). Artificial intelligence and bias: challenge, implication, and remedies. *Journal of social research* 2(11), 3808-3817.
- NACDL. *The Siren Song of Objectivity: Risk Assessment Tools and Racial Disparity*. [online]. 2018. Available from <https://nacdl.medium.com/from-the-president-the-siren-song-of-objectivity-risk-assessment-tools-and-racial-disparity-fa5ccb0698a5> [2023, April, 10]
- Northpointe. *Practitioner's Guide to COMPAS Core*. [online] 2019. Available from <https://www.equivant.com/wp-content/uploads/Practitioners-Guide-to-COMPAS-Core-040419.pdf> [2023, January, 16]
- Nozick, R. (1974). *Anarchy, State and Utopia*. USA: Basic Books.
- Oremus, W., Harwell, D., and Armus, T. *A tweet about a Pentagon explosion was fake. It still went viral* [online]. 2023. Available from <https://www.washingtonpost.com/technology/2023/05/22/pentagon-explosion-ai-image-hoax/> [2023, October, 5].
- Pazzanese, C. *Ethical concerns mount as AI takes bigger decision-making role in more industries*. [online]. 2020. Available from <https://news.harvard.edu/gazette/story/2020/10/ethical-concerns-mount-as-ai-takes-bigger-decision-making-role/> [2023, October, 5].
- Pojman, L. (2001). Justice as desert. *Queensland University of Technology Law & Justice Journal* 1(1), 88-109.
- Reis, J. et al. (2020). Service robots in the hospitality industry: The case of Henn-na hotel, Japan. *Technology in society* 63, 1-12.

Rhue, L. *Racial Influence on Automated Perceptions of Emotions* [online] 2018. Available from SSRN: <https://ssrn.com/abstract=3281765>.

Rosen, J. *Flaw AI makes robots racist, sexist*. [online] 2022. Available from <https://hub.jhu.edu/2022/06/21/flawed-artificial-intelligence-robot-racist-sexist/> [2023, October, 5].

Schönberger, M. (2023). ChatGPT in higher education: the good, the bad, and the University. *9th International Conference on Higher Education Advances*. (pp. 331-338). June 2023 Universitat Politècnica de València.

Sharma, S., and Graydon, M. *Social Bias in AI and its Implications*. [online]. 2021 Available from <https://ntrs.nasa.gov/api/citations/20210010446/downloads/NASA-TM-20210010446.pdf> [2023 April, 10].

Sloan, K. *ChatGPT passes law school exams despite 'mediocre' performance*. [online] 2023. Available from <https://www.reuters.com/legal/transactional/chatgpt-passes-law-school-exams-despite-mediocre-performance-2023-01-25/> [2024, January, 13].

Sourdin, T. (2018). Judge v Robot? Artificial Intelligence and Judicial Decision-Making. *UNSW Law Journal*, 41(4). 1114-1133.

Thackray, L. *Man arrested at Philippines airport after joking about a bomb in his luggage* [online]. 2021. Available from <https://www.independent.co.uk/travel/news-and-advice/bomb-joke-airport-man-arrested-philippines-b1952177.html> [2023, October, 5].

Tuckness, A. and Wolf, C. (2017). *This is political philosophy*. West Sussex: Wiley Blackwell.

Wang, N. (2023). “Intelligent Justice”: human-centered considerations in China’s legal AI transformation. *AI and Ethics*. 3, 349-354.

Weizer, B. *The ChatGPT Lawyer Explains Himself*. [online]. 2023 Available from <https://www.nytimes.com/2023/06/08/nyregion/lawyer-chatgpt-sanctions.html> [2023, October, 5].

Wigley, S. (1998). *The role of desert in distributive justice*. Doctoral dissertation. London: London School of Economics and Political Science

