วารสารสังคมศาสตร์ มหาวิทยาลัยนเรศวร ปีที่ 19 ฉบับที่ 2 (กรกฎาคม – ธันวาคม 2566): 31–57

Received: 7 Dec. 2022; Revised: 26 Jun. 2023; Accepted: 29 Jun. 2023

ปัญญาประดิษฐ์กับความรับผิดชอบทางจริยธรรม: ประเด็นปัญหาทางปรัชญา

ปิยณัฐ ประถมวงษ์[1]

บทคัดย่อ

ความก้าวหน้าของปัญญาประดิษฐ์ได้ก่อให้เกิดข้อถกเถียงเรื่องความรับผิดชอบทางจริยธรรม มนุษย์มีแนวโน้มที่จะใช้ให้ปัญญาประดิษฐ์ทำงานแทนตนเองมากขึ้น ในแง่หนึ่งปัญญาประดิษฐ์สามารถเรียนรู้และตัดสินใจได้เอง ทว่ามันก็ไม่สามารถรับผิดชอบทางจริยธรรมอันเป็นอิสระจากมนุษย์ได้ คำถามที่น่าสนใจก็คือว่าหากปฏิบัติการของปัญญาประดิษฐ์นำไปสู่ผลร้ายทางจริยธรรมที่คาดไม่ถึง เราจะทำความเข้าใจเรื่องความรับผิดชอบทางจริยธรรมของปัญญาประดิษฐ์ได้อย่างไร บทความวิจัยนี้ใช้วิธีการวิจัยเชิงเอกสาร มีวัตถุประสงค์มุ่งศึกษาและวิเคราะห์ประเด็นปัญหาเรื่องปัญญาประดิษฐ์กับความรับผิดชอบทางจริยธรรม และอภิปรายเพื่อนำเสนอมุมมองทางปรัชญาของผู้เขียนเพื่อพยายามแก้ปัญหาเรื่องช่องว่างของความรับผิดชอบทางจริยธรรม ผลการวิจัยมีข้อเสนอหลัก 2 ข้อ กล่าวคือ (1) ผู้เขียนสนับสนุนแนวคิดที่ว่าความรับผิดชอบทางจริยธรรมจำเป็นต้องยึดโยงกับการเป็นผู้กระทำการทางจริยธรรมในแบบที่ไม่ใช้กรอบคิดแบบดั้งเดิม กล่าวคือ ไม่ได้กำหนดความรับผิดชอบไปที่ปัจเจกบุคคลและคุณลักษณะแบบมนุษย์ของปัจเจกบุคคล และ (2) สืบเนื่องจากเหตุผลแรก ผู้เขียนจึงเสนอว่ามโนทัศน์เรื่องการเป็นผู้กระทำการทางจริยธรรมในฐานะที่เป็นปฏิสัมพันธ์ร่วมกันระหว่างมนุษย์กับปัญญาประดิษฐ์เป็นสิ่งจำเป็นต่อความรับผิดชอบทางจริยธรรม แม้ว่าแนวคิดนี้จะยังคงมีแนวโน้มยึดโยงอยู่กับการเป็นผู้กระทำการ และความเป็นสาเหตุซึ่งสอดคล้องกับวิธีคิดแบบมองย้อนกลับหลัง แต่เราก็ควรพยายามปรับให้แนวคิดนี้ให้มีความสอดคล้องกับวิธิคิดเรื่องความรับผิดชอบทางจริยธรรมในแบบที่มองไปข้างหน้าด้วยเช่นกัน

คำสำคัญ: ปัญญาประดิษฐ์ ความรับผิดชอบทางจริยธรรม การเป็นผู้กระทำการทางจริยธรรม ช่องว่างของความรับผิดชอบทางจริยธรรม

---

[1] อาจารย์ประจำภาควิชามนุษยศาสตร์ คณะสังคมศาสตร์และมนุษยศาสตร์ มหาวิทยาลัยมหิดล <Email: piyanat.pra@mahidol.ac.th>

Journal of Social Sciences Naresuan University, Volume 19 Number 2 (July–December, 2023): 31–57

Received: 7 Dec. 2022; Revised: 26 Jun. 2023; Accepted: 29 Jun. 2023

Artificial Intelligence and Moral Responsibility: A Philosophical Problem

Piyanat Prathomwong[2]

ABSTRACT

The progress of artificial intelligence (AI) has raised debates on moral responsibility. Humans tend to delegate tasks and practices to AI. Whereas AI can learn and make decision by itself, it cannot be morally responsible and independent from the human being. The crucial question is how can we understand and articulate the moral responsibility of AI if the practices of AI have unintended consequences or are morally wrong? According to a documentary research method, this article aims to study and analyze the problem of AI and moral responsibility. It then discusses and proposes arguments attempting to tackle the problem of the responsibility gap. As a result, it offers two thesis statements: (1) it advocates that moral responsibility demands moral agency in a non–traditional approach, i.e., it does not locate responsibility in the individual and its human–like attribution. Furthermore, (2) it argues, following the previous statement, that the concept of moral agency as a mutual interaction between humans and AI is necessary for moral responsibility. Although the argument tends to deal with agency and causality—which is consistent with a backward–looking approach, we should also adjust it to consider moral responsibility in terms of a forward–looking approach.

Keywords: Artificial Intelligence, Moral Responsibility, Moral Agency, Responsibility Gap

[2] Lecturer, Department of Humanities, Faculty of Social Science and Humanities, Mahidol University <Email: piyanat.pra@mahidol.ac.th>

บทนำ

ในช่วงของการเข้าสู่ศตวรรษที่ 21 เทคโนโลยีได้พัฒนาอย่างก้าวกระโดด มีการพัฒนาปัญญาประดิษฐ์ (Artificial Intelligence: AI) ด้วยเทคนิคที่ซับซ้อนขึ้น เช่น การเรียนรู้ของเครื่อง (machine learning) และการเรียนรู้เชิงลึก (deep learning) ซึ่งคอมพิวเตอร์หรือจักรกลสามารถเรียนรู้และตัดสินใจได้เอง ในขณะเดียวกัน มนุษย์มีแนวโน้มที่จะยอมให้ AI เข้ามามีบทบาทในการปฏิบัติงานหรือทำกิจกรรมต่าง ๆ แทนที่มนุษย์มากขึ้น จนอาจกล่าวได้ว่าการกระทำหนึ่ง ๆ อาจเป็นการกระทำร่วมกันระหว่าง AI และมนุษย์ ในปัจจุบันและอนาคตอันใกล้เราจะเห็นปฏิสัมพันธ์ระหว่างมนุษย์และจักรกล (human–machine interaction) มากขึ้น ดังนั้น ทั้งมนุษย์และ AI จึงมีส่วนร่วมกันในระบบความสัมพันธ์ของการกระทำการบางอย่างเสมอ

การศึกษาวิเคราะห์ว่าในโลกที่มนุษย์ดำรงอยู่ร่วมกับ AI อย่างใกล้ชิดเช่นนี้ เราจะนับใครและสิ่งใดบ้างว่าควรเป็นผู้รับผิดชอบทางจริยธรรม หากเกิดกรณีที่เทคโนโลยีส่งผลร้ายหรือผลลัพธ์ที่ไม่คาดคิดต่อมนุษย์ การระบุตัวผู้รับผิดชอบทางจริยธรรมโดยเพียงการชี้ไปที่ปัจเจกบุคคลที่เป็น ‘ผู้ใช้’ ‘ผู้ออกแบบ’ หรือ ‘ผู้สร้าง’ เพียงอย่างเดียว อาจไม่เพียงพอในการทำความเข้าใจความซับซ้อนของปัญหาอีกต่อไป ในหลายกรณีผู้ใช้ที่ใช้ AI ก็ไม่อาจล่วงรู้ว่า AI จะตัดสินใจอย่างไร ยิ่งไปกว่านั้น ผู้ออกแบบหรือผู้สร้าง AI เช่น วิศวกรคอมพิวเตอร์ก็ไม่สามารถคาดการณ์ได้ว่าอัลกอริทึมและโมเดลของ AI ที่ตนเองสร้างขึ้นมาจะตัดสินใจอย่างไรหรือจะทำอะไรในอนาคต เนื่องจาก AI สามารถเรียนรู้และตัดสินใจได้เอง

บทความวิจัยนี้จึงมุ่งศึกษา วิเคราะห์ และอภิปรายข้อถกเถียงเรื่อง AI กับความรับผิดชอบทางจริยธรรม ทั้งนี้ ผู้เขียนมุ่งอภิปรายและนำเสนอมุมมองทางปรัชญาที่อาจช่วยแก้ปัญหาข้อถกเถียงเรื่องปัญญาประดิษฐ์และความรับผิดชอบทางจริยธรรมได้

วัตถุประสงค์

1. เพื่อศึกษาวิเคราะห์ประเด็นปัญหาทางปรัชญาเรื่องปัญญาประดิษฐ์กับความรับผิดชอบทางจริยธรรม

2. เพื่ออภิปรายและนำเสนอมุมมองทางปรัชญา โดยมุ่งนำเสนอความเป็นไปได้ในการแก้ปัญหาเรื่องความรับผิดชอบทางจริยธรรมของปัญญาประดิษฐ์

วิธีดำเนินการวิจัย

โครงการวิจัยนี้เป็นการวิจัยเชิงเอกสาร (documentary research) ตามลักษณะของการวิจัยในสาขาปรัชญา ซึ่งเป็นสาขาย่อยในสาขามนุษยศาสตร์ กล่าวคือเป็นการมุ่งเน้นการสืบค้นเอกสาร ทำความเข้าใจ แสดงข้อถกเถียง ตีความ วิเคราะห์มโนทัศน์เชิงปรัชญา อภิปรายข้อถกเถียงต่าง ๆ พร้อมกับแสดงทัศนะทางวิชาการและนำเสนอมุมมองทางปรัชญาของผู้เขียน

ขอบเขตการวิจัย

บทความวิจัยนี้เป็นการวิจัยเชิงเอกสาร ผู้เขียนได้กำหนดขอบเขตการวิจัยโดยมุ่งค้นคว้าเอกสาร หนังสือ และงานวิจัยที่เกี่ยวกับ AI กับความรับผิดชอบทางจริยธรรม โดยเน้นงานวิชาการและงานวิจัยของนักปรัชญาเทคโนโลยี นักจริยศาสตร์เทคโนโลยี และนักจริยศาสตร์ปัญญาประดิษฐ์ดังที่แสดงในรายการเอกสารอ้างอิง ประการสำคัญ ผู้เขียนพยายามมุ่งค้นหาและสำรวจเอกสาร หนังสือ และงานวิจัยที่ตีพิมพ์ในปีใกล้ปัจจุบันมากที่สุด เพื่อให้มั่นใจว่าขอบเขตของการวิจัยมีความเท่าทันต่อข้อถกเถียงในแวดวงวิชาการในปัจจุบัน

ผลการวิจัย

บทความวิจัยนี้มุ่งศึกษาและวิเคราะห์เชิงมโนทัศน์เรื่อง AI กับความรับผิดชอบทางจริยธรรม โดยพิจารณาว่ามโนทัศน์ดังกล่าวสามารถถูกเข้าใจอย่างไรได้บ้าง นักปรัชญามีข้อถกเถียงอย่างไรบ้าง จากนั้นบทความได้อภิปรายเพื่อนำเสนอมุมมองทางปรัชญาของผู้เขียนที่พยายามเสนอความเป็นไปได้ในการแก้ปัญหาเรื่องช่องว่างของความรับผิดชอบทางจริยธรรม โดยที่ผู้เขียนมีข้อเสนอหลัก (thesis statement) 2 ข้อ ซึ่งเป็นผลจากการวิจัยดังนี้

1. ความรับผิดชอบทางจริยธรรมของ AI จำเป็นต้องยึดโยงกับการเป็นผู้กระทำการทางจริยธรรมในแบบที่ไม่ใช้กรอบคิดแบบดั้งเดิม

2. สืบเนื่องจากข้อแรก นำมาสู่การเสนอว่า มโนทัศน์เรื่องการเป็นผู้กระทำการทางจริยธรรมในฐานะที่เป็นปฏิสัมพันธ์ร่วมกันระหว่างมนุษย์กับปัญญาประดิษฐ์เป็นสิ่งจำเป็นต่อความรับผิดชอบทางจริยธรรมของ AI

อย่างไรก็ตาม ก่อนที่จะนำเสนอการอภิปรายในรายละเอียดของข้อเสนอหลักในส่วนต่อไป ผู้เขียนจำเป็นต้องพิจารณามโนทัศน์ในทางปรัชญาเรื่องความรับผิดชอบทางจริยธรรมตามกรอบคิดแบบดั้งเดิมในส่วนที่เกี่ยวข้องเฉพาะกับมนุษย์เสียก่อน

หลังจากนั้นจึงระบุปัญหาว่ากรอบคิดแบบดั้งเดิมดังกล่าวมีปัญหาอย่างไรเมื่อนำมาปรับใช้กับกรณีของ AI ต่อจากนั้น ผู้เขียนจึงกล่าวถึงประเด็นปัญหาเรื่อง AI กับช่องว่างของความรับผิดชอบทางจริยธรรม (responsibility gap) ซึ่งเป็นปัญหาหลักที่ทำให้เกิดข้อถกเถียงและข้อเสนอต่าง ๆ ตามมา และในท้ายที่สุดผู้เขียนจะอภิปรายเพื่อนำเสนอข้อเสนอหลัก 2 ข้อ ซึ่งเป็นผลจากการวิจัยนี้

ความรับผิดชอบทางจริยธรรมของมนุษย์ตามกรอบคิดแบบดั้งเดิม

ตามกรอบคิดปรัชญากรีกรวมถึงปรัชญาสมัยใหม่ ความเป็นมนุษย์ผูกโยงอยู่กับแนวคิดแบบปัจเจกชนนิยม (individualism) และการมีเหตุผล (rationality) มนุษย์คือสิ่งมีชีวิตที่สามารถเลือกกระทำการได้จากความมุ่งหมายของตนเอง ความมุ่งหมายคือสารัตถะหนึ่งของการอธิบายการกระทำว่าการกระทำต่าง ๆ นั้นเกิดขึ้นจากความจงใจของผู้กระทำ พูดอีกอย่างหนึ่งได้ว่ามนุษย์มีเสรีภาพที่จะคิดและกระทำได้จากเจตนาของตน และเมื่อมนุษย์ดำรงชีวิตอยู่ในสังคมอันเป็นพื้นที่ที่ประกอบไปด้วยตนเองและผู้อื่น การกระทำของมนุษย์คนหนึ่งหรือกลุ่มหนึ่งจึงมักจะส่งผลต่อผู้อื่นเสมอทั้งในทางที่ดีและทางที่ร้าย ลักษณะเช่นนี้กล่าวได้ว่ามนุษย์คือผู้กระทำการเชิงจริยธรรม และด้วยเหตุนี้ผู้กระทำการเชิงจริยธรรมจำเป็นต้องมีความรับผิดชอบทางจริยธรรมในสิ่งที่ตนกระทำ

เมื่อพิจารณาตามมุมมองแบบอาริสโตเติลที่นำเสนอในงานปรัชญาชิ้นสำคัญเรื่อง *Nicomachean Ethics* เราจะพบข้อเสนอว่า เงื่อนไขของความรับผิดชอบทางจริยธรรมมีอยู่ 2 ประการ คือ เงื่อนไขเชิงสาเหตุ และเงื่อนไขเชิงความรู้ โดยสามารถสรุปได้ดังนี้ (ก) เงื่อนไขเชิงสาเหตุคือการที่บุคคลสามารถควบคุม (control) การกระทำของตนเองได้ โดยที่การกระทำของตนนั้นเป็นสาเหตุของผลที่ตามมา การที่บุคคลสามารถควบคุมการกระทำของตนเองได้จึงหมายถึงการกระทำที่เกิดขึ้นผ่านการควบคุมความเป็นสาเหตุของผลที่เกิดขึ้นจากการกระทำนั้น เช่น ความสามารถในการควบคุมร่างกายของตนเองได้ว่าจะกำหมัดแล้วชกไปที่ใบหน้าของคนที่ยืนอยู่ตรงหน้าหรือไม่ ในสภาวะปกติมนุษย์ที่ปกติมักควบคุมร่างกายตนเองได้เสมอว่าจะกระทำหรือไม่กระทำอะไร และ (ข) เงื่อนไขเชิงความรู้ คือ การที่บุคคลมีความรู้เกี่ยวกับผลของการกระทำ พูดอีกอย่างหนึ่งได้ว่า บุคคลมีความรู้ว่าการกระทำนั้นจะนำไปสู่ผลอะไร รวมถึงบุคคลนั้นมีความรู้หรือความตระหนักรู้ว่าตนเองกำลังกระทำสิ่งใดอยู่ เช่น การกระทำหยิบอาวุธไปยิงบุคคลอื่นเป็นเรื่องที่รู้ได้ว่าคนที่ถูกยิงนั้นอาจเสียชีวิตได้ และรู้ว่าตนเองกำลังกระทำหรือไม่ได้กระทำสิ่งนั้นอยู่ (Coeckelbergh, 2020a, p. 111)

นอกจากการพิจารณาการกระทำจากเงื่อนไข 2 ข้อ ตามมุมมองแบบดั้งเดิมของอาริสโตเติลแล้ว ยังมีอีกหนึ่งเงื่อนไขที่ทำให้ปัจเจกบุคคลจะต้องรับผิดชอบในสิ่งที่ตนกระทำคือเจตนาหรือความมุ่งหมายซึ่งเป็นเนื้อหาจิตที่เราสร้างขึ้นภายในใจก่อนที่จะส่งผลไปสู่การกระทำต่าง ๆ ของเรา ในแง่นี้ การกระทำที่จะเรียกได้ว่าเป็นการกระทำจากเจตจำนงของปัจเจกบุคคลจริง ๆ จะเกิดขึ้นได้ก็ต่อเมื่อปัจเจกบุคคลสร้างเนื้อหาในจิตที่มีความมุ่งหมายว่าปรารถนาจะกระทำอะไรบางอย่าง แล้วจิตในสมองจึงสั่งให้ร่างกายกระทำตามนั้น และบุคคลนี้ก็รู้ตัวว่าตนเองกำลังทำอะไร

วิธีคิดแบบสืบย้อนสาเหตุไปยังเนื้อหาของจิตและความจงใจ รวมถึงวิธีการพิจารณาเงื่อนไข 2 ประการของอาริสโตเติลมักพบเห็นได้ในข้อเสนอทางปรัชญาแบบดั้งเดิม หากลองตั้งข้อสังเกตในประเด็นนี้จะพบว่าวิธีคิดตามมุมมองแบบดั้งเดิมในเรื่องความรับผิดชอบทางจริยธรรมจะมุ่งเน้นแนวคิดแบบภายใน (internalist approach) กล่าวคือดูเจตนาในจิตของผู้กระทำการ ซึ่งจุดเด่นของวิธีคิดแบบนี้คือเป็นข้อเสนอที่เคร่งครัดและมีเหตุผล เนื่องจากผู้รับผิดชอบจะสามารถรับผิดชอบได้เฉพาะก็ต่อเมื่อเขามีเจตนาจากภายในจิตที่จะทำสิ่งต่าง ๆ รวมถึงมีความสามารถในการควบคุมการกระทำและความรู้ต่อการกระทำนั้น ๆ

ปัญหาของความรับผิดชอบทางจริยธรรมตามกรอบคิดแบบดั้งเดิมเมื่อนำมาปรับใช้กับ AI

กรอบคิดแบบดั้งเดิมดังที่กล่าวมามีปัญหาและมีข้อจำกัดเมื่อนำมาใช้ทำความเข้าใจเรื่องความรับผิดชอบทางจริยธรรมของ AI เมื่อเราพิจารณาเรื่องความรับผิดชอบทางจริยธรรมตามกรอบคิดแบบดั้งเดิมจะพบว่าเป็นการรับผิดชอบในสิ่งที่เกิดขึ้นแล้ว (retrospective responsibility) กล่าวอีกอย่างหนึ่งคือ เป็นการรับผิดชอบและการรับผิด (accountability and liability) ทั้งนี้ Porter et al. (2018) และ Zerilli et al. (2021) ระบุตรงกันว่า นักปรัชญาส่วนใหญ่ที่ยึดแนวคิดทางปรัชญาแบบดั้งเดิมนี้เมื่อพิจารณาประเด็นความรับผิดชอบทางจริยธรรมก็มักมีแนวโน้มใช้มิติในข้อนี้ แนวโน้มเช่นนี้สะท้อนถึงการให้ความสำคัญกับการสืบย้อนกลับสู่สาเหตุ (causation) ซึ่งเป็นการสืบย้อนสาเหตุสู่เนื้อหาของจิตและเจตนาภายในจิต รวมถึงความสามารถในการควบคุมการกระทำและความรู้ต่อการกระทำนั้น ๆ

อย่างไรก็ตาม แนวคิดทางปรัชญาแบบดั้งเดิมที่มีรากฐานมาจากอาริสโตเติลเป็นแนวคิดที่มีข้อจำกัดบางประการเมื่อเรากำลังพิจารณาเรื่องความรับผิดชอบทางจริยธรรมของ AI ด้วยสาเหตุดังที่ Ceockelbergh (2020b, p. 2053) ระบุว่าแนวคิด

แบบดั้งเดิมเน้นมิติของตัวมนุษย์ที่เป็นผู้กระทำการมากเกินไป และมีนัยว่าตั้งอยู่บนวิธีคิดที่มองเทคโนโลยีว่าเป็นเพียงเครื่องมือที่มีความเป็นกลาง ซึ่งปัจเจกบุคคลที่เป็นมนุษย์เท่านั้นที่จะเป็นผู้รับผิดชอบได้ ราวกับว่า AI ไม่ได้มีส่วนในการตัดสินใจและกระทำสิ่งต่าง ๆ เลย

แนวคิดแบบดั้งเดิมเช่นนี้แม้จะฟังดูสมเหตุสมผลในแง่ของการสืบย้อนกลับสู่สาเหตุของการกระทำที่นำไปสู่ผล แต่ก็อาจเป็นอุปสรรคในการทำความเข้าใจเรื่องความรับผิดชอบทางจริยธรรมของ AI และถือเป็นข้อท้าทายต่อการสร้างแนวคิดใหม่ ๆ ที่สอดคล้องกับความก้าวหน้าของ AI ที่สามารถเรียนรู้และตัดสินใจได้เอง จนก่อให้เกิดปัญหาเรื่องช่องว่างของความรับผิดชอบทางจริยธรรม ดังที่ผู้เขียนจะกล่าวถึงในส่วนถัดไป

ปัญหาเรื่องช่องว่างของความรับผิดชอบทางจริยธรรมของปัญญาประดิษฐ์

ก่อนที่ผู้เขียนจะอภิปรายและนำเสนอข้อเสนอหลักของตนเอง ในส่วนนี้ผู้เขียนจำเป็นต้องแสดงให้เห็นประเด็นปัญหาหลักในข้อถกเถียงเรื่องดังกล่าวเสียก่อน ข้อท้าทายในอภิจริยศาสตร์ที่เกี่ยวข้องกับ AI คือปัญหาเรื่องการถ่ายโอนการเป็นผู้กระทำการจากมนุษย์ไปสู่ AI ซึ่งแน่นอนว่าส่งผลต่อคำถามเรื่องความรับผิดชอบทางจริยธรรม ปัญหานี้จะไม่เกิดขึ้นในกรณีของการใช้เทคโนโลยีประเภทอื่นหรือเทคโนโลยีอัลกอริทึมแบบเก่าที่สามารถสืบกลับไปสู่สาเหตุได้ในเชิงเทคนิค กล่าวคือสำหรับอัลกอริทึมแบบเก่า เช่นวิธีการที่เรียกว่าต้นไม้ของการตัดสินใจ (decision tree) เรายังสามารถสืบย้อนกลับไปสู่กระบวนการและสาเหตุของการตัดสินใจนั้นได้ เมื่อเกิดผลเลวร้ายขึ้นเราก็สามารถสืบย้อนกลับไปสู่ต้นเหตุและแก้ปัญหาได้ (รวมถึงอาจจะหาผู้รับผิดชอบได้ เช่น ผู้ที่เขียนโปรแกรมนั้นขึ้นมา หรือหาจุดที่ผิดพลาดเชิงเทคนิคได้) กรณีเช่นนี้จะต่างจากกรณีในปัจจุบันที่เรามักใช้ AI ด้วยเทคนิคการเรียนรู้ของเครื่องโดยเฉพาะอย่างยิ่งวิธีการแบบการเรียนรู้เชิงลึกที่มีความซับซ้อนและลำดับชั้น (layers) มากมาย ดังที่ Gunkel (2020, pp. 311–312) ระบุว่าการสืบย้อนกลับดังกล่าวกลายเป็นสิ่งที่ทำยาก แม้ผู้เขียนโปรแกรมจะมีความรู้โดยทั่วไปว่าระบบทำงานอย่างไร แต่ผู้เขียนโปรแกรมเองก็ไม่สามารถเข้าใจ รวมถึงไม่สามารถอธิบายและคาดการณ์ได้ว่าเหตุใดระบบการเรียนรู้ของเครื่องจึงทำการตัดสินใจในแบบที่มันตัดสินใจ เราจึงไม่สามารถล่วงรู้หรือคาดการณ์ได้ว่า AI จะตัดสินใจอย่างไร รวมถึงเราไม่สามารถควบคุมและอธิบายสิ่งที่จะเกิดขึ้นจากการกระทำของ AI ได้ นอกจากนี้ Coeckelbergh (2020a, pp. 116–117) เห็นว่าประเด็นนี้ถือเป็น

ปัญหาด้านความโปร่งใส (transparency) ของ AI เนื่องจาก AI มีลักษณะที่เรียกว่ากล่องดำ (black box) มันจึงเป็นสิ่งที่มีความคลุมเครือ (opaqueness) ต่อการรับรู้และความเข้าใจของมนุษย์ เราไม่สามารถรู้ได้อย่างแน่ชัดว่า AI ประมวลผลอย่างไรมันจึงให้คำตอบหรือตัดสินใจหรือกระทำในแบบที่มันทำ

ดังนั้น หากเกิดกรณีที่ AI สร้างความเสียหายทางจริยธรรม เช่น มีคนบาดเจ็บเสียชีวิต หรือหากไม่ใช่เรื่องรุนแรงทางกายภาพแต่ก็เป็นเรื่องสำคัญ เช่น ความอยุติธรรม การเหยียดเพศ การเหยียดเชื้อชาติ ฯลฯ คำถามคือเราจะทำความเข้าใจเรื่องความรับผิดชอบทางจริยธรรมอย่างไร เราจะให้ AI รับผิดชอบอย่างสมบูรณ์ (อย่างเป็นอิสระจากมนุษย์) ได้หรือไม่ (สามัญสำนึกของเราโดยส่วนใหญ่มักตอบคำถามนี้ว่า ไม่) หรือไม่เช่นนั้น ใครควรจะรับผิดชอบ ผู้ใช้ ผู้เขียนโปรแกรม ผู้บริหารบริษัท วิศวกร หรือควรจะเป็นใคร และควรเป็นปัจเจกบุคคลหรือควรจะเป็นกลุ่มบุคคล และถ้าเป็นกลุ่ม จะต้องเป็นกลุ่มใด กลุ่มนั้นต้องมีเฉพาะมนุษย์ใช่หรือไม่ เป็นกลุ่มที่มีมนุษย์อยู่ร่วมกับจักรกลที่เป็น AI ได้หรือไม่

สภาวะที่ AI สามารถตัดสินใจได้เอง จนเกิดปัญหาเรื่องช่องว่างของความรับผิดชอบสะท้อนสภาวะที่มนุษย์ไม่สามารถควบคุมเทคโนโลยีได้ ดังที่ Matthias (2004) เสนอว่าเทคโนโลยีประเภทที่ตัดสินใจเองได้โดยปราศจากการแทรกแซงโดยมนุษย์ อย่างเช่นวิธีการเรียนรู้ของเครื่องมักจะก่อให้เกิดปัญหาช่องว่างของความรับผิดชอบ เนื่องจากไม่มีใครที่จะสามารถควบคุมการกระทำของการเรียนรู้ของเครื่องได้ จึงส่งผลไปสู่ปัญหาที่ว่าเราไม่รู้ว่าจะให้ใครหรือผู้ใดรับผิดชอบการกระทำดังกล่าว

ประเด็นนี้เป็นปัญหาสำคัญที่เป็นแกนกลางและจุดตั้งต้นของบทความวิจัยนี้ ซึ่งในทางวิชาการเรียกว่า ‘ปัญหาเรื่องช่องว่างของความรับผิดชอบทางจริยธรรม’ กล่าวคือ ยิ่ง AI ฉลาดและทำงานได้แม่นยำขึ้นมากเท่าไหร่ มนุษย์ยิ่งมอบหมายให้ AI ทำงานแทนมากขึ้นเท่านั้น ในขณะเดียวกันมนุษย์ก็จะยิ่งมีอำนาจในการควบคุมและมีความรู้ต่อ AI น้อยลงไปเรื่อย ๆ (Matthias, 2004) หากพิจารณาโดยสามัญสำนึกแล้วมนุษย์ไม่สามารถไปตำหนิหรือให้ AI เป็นผู้รับผิดชอบทางจริยธรรมได้เนื่องจากว่า AI เป็นวัตถุที่ไม่มีชีวิต ไม่มีความรู้สำนึกและไม่มีเสรีภาพเหมือนมนุษย์ หรือพูดอีกอย่างหนึ่งคือ AI ไม่รู้ผิดชอบชั่วดี แต่ครั้นจะให้มนุษย์รับผิดชอบเองก็ไม่สามารถทำได้อีก เนื่องจากมนุษย์ขาดทั้งอำนาจควบคุมและขาดทั้งความรู้ต่อ AI ที่ฉลาด

มากขึ้นทุกวัน เราจึงไม่รู้ว่าเราควรจะระบุหรือกำหนดให้ 'ใคร' หรือ 'สิ่งใด' มารับผิดชอบทางจริยธรรมต่อผลที่เกิดขึ้น ในทางทฤษฎีจึงเกิดปัญหาช่องว่างที่มีแนวโน้มจะขยายใหญ่ขึ้นเรื่อย ๆ

ปัญหาสำคัญอีกประการหนึ่งที่เกิดจากช่องว่างของความรับผิดชอบทางจริยธรรมคือมันเปิดทางให้ปัจเจกบุคคลอ้างประเด็นนี้เพื่อหลบเลี่ยงความรับผิดชอบทางจริยธรรมที่เกิดขึ้นจากการใช้เทคโนโลยี ปัจเจกบุคคลอาจจะปฏิเสธความรับผิดชอบด้วยการอ้างว่าการตัดสินใจและการกระทำทั้งหมดล้วนเกิดจากระบบ AI มนุษย์จึงไม่สามารถคาดการณ์หรือแม้กระทั่งรับผิดชอบได้ เนื่องจากเมื่อใดก็ตามที่มนุษย์มอบหมายงานให้ AI ทำแทน เมื่อนั้นมนุษย์ก็ไม่สามารถควบคุมสิ่งต่าง ๆ ได้ มนุษย์หลุดจากการควบคุม (out of control) ไปเสียแล้ว และมนุษย์ก็ขาดความรู้ว่าข้างในกล่องดำของระบบ AI เกิดอะไรขึ้นกันแน่ เงื่อนไขที่กล่าวมานี้ทำให้เกิดปัญหาและความซับซ้อนในการทบทวนย้อนกลับเชิงสาเหตุของการกระทำต่าง ๆ มนุษย์จึงรู้สึกว่าตนไม่สามารถรับผิดชอบทางจริยธรรมได้

นอกจากลักษณะปัญหาเรื่องช่องความความรับผิดชอบทางจริธรรมดังที่กล่าวมาแล้ว เรายังพบลักษณะความซับซ้อนของเทคโนโลยีสมัยใหม่ที่ก่อให้เกิดปัญหาด้านการกำหนดความรับผิดชอบทางจริยธรรมในอีกแง่มุมหนึ่ง ซึ่งถือว่ามีส่วนทำให้ปัญหาเรื่องช่องว่างของความรับผิดชอบยิ่งขยายใหญ่ขึ้นไปอีก กล่าวคือเราไม่สามารถระบุไปที่ตัวปัจเจกบุคคลในฐานะเป็นผู้รับผิดชอบแต่เพียงผู้เดียวได้อีกต่อไป การมีผู้เกี่ยวข้องเป็นจำนวนมากกับเทคโนโลยีที่ซับซ้อนเช่นนี้ก่อให้เกิดปัญหาที่เรียกว่า 'ปัญหาเรื่องการมีหลายบุคคลเกี่ยวข้อง' (problem of many hands) (Jonas, 1984) จากปัญหาที่เกี่ยวเนื่องกับช่องว่างของความรับผิดชอบทางจริยธรรมดังที่กล่าวมาแล้ว เมื่อมนุษย์ไม่สามารถกล่าวโทษอุปกรณ์เทคโนโลยีได้ รวมถึงไม่สามารถกล่าวโทษปัจเจกบุคคลได้ มนุษย์ก็อาจอ้างปัญหาการมีหลายบุคคลเกี่ยวข้อง โดยอ้างว่าความรับผิดชอบนั้นควรจะเป็นของคนอื่นด้วย แทนที่จะเป็นความรับผิดชอบของเขาที่ปัจเจกบุคคลเพียงคนเดียว เนื่องจากเหตุการณ์เช่นนี้มีมนุษย์หลายคนเกี่ยวข้อง ผู้ปฏิบัติการที่เกี่ยวข้องอาจจะถามได้ว่าเหตุใดจึงต้องโทษเขาเพียงคนเดียวหรือเหตุใดความรับผิดชอบจึงตกอยู่ที่เขาเพียงคนเดียว ยิ่งไปกว่านั้นในกรณีของ AI ที่มีความเกี่ยวข้องกับข้อมูลและโลกวัตถุที่ซับซ้อนที่ไม่ใช่มนุษย์ เช่น ชุดข้อมูล ระบบอัลกอริทึมที่ตัดสินใจได้เอง เครือข่ายอินเตอร์เน็ต ระบบคลาวด์ ลักษณะเชิงวัตถุที่เข้ามาเกี่ยวข้องกับกระบวนการตัดสินใจ รวมถึงการกระทำที่เกิดขึ้นจากปฏิสัมพันธ์ระหว่างมนุษย์และจักรกล ฯลฯ เราจึงอาจเรียกลักษณะปัญหา

เชิงวัตถุนี้ได้ว่า ‘ปัญหาการมีหลายสิ่งเกี่ยวข้อง’ (problem of many things) (Coeckelbergh, 2020a, pp. 113–114) ทั้งนี้ Tigard (2020) ตั้งข้อสังเกตว่าปัญหาเรื่องช่องว่างของความรับผิดชอบทางจริยธรรมของ AI เกิดขึ้นเนื่องจากมุมมองแบบปัจเจกชนนิยมและมโนทัศน์เรื่องความรับผิดชอบทางจริยธรรมแบบดั้งเดิม ความรับผิดชอบทางจริยธรรมจึงผูกโยงอยู่กับความเป็นสาเหตุ (causality) กล่าวคือ ผู้กระทำการได้กระทำสิ่งต่าง ๆ อันเกิดมาจากสาเหตุและความมุ่งหมายของเนื้อหาจิต หรือแม้จะไม่มีความมุ่งหมายของจิต แต่เราก็อาจพิจารณาได้ว่าผู้กระทำการนั้นเป็นต้นเหตุของผลที่ตามมา แม้ว่าแบบหลังนี้จะเป็นนิยามที่อ่อนกว่าแบบแรกก็ตาม แต่ก็ถือว่าเป็นวิธีคิดแบบเชิงสาเหตุทั้งสิ้น นอกจากนี้ Matthias (2004) ยังชี้ว่า ปัญหาเรื่องช่องว่างของความรับผิดชอบทางจริยธรรมของ AI ไม่สามารถแก้ปัญหาได้ด้วยมโนทัศน์แบบดั้งเดิมที่มีฐานคิดวางอยู่ในหลักการเชิงสาเหตุ

จากที่ได้กล่าวมา จะเห็นว่าประเด็นปัญหาทางปรัชญาเรื่อง AI กับความรับผิดชอบทางจริยธรรมโดยเฉพาะปัญหาเรื่องช่องว่างของความรับผิดชอบทางจริยธรรมเป็นเรื่องที่ต้องได้รับการแก้ไข ในส่วนต่อไปของบทความ ผู้เขียนจะมุ่งอภิปรายข้อถกเถียงต่าง ๆ เพื่อนำเสนอมุมมองทางปรัชญาที่พยายามเสนอความเป็นไปได้ในการแก้ปัญหาดังกล่าว

สรุปและอภิปรายผล

ในการพิจารณาประเด็นเรื่องช่องว่างของความรับผิดชอบทางจริยธรรมของ AI นั้น วิธีคิดตั้งต้นที่แตกต่างกันจะทำให้เรามองเห็นปัญหาเรื่องนี้ไปต่างกัน เช่น เราอาจมีวิธีคิดแบบการมองโลกในแง่ดีของเทคโนโลยี (techno-optimism) ที่เห็นว่าปัญหาเรื่องช่องว่างของความรับผิดชอบทางจริยธรรมสามารถแก้ไขได้ (Tigard, 2020) เป้าหมายของบทความวิจัยนี้คือการมองหาความเป็นไปได้ในการแก้ปัญหาเรื่องช่องว่างของความรับผิดชอบทางจริยธรรม ผู้เขียนจึงขอยึดวิธีคิดแบบมองโลกในแง่ดีเป็นจุดตั้งต้น โดยที่ในส่วนนี้ ผู้เขียนจะอภิปรายเพื่อนำเสนอมุมมองทางปรัชญาที่พยายามเสนอความเป็นไปได้ในการแก้ปัญหาเรื่องช่องว่างของความรับผิดชอบทางจริยธรรม โดยที่ผู้เขียนมีข้อเสนอหลัก 2 ข้อ ดังนี้

1. ความรับผิดชอบทางจริยธรรมของ AI จำเป็นต้องยึดโยงกับการเป็นผู้กระทำการทางจริยธรรมในแบบที่ไม่ใช้กรอบคิดแบบดั้งเดิม

2. การนำเสนอมโนทัศน์เรื่องการเป็นผู้กระทำการทางจริยธรรมในฐานะที่เป็นปฏิสัมพันธ์ร่วมกันระหว่างมนุษย์กับปัญญาประดิษฐ์เป็นสิ่งจำเป็นต่อความรับผิดชอบ

ทางจริยธรรมของ AI

ข้อเสนอหลักข้อที่ 1: ความรับผิดชอบทางจริยธรรมของ AI จำเป็นต้องยึดโยงกับการเป็นผู้กระทำการทางจริยธรรมในแบบที่ไม่ใช้กรอบคิดแบบดั้งเดิม

การที่เราจะเห็นด้วยกับแนวคิดที่ว่าความรับผิดชอบทางจริยธรรมยึดโยงอยู่กับการเป็นผู้กระทำการนั้น เราต้องมีข้อสันนิษฐานเริ่มต้นว่าความรับผิดชอบทางจริยธรรมต้องการการเป็นผู้กระทำการมาเป็นสิ่งรองรับ หรือกล่าวอีกอย่างหนึ่งคือ การเป็นผู้กระทำการคือเงื่อนไขหรือสารัตถะสำคัญของความรับผิดชอบทางจริยธรรม นั่นหมายความว่า ถ้าเราสามารถระบุการเป็นผู้กระทำการทางจริยธรรมได้ เราก็สามารถกำหนดความรับผิดชอบทางจริยธรรมให้ผูกพันกับผู้กระทำการนั้นได้

ในแวดวงการวิจัย AI มีข้อเสนอว่าการวิจัยและถกเถียงเรื่องจริยศาสตร์ปัญญาประดิษฐ์ นอกจากการมุ่งเน้นที่ตัวมนุษย์แล้ว เราควรมีทิศทางที่ให้ความสำคัญกับการพิจารณาทางจริยธรรมของสิ่งที่ไม่ใช่มนุษย์ด้วยเช่นกัน เช่น สัตว์ สิ่งมีชีวิตทางชีววิทยา สิ่งแวดล้อม และ AI (Owe & Baum, 2021) ในทำนองเดียวกันนี้ เราก็พบนักวิจัยที่พยายามเสนอการสร้าง AI เพื่อกำหนดให้ AI สามารถมีความรับผิดชอบทางจริยธรรมได้ ข้อเสนอเช่นนี้สามารถแบ่งได้เป็น 2 แบบ คือ การสร้างผู้กระทำการทางจริยธรรมโดยใช้กรอบคิดแบบดั้งเดิม และการสร้างผู้กระทำการทางจริยธรรมโดยไม่ใช้กรอบคิดแบบดั้งเดิม แบบแรกเป็นสิ่งที่ผู้เขียนปฏิเสธ ส่วนแบบที่สองเป็นสิ่งที่ผู้เขียนเห็นด้วย

ผู้เขียนขอเริ่มจากการกล่าวถึงแนวคิดในการสร้าง AI ให้เป็นผู้กระทำการทางจริยธรรมโดยใช้กรอบคิดแบบดั้งเดิม ซึ่งเป็นแนวคิดที่ผู้เขียนไม่เห็นด้วย การสร้างผู้กระทำการทางจริยธรรมโดยใช้กรอบคิดแบบดั้งเดิม คือการพยายามสร้างผู้กระทำการทางจริยธรรม (moral agent) ให้เหมือนกับมนุษย์ การใช้คำว่า ‘กรอบคิดแบบดั้งเดิม’ ในบริบทนี้ผู้เขียนหมายถึงมุมมองที่เห็นว่า มนุษย์มีคุณลักษณะของการเป็นผู้กระทำการทางจริยธรรมที่มีความรับผิดชอบทางจริยธรรมเช่นใด AI ก็ต้องมีคุณลักษณะในแบบเดียวกันทุกประการ ผู้เขียนตั้งข้อสังเกตว่าความเข้าใจต่อมโนทัศน์เรื่อง AI ส่งผลต่อความคิดเรื่องความรับผิดชอบทางจริยธรรม กล่าวคือ หากเรามีความเชื่อว่าเราจะสามารถสร้าง AI หรือหุ่นยนต์ได้เหมือนมนุษย์จริง ๆ เราก็มีแนวโน้มที่จะคิดแบบเทียบเคียงกันตรง ๆ ระหว่าง AI กับมนุษย์ หรือมีวิธีคิดแบบมานุษยรูปนิยม (anthropomorphism) กล่าวคือ หากเราเชื่อว่าเราสามารถสร้าง

AI ให้มีเนื้อหาจิต มีจิตสำนึก และมีเจตจำนงเสรีได้จริง เราก็สามารถคิดถึง AI ได้ในระนาบเดียวกับมนุษย์ คือมองว่า AI เป็นเหมือนมนุษย์แบบหนึ่ง รวมถึงอาจพิจารณา AI แบบนี้ว่ามีลักษณะเป็นผู้กระทำการทางจริยธรรมเหมือนกับที่มนุษย์เป็น AI จึงสามารถตัดสินใจและกระทำการในสิ่งต่าง ๆ อย่างเป็นอิสระจากมนุษย์ได้ และ AI จึงต้องรับผิดชอบทางจริยธรรมในสิ่งที่มันกระทำได้ด้วยตัวเองอย่างเป็นอิสระจากมนุษย์ได้เช่นกัน

แนวคิดแบบนี้มีจุดเด่นในเรื่องความสมเหตุสมผลในการพิจารณาเรื่องความรับผิดชอบทางจริยธรรม โดยมีความสอดคล้องกับกรอบคิดแบบดั้งเดิมของอาริสโตเติลอันเป็นจุดแข็งทางปรัชญามาโดยตลอด ดังนั้น ถ้าหากว่าข้อเท็จจริงในโลกอนาคตเป็นไปตามฉากทัศน์แบบนี้ได้จริง ๆ นั่นก็หมายความว่ามนุษย์อาจต้องยกสถานะให้ AI หรือหุ่นยนต์เป็นพลเมืองหรือมีสิทธิทางกฎหมายและต้องจ่ายภาษีเหมือนกับมนุษย์ เนื่องจากเราพิจารณาว่า AI หรือหุ่นยนต์ตัวหนึ่งก็เป็นเหมือนมนุษย์คนหนึ่ง ท้ายที่สุดแล้ว ข้อท้าทายสำหรับวิธีคิดนี้ขึ้นอยู่กับหลักฐานเชิงประจักษ์และความก้าวหน้าในการสร้าง AI และเกี่ยวข้องกับข้อถกเถียงเชิงทฤษฎีในเรื่องความเป็นไปได้ที่จะสร้าง AI ที่มีจิตสำนึกและเจตจำนงเสรี

อย่างไรก็ตาม แนวคิดนี้ก็ยังมีจุดอ่อน เนื่องจากต้องเผชิญคำถามและความสงสัยเรื่องความเป็นไปได้ที่จะสร้าง AI ที่มีคุณลักษณะต่าง ๆ เหมือนกับมนุษย์ทุกประการ ดังที่กล่าวมาแล้วว่าวิธีคิดเช่นนี้เป็นการยึดแนวคิดเรื่องความรับผิดชอบทางจริยธรรมตามกรอบแบบดั้งเดิม เพียงแค่นำคุณลักษณะของผู้กระทำการเชิงจริยธรรมของมนุษย์ไปสวมครอบไว้บน AI ที่ถูกสร้างขึ้นมาเพื่อให้มีคุณลักษณะเหมือนมนุษย์ทุกประการ (human–like attribution) ซึ่งแนวคิดนี้เป็นสิ่งที่ผู้เขียนปฏิเสธด้วยเหตุผล 2 ประการ

ประการแรก ผู้เขียนปฏิเสธความเป็นไปได้ในการสร้าง AI ที่มีคุณลักษณะเหมือนมนุษย์ทุกประการ เนื่องจากลักษณะเฉพาะทางชีววิทยาของมนุษย์ที่ผ่านการวิวัฒนาการตามธรรมชาติมาอย่างยาวนานน่าจะเป็นสิ่งที่เลียนแบบหรือสร้างขึ้นเองได้ยาก นอกจากนี้ การประมวลผลของ AI ไม่ได้กระทำในลักษณะเดียวกับสมองมนุษย์ทุกประการ หากลักษณะการประมวลผลของทั้งสองสิ่งมีความแตกต่างกันก็เป็นไปได้ยากที่จะสร้าง AI ที่มีคุณลักษณะเหมือนกับมนุษย์ทุกประการ นอกจากนี้ยังมีประเด็นข้อถกเถียงเรื่องความเป็นไปได้ที่จะสร้าง AI ที่มีจิตสำนึกเหมือนมนุษย์ กล่าวคือ ยังมีปัญหาช่องว่างเชิงญาณวิทยาที่แก้ไขยาก จิตสำนึกเป็นมุมมองเชิงอัตวิสัย (subjective) ในขณะที่การวิจัยทางวิทยาศาสตร์เป็นการมุ่งหาหลักฐานเชิงประจักษ์

เป็นมุมมองเชิงวัตถุวิสัย (objective) ดังนั้น เราจึงไม่สามารถใช้วิธีการเชิงวัตถุวิสัยไปตรวจหาหรือรับรองการมีอยู่ของจิตสำนึกซึ่งเป็นสิ่งอัตวิสัยได้ เช่นเดียวกับการที่เราไม่สามารถสร้าง AI ที่มีจิตสำนึกได้หากเรายังไม่สามารถเข้าใจจิตสำนึกอย่างเป็นวัตถุวิสัยหรืออย่างเป็นวิทยาศาสตร์ได้ เหตุผลประการแรกนี้นำมาสู่การปฏิเสธความเชื่อที่ว่า AI ในตัวมันเองจะสามารถรับผิดชอบทางจริยธรรมได้อย่างเป็นอิสระจากมนุษย์ เนื่องจาก AI ไม่มีความรู้สำนึกและไม่มีเจตจำนงเสรีเหมือนกับที่มนุษย์มี (Johnson, 2015; Sharkey, 2017; Hakli & Mäkelä, 2019; Bernáth, 2021) ดังนั้น การพยายามสร้าง AI ให้เป็นผู้กระทำการทางจริยธรรมในแบบที่เหมือนมนุษย์เพื่อให้มันรับผิดชอบได้อย่างเป็นอิสระจากมนุษย์จึงเป็นคำกล่าวอ้างที่ ‘ไปไกลเกินไป’ (going too far) (Gunkel, 2020, p. 313)

ประการที่สอง ผู้เขียนปฏิเสธแนวทางการตั้งเป้าหมายการสร้าง AI ให้มีคุณลักษณะเหมือนมนุษย์ ข้อนี้เป็นความคิดเห็นในเชิงปทัสถาน (normative) ของผู้เขียน กล่าวคือ ผู้เขียนคิดว่ามนุษย์ไม่ควรตั้งเป้าหมายที่จะสร้าง AI ให้เหมือนมนุษย์ เนื่องจากการทำเช่นนี้มีแนวโน้มนำไปสู่การลดทอนคุณค่าและศักดิ์ศรีความเป็นมนุษย์ หากจะสร้าง AI เราควรสร้าง AI ขึ้นมาเพื่อให้มันเป็น AI เท่านั้น กล่าวคือ เราไม่ควรสร้าง AI ขึ้นมาเพื่อให้มันเป็น (เหมือน) มนุษย์ การกระทำเช่นนี้มีแนวโน้มสร้างความเสี่ยงต่อการดำรงอยู่ของมนุษย์ (existential risk) นอกจากนี้ การพยายามสร้าง AI ให้เหมือนมนุษย์อาจจะยิ่งทำให้มนุษย์เข้าใจสภาพการณ์ของ AI ผิดเพี้ยนไปจากความเป็นจริงในเชิงวิทยาศาสตร์ ดังที่ Tigard, Conradie, & Nagel (2020) ชี้ว่า ยิ่งเราพยายามสร้าง AI ที่สามารถรับผิดชอบได้เองหรือสร้างวิธีคิดที่จะทำให้ AI มีสถานะทางจริยธรรมบางอย่างเหมือนมนุษย์มากเท่าไหร่มันก็จะยิ่งทำให้ปัญหาต่าง ๆ รุนแรงขึ้นไปอีก กล่าวคือ มันจะทำให้เราจินตนาการไปว่าจักรกล (ซึ่งไม่มีชีวิต) กลายเป็นอะไรบางอย่างที่มากกว่าหรือเกินไปจากสิ่งที่มันเป็นอยู่จริง ๆ เช่น เราอาจจินตนาการไปว่าจักรกลที่มีความรับผิดชอบทางจริยธรรมเป็นจักรกลที่มีอารมณ์ความรู้สึก จินตนาการเช่นนี้ทำให้ข้อสันนิษฐานเบื้องต้นของเราหลุดลอยจากความเป็นจริงเชิงประจักษ์และอาจทำให้เรามุ่งแก้ปัญหาต่าง ๆ ในทิศทางที่ผิดเนื่องจากเราเข้าใจสภาพการณ์ที่แท้จริงของความเป็น AI ผิดเพี้ยนไป

ในส่วนต่อไป ผู้เขียนจะอภิปรายเพื่อสนับสนุนข้อเสนอหลักข้อที่ 1 ในเรื่องการสร้าง AI ให้เป็นผู้กระทำการทางจริยธรรมโดยไม่ใช้กรอบคิดแบบดั้งเดิม หากเราปฏิเสธจินตนาการแบบมานุษยรูปนิยมที่คิดว่า AI มีคุณลักษณะทุกประการเหมือนมนุษย์ออกไป เราก็ควรจะเชื่อว่า AI หรือหุ่นยนต์เป็นได้แค่เพียงระบบ

การวิเคราะห์และคาดการณ์จากการประมวลผลข้อมูลเท่านั้น เรียกได้ว่าเป็นซอฟต์แวร์รูปแบบหนึ่งที่ทำงานบนฐานของโครงสร้างฮาร์ดแวร์หรือวัตถุรูปแบบหนึ่ง แม้จะมีศักยภาพมากเพียงใดก็ไม่สามารถกล่าวได้ว่าเป็นผู้กระทำการทางจริยธรรมได้อย่างสมบูรณ์เหมือนกับที่มนุษย์เป็น

มีงานวิจัยจำนวนหนึ่ง เช่น Floridi and Sanders (2004) Wallach and Allen (2009) และ Anderson and Anderson (2011) ที่พยายามปรับเปลี่ยนนิยามหรือมโนทัศน์ของคำว่าผู้กระทำการทางจริยธรรมใหม่ เพื่อหลีกหนีกรอบคิดแบบดั้งเดิม นั่นคือมีการเสนอให้สร้าง ‘ผู้กระทำการทางจริยธรรมแบบประดิษฐ์’ (Artificial Moral Agent: AMA) ซึ่งเป็นการพยายามสร้างกรอบคิด สร้างคำนิยามและคำอธิบายแบบใหม่ เพื่อนำไปอธิบาย AMA ที่แม้ไม่ใช่ผู้กระทำการทางจริยธรรมที่เหมือนมนุษย์ แต่ก็อาจถือได้ว่าเป็นผู้กระทำการทางจริยธรรมได้ในบางระดับโดยไม่จำเป็นต้องใช้กรอบคิดดั้งเดิมที่ผูกโยงอยู่กับการมีจิตสำนึก เจตจำนงเสรี และเนื้อหาของจิต แนวคิดนี้เป็นแนวคิดที่ค่อนไปทางด้านการต่อต้านมานุษยรูปนิยม (anti–anthropomorphism) รวมถึงต่อต้านแนวคิดมนุษย์เป็นศูนย์กลาง (anti–anthropocentrism)

แนวคิดนี้มีการนำเสนอให้ปรับมโนทัศน์ใหม่ (reconceptualization) ของการเป็นผู้กระทำการทางจริยธรรม โดยเพิ่มมโนทัศน์คำว่า ‘แบบประดิษฐ์’ (artificial) เข้าไป กลายเป็นคำใหม่ว่า ‘ผู้กระทำการทางจริยธรรมแบบประดิษฐ์’ ผู้เขียนมีข้อสังเกตว่าโดยปกติคำว่า ‘ผู้กระทำการเชิงจริยธรรม’ หมายถึงผู้มีคุณลักษณะในด้านการกระทำเชิงจริยธรรมซึ่งเป็นมนุษย์เท่านั้น เหมือนกับกรณีคำว่า ‘ปัญญา’ หรือ intelligence ที่โดยความหมายปกติจะหมายถึงปัญญาหรือความฉลาดของมนุษย์เท่านั้น และเมื่อมีการประดิษฐ์หรือสร้างปัญญาขึ้นมา จึงมีคำเรียกว่า ‘ปัญญาประดิษฐ์’ ความน่าสนใจอยู่ตรงที่ในการนิยามคำว่าปัญญาประดิษฐ์นั้น เราสามารถนิยามว่าเป็นความฉลาดหรือปัญญาที่ไม่ใช่แบบมนุษย์ก็ได้ แม้ว่าเทคนิคการเรียนรู้เชิงลึกจะได้แรงบันดาลใจมาจากการทำงานของสมองมนุษย์ก็ตาม แต่มันก็ไม่ได้ทำงานเหมือนกับสมองมนุษย์ การที่ AI สามารถแยกภาพหมาจิ้งจอกออกจากหมาพุดเดิลได้นั้น ระบบด้านความฉลาดหรือด้านปัญญาที่กระทำเช่นนั้นไม่ได้มีลักษณะการประมวลผลเหมือนกับตอนที่สมองมนุษย์ทำการแยกภาพหมาจิ้งจอกออกจากหมาพูเดิล ในแง่นี้ กล่าวได้ว่าคำว่า ‘ผู้กระทำการทางจริยธรรมแบบประดิษฐ์’ (หรือ AMA) จึงไม่จำเป็นต้องนิยามเหมือนกับ ‘ผู้กระทำการทางจริยธรรม’ (แบบมนุษย์) เสมอไป

ตัวอย่างการเสนอแนวคิด AMA เช่น Anderson and Anderson (2011) ที่เสนอว่าจักรกลอาจจะทำตามกฎทางจริยธรรมได้ดีกว่ามนุษย์ และการสร้าง AI ที่มีความสามารถในการใช้เหตุผลอย่างไม่มีอคตินั้นมีความเป็นไปได้ ข้อเสนอทำนองนี้ตั้งอยู่บนสมมติฐานที่ว่า AI ที่ถูกฝึกมาอย่างดีจะสามารถเข้าใจและทำตามกฎจริยธรรมได้ดีกว่ามนุษย์ เนื่องจากมนุษย์มักจะอ่อนแอและยอมให้อารมณ์อยู่เหนือเหตุผล (Anderson & Anderson, 2011; Coeckelbergh, 2020c, p. 154; Gunkel; 2020) ในขณะที่ Wallach and Allen (2009) เสนอแนวคิดเรื่องจักรกลที่มีจริยธรรม (moral machine) และยืนยันว่าการสร้าง AMA เป็นสิ่งจำเป็น โดยหลีกเลี่ยงมุมมองแบบดั้งเดิมด้วยการปฏิเสธการพยายามสร้าง AI ที่มีจิตสำนึกหรือเจตจำนงเสรี ทว่าเสนอให้ใช้มุมมองทางจริยธรรมเชิงหน้าที่ (functional morality approach) ด้วยการพิจารณาว่า AI สามารถตอบสนองหรือแสดงถึงพฤติกรรมที่มีความไวทางจริยธรรม (moral sensitivity) บางประการได้หรือไม่ ถ้าทำได้แสดงว่า AI มีลักษณะของ AMA นั่นคือ AI มีพฤติกรรมภายนอกที่แสดงออก ‘ราวกับว่า’ (as–if) มันเป็นผู้กระทำการทางจริยธรรมจริง ๆ (Coeckelbergh, 2020c, pp. 154–155) นอกจากนี้ เรายังพบข้อเสนอเรื่อง AMA โดย Floridi and Sanders (2004) ที่พยายามสร้างกรอบมโนทัศน์ขึ้นใหม่อย่างชัดเจน กล่าวคือ ให้ตัดการพิจารณาเรื่องจิตรู้สำนึก เจตจำนงเสรี และเนื้อหาจิตออกไปจากคำนิยามของการเป็น AMA และเสนอกรอบเงื่อนไข 3 ประการสำหรับนิยามใหม่นี้ คือ (ก) ความสามารถเปลี่ยนสภาวะตนเองด้วยการตอบสนองต่อสิ่งเร้า (ข) ความสามารถเปลี่ยนสภาวะตนแม้จะไม่มีสิ่งเร้า และ (ค) ความสามารถเปลี่ยนกฎเกณฑ์บางประการโดยที่สภาวะของมันก็เปลี่ยนแปลงไปด้วย หาก AI มีเงื่อนไขครบ 3 ข้อนี้ก็ถือว่าเรียกได้เป็น AMA

แม้ว่าผู้เขียนจะสนับสนุนแนวคิดที่ว่าความรับผิดชอบทางจริยธรรมของ AI จำเป็นต้องยึดโยงกับการเป็นผู้กระทำการทางจริยธรรม แต่ในขณะเดียวกัน ผู้เขียนก็ไม่เห็นด้วยในกรณีที่อาจมีความเป็นไปได้ในการนำมโนทัศน์เรื่อง AMA ไปผูกโยงกับความรับผิดชอบทางจริยธรรมที่เป็นอิสระจากมนุษย์ เนื่องด้วยผู้เขียนปฏิเสธว่า AI จะทำงานอย่างเป็นอิสระจากมนุษย์และสิ่งอื่นได้ ดังที่ Dodig–Crnkovic, & Persson (2008) ชี้ว่าปฏิบัติการของ AI เป็นระบบแบบเทคโนสังคม (socio–technological systems) ปฏิบัติการและสภาวะเกี่ยวข้องดังกล่าวจึงไม่สามารถแยกขาดจากมนุษย์และสังคมได้ แม้ว่าผู้เขียนจะเห็นว่าในกรณีของ AI ‘การเป็นผู้กระทำการทางจริยธรรม’ เป็นเงื่อนไขจำเป็นของ ‘ความรับผิดชอบทางจริยธรรม’ ทว่า ‘ผู้กระทำการทางจริยธรรม’ ดังกล่าวนี้ ต้องไม่ใช่ AI ที่เป็นอิสระจากมนุษย์ แต่เป็นปฏิสัมพันธ์

ร่วมกันระหว่าง AI กับมนุษย์ ดังที่จะอภิปรายต่อไปในข้อเสนอหลักข้อที่ 2 ในส่วนท้ายของบทความ

อย่างไรก็ตาม ในตอนนี้ผู้เขียนขอใช้พื้นที่ตรงนี้อภิปรายต่อถึงประเด็นที่ผู้เขียนกำลังอ้างว่าในกรณีของ AI นั้น 'การเป็นผู้กระทำการทางจริยธรรม' เป็นเงื่อนไขจำเป็นของ 'ความรับผิดชอบทางจริยธรรม' มาถึงตรงนี้ ผู้อ่านอาจจะรู้สึกว่ามีข้อโต้แย้งต่อข้ออ้างนี้ของผู้เขียนใน 2 ประการ ดังนี้

ข้อโต้แย้งประการแรก ในกรณีที่ลองย้อนกลับไปใช้กรอบคิดดั้งเดิมแบบอาริสโตเติลก็อาจกล่าวได้ว่า สมมุติว่าแม้เราจะสร้าง AI ที่สามารถมีความรู้และมีเจตจำนงเสรีได้สำเร็จตามวิธีคิดแบบมานุษยรูปนิยม กระทั่งเราเข้าใจว่า AI เป็นผู้กระทำการทางจริยธรรมได้จริง ๆ เนื่องจาก AI รู้ว่าตัวเองทำอะไรอยู่และเลือกการกระทำได้อย่างอิสระ (แบบที่มนุษย์เป็นและเข้าเงื่อนไขแบบอาริสโตเติล) แต่นี่ก็ไม่ได้รับประกันว่า AI จะสามารถรับผิดชอบทางจริยธรรมในสิ่งที่มันกระทำได้ (Tigard, 2021, p. 436) ดังนั้น ความสามารถในการกำหนดการเป็นผู้กระทำการทางจริยธรรม (ตามกรอบคิดแบบดั้งเดิม) ให้กับ AI ได้นั้นไม่ได้นำไปสู่ผลลัพธ์ที่ว่าเราจะสามารถกำหนดความรับผิดชอบทางจริยธรรมให้ AI ได้ด้วย

ข้อโต้แย้งประการที่สอง ในกรณีที่พยายามสร้าง AMA ที่ไม่จำเป็นต้องเหมือนมนุษย์ การสร้าง AMA ได้สำเร็จก็ไม่ได้นำไปสู่ความรับผิดชอบทางจริยธรรม กล่าวคือ การที่ AMA แสดงออกถึงการเป็นผู้กระทำการในเชิงจริยธรรมนั้น ไม่ได้ส่งผลผูกพันว่ามันจะต้องมีความรับผิดชอบทางจริยธรรมอย่างเป็นอิสระจากมนุษย์ด้วยเสมอไป เราไม่แน่ใจว่า AI จะสามารถมีความรับผิดชอบทางจริยธรรมได้หรือไม่ แม้ว่ามันจะเรียกได้ว่าเป็น 'ผู้กระทำการทางจริยธรรมแบบประดิษฐ์' หรือ AMA ก็ตาม ดังที่ Nyholm (2020, p. 55) ชี้ว่า แม้การสร้าง AMA อาจจะสำเร็จในแง่ที่ออกแบบโปรแกรมให้ AI สามารถตัดสินใจทางจริยธรรมตามหลักการทางจริยธรรมที่ยอมรับได้ แต่ก็ไม่ได้หมายความว่า AMA เช่นนี้จะสามารถรับผิดชอบทางจริยธรรมได้ด้วย เนื่องจากเป็นเพียงการสร้างเครื่องจักรที่ปฏิบัติตามหลักเกณฑ์ทางจริยธรรมเท่านั้น โดยไม่ได้รับประกันว่าเครื่องจักรเหล่านี้จะสามารถมีความรับผิดชอบทางจริยธรรมได้ด้วย การที่จักรกลสามารถตัดสินใจและปฏิบัติตามหลักเกณฑ์ทางจริยธรรมได้เป็นเพียงการบ่งชี้ว่ามันสามารถคิดได้อย่างมีเหตุผลตามหลักเกณฑ์ทางจริยธรรม แต่ไม่ได้เกี่ยวข้องอะไรกับการบอกว่ามันจะสามารถรับผิดชอบในสิ่งที่มันตัดสินใจได้ด้วยตัวของมันเอง

ผู้เขียนเห็นว่าข้อโต้แย้งข้างต้นอาจจะฟังขึ้นก็ต่อเมื่อมันเป็นข้อโต้แย้งที่มีต่อ (i) ฐานคิดแบบดั้งเดิมของอาริสโตเติล และ/หรือ เป็นข้อโต้แย้งที่มีต่อ (ii) ฐานคิดที่จะสร้าง AMA ที่สามารถรับผิดชอบอย่างเป็นอิสระจากมนุษย์ อย่างไรก็ตาม ผู้เขียนเห็นว่าข้อโต้แย้งนี้ไม่ใช่สิ่งที่แย้งกับข้อเสนอหลักข้อที่ 1 ของผู้เขียนโดยตรง และในทางตรงกันข้ามกลับกลายเป็นข้อโต้แย้งที่ผู้เขียนค่อนข้างเห็นด้วย เนื่องจากมันเป็นข้อโต้แย้งเพื่อหักล้างต่อทั้งฐานคิด (i) และ (ii) ซึ่งเป็นฐานคิดที่ผู้เขียนปฏิเสธเช่นกัน

ประการสำคัญ ผู้เขียนขอเน้นย้ำว่าแท้จริงแล้ว ‘ความรับผิดชอบทางจริยธรรมอย่างเป็นอิสระจากมนุษย์’ ต่างหากคือสิ่งที่ผู้เขียนปฏิเสธ หมายความว่าเมื่อพิจารณาในเรื่อง AI กับความรับผิดชอบทางจริยธรรม ผู้เขียนไม่ได้ปฏิเสธความจำเป็นในการเชื่อมโยง ‘การเป็นผู้กระทำการทางจริยธรรม’ เข้ากับความรับผิดชอบทางจริยธรรมในตัวมันเอง ทว่า ผู้เขียนมุ่งปฏิเสธเฉพาะการเป็นผู้กระทำการทางจริยธรรมและความรับผิดชอบทางจริยธรรมที่เป็นอิสระจากมนุษย์ พูดอีกอย่างหนึ่งคือ ผู้เขียนไม่ได้สนับสนุนว่า ‘การเป็นผู้กระทำการทางจริยธรรม (ที่เป็น AI ที่มีจิตสำนึกและเจตจำนงเสรีที่เป็นอิสระจากมนุษย์ และ/หรือ เป็น AMA ที่เป็นอิสระจากมนุษย์)’ เป็นเงื่อนไขจำเป็นของ ‘ความรับผิดชอบทางจริยธรรมอย่างเป็นอิสระจากมนุษย์’ แต่ผู้เขียนกำลังสนับสนุนแนวคิดที่ว่า ‘การเป็นผู้กระทำการทางจริยธรรม (แบบปฏิสัมพันธ์ร่วมกันระหว่าง AI กับมนุษย์)’ เป็นเงื่อนไขจำเป็นของ ‘ความรับผิดชอบทางจริยธรรมที่ไม่ได้เป็นอิสระจากมนุษย์’

นอกจากนี้ ทั้งฐานคิดแบบ (i) และ (ii) ที่กล่าวมาด้านบน (ซึ่งเป็นวิธีคิดที่ผู้เขียนปฏิเสธ) ล้วนแต่พุ่งเป้าไปที่ผู้กระทำการที่สามารถรับผิดชอบด้วยตนเองอย่างเป็นอิสระ โดยเฉพาะในแง่ที่เป็นปัจเจกบุคคลอย่างในกรณี ‘ผู้เดียว’ หรือ ‘คน ๆ เดียว’ หรือ ‘สิ่ง ๆ เดียว’ การพยายามสร้างแนวคิดเพื่อแก้ปัญหาช่องว่างทางจริยธรรมทั้งหมดด้านบนที่กล่าวมาจนถึงตอนนี้ยังไม่ได้พิจารณาแนวคิดที่พยายามสร้างหน่วยที่มารับผิดชอบทางจริยธรรมในลักษณะแบบรวมหมู่ (collective) หรือแบบร่วมกัน (shared) ซึ่งวิธีคิดแบบนี้ยิ่งหนีห่างออกจากความคิดดั้งเดิมทั้งหมดที่กล่าวมาในบทความนี้ แน่นอนว่าหนีห่างจากกรอบคิดดั้งเดิมแบบอาริสโตเติลที่มีลักษณะให้ความสำคัญกับปัจเจกบุคคลและมุ่งเน้นการใช้เฉพาะกับมนุษย์ การกล่าวถึงความรับผิดชอบทางจริยธรรมในลักษณะแบบรวมหมู่หรือแบบร่วมกันถือเป็นอีกมโนทัศน์ทางเลือกหนึ่งที่สำคัญในทางปรัชญา ดังเช่นที่พบในงานวิจัยที่ถกเถียงเรื่องการมุ่งสู่เป้าหมายของจิตแบบรวมหมู่ (collective intentionality) (Schweikard & Schmid, 2021) อย่างไรก็ตาม ข้อถกเถียงของนักปรัชญาในกลุ่มนี้มักจะเน้นเฉพาะ

ลักษณะร่วมหมู่ระหว่าง ‘มนุษย์กับมนุษย์’ ด้วยกันเองเท่านั้น ยังไม่ได้กล่าวถึงปฏิสัมพันธ์ระหว่างมนุษย์กับสิ่งอื่น เช่น ‘มนุษย์กับ AI’ ดังนั้น หากเราลองใช้แนวโน้มเรื่องลักษณะรวมหมู่ดังกล่าวเป็นจุดตั้งต้น แล้วลองขยายขอบเขตของความรวมหมู่จากลักษณะรวมหมู่ระหว่างมนุษย์กับมนุษย์ไปสู่ลักษณะรวมหมู่ระหว่างมนุษย์กับ AI โดยเชื่อมโยงกับประเด็นเรื่องการเป็นผู้กระทำการทางจริยธรรมในฐานะที่เป็นปฏิสัมพันธ์ร่วมกันระหว่างมนุษย์กับ AI เราอาจพบความเป็นไปได้ที่อาจจะสามารถช่วยค้นหามโนทัศน์ทางเลือกเรื่องความรับผิดทางจริยธรรมของ AI ในแบบอื่นเพื่อแก้ปัญหาเรื่องช่องว่างของความรับผิดชอบ ดังที่ผู้เขียนจะอภิปรายในส่วนถัดไป

ข้อเสนอหลักข้อที่ 2: มโนทัศน์เรื่องการเป็นผู้กระทำการทางจริยธรรมในฐานะที่เป็นปฏิสัมพันธ์ร่วมกันระหว่างมนุษย์กับปัญญาประดิษฐ์เป็นสิ่งจำเป็นต่อความรับผิดชอบทางจริยธรรมของ AI

ในส่วนนี้ ผู้เขียนจะนำเสนอการสนับสนุนมโนทัศน์เรื่องการเป็นผู้กระทำการทางจริยธรรมในฐานะที่เป็นปฏิสัมพันธ์ร่วมกันระหว่างมนุษย์กับ AI และแสดงให้เห็นว่าเป็นมโนทัศน์ที่จำเป็นต่อความรับผิดชอบทางจริยธรรม ข้อเสนอดังกล่าวจำเป็นต้องพิจารณากรอบคิดทางเลือกแบบอื่น ซึ่งผู้เขียนเห็นว่า แม้เราจะยังคงมุ่งเน้นไปที่การเป็นผู้กระทำการทางจริยธรรม แต่เราก็ไม่จำเป็นต้องยึดถือในเรื่องการพิจารณาเนื้อหาของจิตเป็นสำคัญ ดังที่ Floridi (2016) เสนอว่าหากเราจะทำความเข้าใจมโนทัศน์เรื่องความรับผิดชอบทางจริยธรรมแบบปันส่วน (distributed moral responsibility) เราควรยืนอยู่บนวิธีคิดที่ไม่ได้มุ่งเน้นเนื้อหาของจิตและกระทำที่จงใจ (intentional action) ในแง่นี้ พึงสังเกตว่าผู้เขียนไม่ได้มุ่งปฏิเสธความเห็นสาเหตุในตัวมันเอง เนื่องจากผู้เขียนเน้นย้ำว่าผู้กระทำการทางจริยธรรมยังถือเป็นสาเหตุที่นำไปสู่ความรับผิดชอบทางจริยธรรม แต่ผู้เขียนปฏิเสธเฉพาะการสืบกลับเชิงสาเหตุที่มุ่งเน้นแนวคิดแบบภายในของมนุษย์ที่เป็นปัจเจกบุคคลตามแนวคิดอาริสโตเติลและตามการพิจารณาเนื้อหาจิตและความจงใจ

การสร้างมโนทัศน์เรื่องการเป็นผู้กระทำการทางจริยธรรมในฐานะที่เป็นปฏิสัมพันธ์ร่วมกันระหว่างมนุษย์กับ AI ดังกล่าวมีรากฐานทางทฤษฎีต่าง ๆ ที่สอดคล้องกับสิ่งที่ผู้เขียนจะนำเสนอ ยกตัวอย่างเช่น การตีความว่ามิติความสัมพันธ์เชิงวัตถุสามารถมีลักษณะของการเป็นผู้กระทำการได้ในบางความหมาย และเครือข่ายเหล่านี้ (ซึ่งรวมถึงมนุษย์) ได้ร่วมกันกระทำการรังสรรค์สิ่งต่าง ๆ ขึ้นมาเป็นสังคมที่เราอาศัยอยู่ หรือตามที่ Verbeek (2005) เรียกว่า ‘การไกล่เกลี่ยประสานเชิงเทคโนโลยี’ (technological mediation) ที่เสนอว่าเทคโนโลยีได้ก่อรูปความเข้าใจ

การรับรู้ การกระทำ และพฤติกรรมของมนุษย์ นั่นหมายความว่าเทคโนโลยีมีส่วนประสานในการกระทำของมนุษย์ นอกจากนี้ ยังมีกรอบทฤษฎีในลักษณะเดียวกัน เช่น แนวคิดเรื่องการเป็นผู้กระทำการยืดขยาย (extended agency) ของ Hanson (2009) ที่เสนอว่าเทคโนโลยีมีส่วนช่วยส่งเสริมและเป็นส่วนขยายการกระทำของมนุษย์ ซึ่งอาจเรียกรวมกันได้ว่าเป็น 'ผู้กระทำการยืดขยาย' และอาจมีความเป็นไปได้ที่จะพูดถึง 'ความรับผิดชอบร่วม' (joint responsibility) ซึ่งเป็นวิธีคิดที่ตรงข้ามกับวิธีคิดแบบปัจเจกนิยมเชิงจริยธรรม (moral individualism) และยังมีอีกหนึ่งแนวคิด คือเรื่อง 'สิ่งที่ยืดขยายด้วย AI' (AI–extender) ของ Telakivi (2021) ที่ชี้ว่า AI เป็นส่วนขยายประธาน (cognition) ของมนุษย์ เป็นการผสาน AI เข้ากับมนุษย์กลายเป็น 'ระบบผสม' (hybrid system) ที่เข้าคู่กันอย่างแนบแน่น (tightly coupled) ในขณะเดียวกับภายใต้ระบบผสมนี้ AI อาจจะมีการกำกับความคิด (manipulation) ของมนุษย์ในทางใดหนึ่ง สิ่งที่น่าสนใจคือ Telakivi ตั้งคำถามว่าในความเป็นระบบผสมนั้น AI ได้เข้ามาเพิ่มพูนหรือลดทอนความเป็นผู้กระทำการทางจริยธรรมกันแน่ ประการสำคัญ Telakivi ได้ตั้งคำถามถึงความสัมพันธ์ระหว่างระบบผสมดังว่าส่งผลต่อความรับผิดชอบทางจริยธรรมหรือไม่ และอย่างไร นอกจากนี้ ยังมี Nyholm (2020) ที่เสนอว่าเราสามารถนึกถึง AI ในฐานะที่เป็นผู้กระทำการรูปแบบหนึ่งได้ในความหมายของการเป็น 'ผู้กระทำการที่ทำงานร่วมกัน' (collaborative agency) โดยที่จะต้องมีมนุษย์เป็นผู้ร่วมกระทำหลัก (key partner) หมายถึงว่า การเป็นผู้กระทำการที่ทำงานร่วมกันนี้มีรูปแบบของลำดับชั้น (hierarchical model) ในแง่ที่ผู้กระทำการประเภทที่เป็น AI กระทำการร่วมกันในลักษณะที่อยู่ภายใต้การดูแล (supervision) และอยู่ภายใต้อำนาจ (authority) ของผู้กระทำการหลักที่เป็นมนุษย์

ดังที่ได้กล่าวมาแล้วก่อนหน้านี้ จุดเริ่มต้นในข้อเสนอหลักข้อที่ 1 ของบทความนี้ ผู้เขียนเสนอว่าความรับผิดชอบทางจริยธรรมของ AI จำเป็นต้องยึดโยงกับการเป็นผู้กระทำการทางจริยธรรมในแบบที่ไม่ใช้กรอบคิดแบบดั้งเดิม ในขณะเดียวกัน ผู้เขียนยังคงเชื่อว่า 'การเป็นผู้กระทำการทางจริยธรรม' เป็นเงื่อนไขจำเป็นของความรับผิดชอบทางจริยธรรมของ AI คำถามที่สำคัญในตอนนี้คือ หากผู้เขียนปฏิเสธกรอบคิดแบบดั้งเดิมของอาริสโตเติล รวมถึงปฏิเสธการมุ่งพิจารณาที่เนื้อหาจิตและความจงใจแล้วอะไรคือข้อเสนอทางเลือกจะนำมาทดแทนการกำหนดการเป็นผู้กระทำการทางจริยธรรมในความหมายแบบดั้งเดิมนี้ ผู้เขียนขอนำเสนอคุณลักษณะทางเลือกที่จะนำมาทดแทนดังกล่าวโดยแบ่งเป็น 2 ประการ

ประการแรก ผู้เขียนเห็นว่าการหันมาพิจารณาลักษณะเชิงพฤติกรรมภายนอกของประธานหรือผู้กระทำการอาจเป็นทางเลือกที่ดี สอดคล้องกับที่ Gogoshin (2021) เสนอให้พิจารณาลักษณะภายนอกของ AI ที่แสดงให้เห็นถึงความรู้สึกไวทางจริยธรรม (moral sensitivity) เช่น การตอบสนองต่อสถานการณ์ด้วยการแสดงให้เห็นถึงหน้าจอที่กำลังประมวลผลเพื่อตัดสินใจคิดหาทางช่วยเหลือเมื่อเห็นคนถูกทำร้าย หรือการที่ AI แสดงออกถึงพฤติกรรมที่ตอบสนองความคาดหวังของสังคม เช่น AI มีพฤติกรรมที่แสดงให้เห็นถึงความรู้สึกไวที่ตนเองกำลังถูกตำหนิ หรือได้รับการชมเชย หรือไม่ได้รับการยอมรับทางจริยธรรม ฯลฯ พฤติกรรมเช่นนี้ถูกกำหนดได้โดยบรรทัดฐานและปฏิบัติการของสังคม ดังนั้น AI ที่มีพฤติกรรมที่ปฏิบัติตามกฎเกณฑ์ของสังคมก็อาจมีความหมายในบางแง่มุมว่ามันอาจจะสามารถรับผิดชอบทางจริยธรรมได้ด้วย ในแง่นี้ เราจึงอาจนับรวม AI ให้เป็นสมาชิกของชุมชนทางจริยธรรรมของมนุษย์ได้ด้วย แม้ว่าอาจจะยังมีบกพร่องและไม่สมบูรณ์อยู่บ้าง อย่างไรก็ตาม ประเด็นนี้อาจถูกโต้แย้งได้โดยง่ายด้วยการตั้งคำถามว่า AI ที่มีพฤติกรรมที่แสดงความไวทางจริยธรรมดังกล่าวนั้นจะมีลักษณะเป็นเช่นไร และเป็นเรื่องน่าเชื่อถือได้จริงหรือไม่ว่าลำพังวัตถุอย่าง AI จะสามารถแสดงให้คนอื่นเชื่อได้ว่ามันมีความใส่ใจทางจริยธรรมจริง ๆ ยิ่งไม่ต้องกล่าวถึงความรับผิดชอบทางจริยธรรม

ในประเด็นโต้แย้งนี้ ผู้เขียนขอแย้งกลับด้วยการอภิปรายเสริมในสิ่งที่ Gogoshin อาจจะไม่ได้ระบุไว้ กล่าวคือผู้เขียนเสนอว่า AI ที่เรากำลังพูดถึงนี้ต้องเป็น AI ที่กำลังปฏิบัติการในท่ามกลางการมีปฏิสัมพันธ์ร่วมกันกับมนุษย์ เพื่อให้เกิดการแสดงพฤติกรรมร่วมกันในแบบที่มีความเป็นมนุษย์ผสมอยู่มากขึ้น ดังนั้น การนับรวม AI เข้าเป็นสมาชิกของชุมชนทางจริยธรรรมนี้จึงหมายถึงการนับเอา AI รวมเข้าบนสนามแห่งปฏิสัมพันธ์ของมนุษย์ ในแง่นี้ปฏิสัมพันธ์บางประการระหว่าง AI กับมนุษย์จึงเกิดขึ้นได้ และเมื่อผู้กระทำการทางจริยธรรมแบบผสมระหว่างมนุษย์กับ AI ดังกล่าวแสดงพฤติกรรมภายนอกที่บ่งชี้ถึงความใส่ใจในเรื่องจริยธรรมก็จะทำให้น่าเชื่อได้มากขึ้นว่ามันอาจนำไปสู่ความรับผิดชอบทางจริยธรรมได้ด้วย

ผู้เขียนขอยกตัวอย่างโดยการให้ลองจินตนาการถึงแพทย์ที่ใช้หูฟัง AI หรือติดตั้งอุปกรณ์ AI ประเภทใดก็ตามที่ผู้อ่านจะจินตนาการถึงได้ โดย AI เหล่านี้เป็น AI ที่สามารถประมวลผลและตัดสินใจทางด้านการแพทย์ได้ สมมติว่าแพทย์คนนี้กำลังปฏิบัติหน้าที่ด้วยการมีปฏิสัมพันธ์กับ AI ที่อยู่ในอุปกรณ์ต่าง ๆ เหล่านี้ เมื่อแพทย์เผชิญสถานการณ์ทางจริยธรรมบางประการ อุปกรณ์ AI อาจส่งสัญญาณเสียงดัง

หรือเปล่งเสียงสังเคราะห์ออกมาเป็นภาษามนุษย์เพื่อบ่งชี้สภาวะเร่งด่วน หรือแสดงหน้าจอให้เห็นถึงการประมวลผลที่กำลังคิดคำนวณและตัดสินใจเพื่อค้นหาคำตอบที่ถูกต้องตามหลักจรรยาบรรณวิชาชีพหรือถูกหลักการจริยธรรมในสถานการณ์ฉุกเฉิน ในส่วนของมนุษย์ก็จะมีสภาวะแสดงออกทางใบหน้าโดยธรรมชาติของมนุษย์ที่มีความรู้สึกไวทางจริยธรรมว่าตนต้องกระทำอะไรบางอย่าง หรือมีพฤติกรรมแสดงออกถึงการครุ่นคิดเพื่อหาคำตอบและตัดสินใจ สภาวะพฤติกรรมภายนอกทั้งหมดนี้สามารถถูกมองรวม ๆ ได้ว่าเป็นหน่วยของผู้กระทำการแห่งสนามปฏิสัมพันธ์ระหว่างมนุษย์กับ AI ที่อยู่ในสภาวะแสดงออกถึงพฤติกรรมที่มีความใส่ใจทางจริยธรรมและพร้อมที่จะรับผิดชอบทางจริยธรรมต่อผลที่ตามมาจากการตัดสินใจ ซึ่งเป็นการตัดสินใจร่วมกันระหว่างการคิดในสมองของแพทย์และการประมวลผลเพื่อให้คำแนะนำโดย AI (AI recommendation)

ประการที่สอง ผู้เขียนเสนอว่าคุณลักษณะทางเลือกที่จะนำมาทดแทนกรอบคิดแบบดั้งเดิมคือการระบุหน้าที่ (duty) และพันธะหน้าที่ (obligation) ให้กับผู้กระทำการ (ทั้งมนุษย์และ AI) ทางเลือกนี้คือวิธิคิดเรื่องความรับผิดชอบทางจริยธรรมในแบบที่มองไปข้างหน้า (forward-looking approach) เป็นการมอบหมายความรับผิดชอบเพื่อเตรียมพร้อมรับผิดชอบต่อสิ่งที่ยังไม่เกิด เป็นการทำตามหน้าที่หรือพันธะหน้าที่บางประการที่ถูกกำหนดไว้ก่อนแล้วล่วงหน้า เช่น การกำหนดให้บริษัทที่ผลิต AI ต้องรับผิดชอบในบางส่วน (ไม่ใช่ทั้งหมด) ในขณะที่บุคคลผู้ใช้เทคโนโลยีต้องรับผิดชอบอีกส่วน วิธีการนี้อาจกำหนดให้ผู้ใช้ AI ต้องยืนยันเจตจำนงในการใช้ AI นั้น ๆ เป็นเช่นว่า ผู้ใช้ยินยอมต่อการถูกกำกับความคิดโดย AI หรือ ผู้ใช้ยินยอมต่อการถูก AI กำกับและลดทอนอัตตาณัติ หรือ ผู้ใช้จะมีส่วนร่วมในความรับผิดชอบทางจริยธรรมร่วมกับบริษัทผู้ผลิต AI ในกรณีที่ผู้ใช้กระทำการตัดสินใจร่วมกับ AI เช่น การที่ผู้ใช้เลือกที่จะเชื่อตามระบบการให้คำแนะนำโดย AI หรือเชื่อตามระบบการคาดการณ์ของ AI (predictive AI)

นอกจากนี้ ผู้เขียนขอนำเสนอข้อเสนอเชิงปทัสถานซึ่งเกี่ยวกับทิศทางทางออกแบบและสร้างเทคโนโลยีที่เน้นการปรับเทคโนโลยีให้เข้ามาสอดคล้องกับความเป็นมนุษย์ โดยผู้เขียนเรียกว่า 'ทิศทางที่ปรับจาก AI สู่มนุษย์' (AI-to-human direction) ในการกำหนดหน้าที่และพันธะหน้าที่ให้กับ AI (กรณีนี้รวมถึงบริษัทผู้ผลิตสร้าง AI) ถือเป็นการใช้วิธีคิดเรื่องความรับผิดชอบทางจริยธรรมในแบบที่มองไปข้างหน้าเพื่อกำกับให้วิถีทางของการพัฒนา AI สอดคล้องกับเป้าหมายและวัฒนธรรมของ

มนุษย์ กล่าวคือ การให้คุณค่ากับจริยธรรมและความรับผิดชอบทางจริยธรรม วิธีคิดนี้สอดคล้องกับที่ Tegmark (2018) และ Russell (2019) เสนอว่าเราควรต้องสร้าง AI ที่มีประโยชน์ (beneficial AI) ด้วยการกำหนดเป้าหมายของ AI ให้สอดคล้องกับเป้าหมายของมนุษย์ และสอดคล้องกับที่ Nyholm (2020) กล่าวถึงวิธีการคิดเรื่องความสัมพันธ์ระหว่างมนุษย์กับ AI ไว้อย่างน่าสนใจว่า มนุษย์ไม่สามารถปรับตัวหรือวิวัฒน์ตัวเองได้ทันกับความก้าวหน้าอย่างรวดเร็วของเทคโนโลยี ในขณะเดียวกับที่มนุษย์จำเป็นต้องมีปฏิสัมพันธ์และผสานเข้ากับเทคโนโลยีอย่างแยกกันไม่ออก ความเสี่ยงและความท้าทายจึงเกิดขึ้นอย่างหลีกเลี่ยงไม่ได้ ดังนั้น Nyholm จึงเสนอว่าทางออกหนึ่งของการแก้ปัญหาคือการตั้งเป้าหมายที่มุ่งไปสู่การพยายามสร้างและปรับทิศทางของเทคโนโลยีให้สามารถเข้ากันได้กับความเป็นมนุษย์ในปัจจุบัน

ผู้เขียนเห็นว่าในแง่หนึ่ง การพยายามการสร้างมโนทัศน์เรื่องการเป็นผู้กระทำการทางจริยธรรมในฐานะที่เป็นปฏิสัมพันธ์ร่วมกันระหว่างมนุษย์กับ AI ดังที่กล่าวมาข้างต้นเป็นการพยายามดำรงคุณค่าศักดิ์ศรีความเป็นมนุษย์ กล่าวคือ แม้ว่าข้อเสนอในลักษณะนี้จะมีนัยว่ามนุษย์ไม่ได้เป็นศูนย์กลางของทุกสิ่งอีกต่อไป เนื่องจากวัตถุและเทคโนโลยีล้วนมีส่วนกำกับหรือส่งอิทธิพลต่อความคิด การตัดสินใจ และการกระทำของมนุษย์ กระทั่งเรายอมให้ AI มีส่วนร่วมหรือมีฐานะเป็นผู้กระทำการร่วมกับเราได้ แต่ก็ไม่ได้หมายความว่าคุณค่าของมนุษย์จะต้องสูญเสียไปด้วย มนุษย์ยังคงมีศักดิ์ศรีที่จะสร้างสรรค์และนำพาเป้าหมายและคุณค่าต่าง ๆ ของมนุษยชาติอยู่ และการที่มนุษย์มีความสามารถบางอย่าง (หรือหลายอย่าง) ที่สู้ AI ไม่ได้ ประกอบกับบทบาทที่เพิ่มมากขึ้นของ AI ที่เริ่มเข้ามายึดครองความเป็นศูนย์กลางในกิจกรรมต่าง ๆ ของมนุษย์ แต่ก็ไม่ได้หมายความว่าคุณค่าและศักดิ์ศรีมนุษย์จะถูกทำลายไปด้วย (Zerilli et al., 2021, p. xviii) นั่นหมายความว่า การบอกว่ามนุษย์ไม่ได้เป็นศูนย์กลางของสิ่งต่าง ๆ ไม่ได้หมายความว่าคุณค่าเรื่องศักดิ์ศรีความเป็นมนุษย์จะต้องสูญหายไปด้วย และการที่มนุษย์ต้องเป็นผู้มีส่วนร่วมในความรับผิดชอบทางจริยธรรมก็ถือเป็นหนึ่งในเครื่องยืนยันศักดิ์ศรีความเป็นมนุษย์

อย่างไรก็ตาม ข้อเสนอหลักที่ผู้เขียนได้อภิปรายด้านบนอาจมีข้อโต้แย้ง 2 ประการ *ประการแรก* เป็นตามที่ Hakli and Mäkelä (2019) ได้โต้แย้งแนวคิดเรื่องการเป็นผู้กระทำการที่ทำงานร่วมกัน (ระหว่างมนุษย์และ AI) และแนวคิดอื่นที่คล้ายกัน โดยแย้งว่าแม้ว่า AI จะมีอิทธิพลต่อมนุษย์ในการกระทำสิ่งต่าง ๆ แต่ AI ก็อาจมีส่วนร่วมเพียงในเชิงสาเหตุ (casual contribution) ของการกระทำเท่านั้น

ไม่ได้เป็นการมีส่วนร่วมเชิงประกอบสร้าง (constitutive) ดังนั้น เราจึงจะไม่สามารถนับรวมว่า AI เป็นส่วนประกอบสร้างในเชิงสารัตถะของการเป็นผู้กระทำการแบบผสมระหว่างมนุษย์กับ AI ได้ แต่ AI เป็นเพียงสาเหตุที่ส่งอิทธิพลต่อความคิดมนุษย์เท่านั้น ในแง่นี้ ความรับผิดชอบทางจริยธรรมยังคงต้องเป็นบทบาทของมนุษย์เท่านั้นไม่ว่าจะเป็นมนุษย์ในฐานะปัจเจกบุคคลหรือในแบบรวมหมู่ระหว่างมนุษย์กับมนุษย์ด้วยกันเอง ในประเด็นนี้ ผู้เขียนเห็นว่าหากมนุษย์เชื่อการตัดสินของ AI ในทันทีทันใดโดยที่ไม่ได้คิดแบบวิพากษ์จากอัตตาณัติของตนเองเลย หรือแม้มนุษย์จะคิดเชิงวิพากษ์ในระดับหนึ่งแต่ก็เลือกที่จะตัดสินใจหรือกระทำการร่วมกันกับ AI ในภาวะที่ลื่นไหลเป็นธรรมชาติแบบทันทีทันใด ก็อาจจะถือได้ว่าเป็นสภาวะของผู้กระทำการร่วมกันในเชิงประกอบสร้าง แต่ถ้ามนุษย์ไม่ได้เชื่อ AI ทั้งหมดในทันที และมีการใช้สมองตนเองคิดพิจารณาสิ่งที่ AI แนะนำ แล้วนำมาปรับปรุงบางส่วนหรือหลายส่วนแล้วจึงตัดสินใจในภายหลังอย่างรอบคอบในระยะเวลาที่ไม่ได้มีลักษณะฉับพลันแบบนี้ก็อาจถือได้ว่าเป็นสภาวะของการกระทำการที่ AI มีส่วนร่วมในเชิงสาเหตุ (casual) ซึ่งเป็นคุณลักษณะที่อาจยอมรับได้ยากกว่าเป็นปฏิสัมพันธ์ร่วมกันอย่างแท้จริงระหว่างมนุษย์กับ AI ผู้เขียนยอมรับว่าข้อเสนอของผู้เขียนอาจใช้ไม่ได้กับทุกกรณี แต่ผู้เขียนมีความเห็นว่าในอนาคตอันใกล้ ยิ่งมนุษย์พึ่งพา AI ในการทำสิ่งต่าง ๆ แทนมนุษย์หรือทำร่วมกับมนุษย์ในทุกกิจกรรมและอย่างฉับพลันทันทีทันใดในภาวะลื่นไหลเป็นธรรมชาติมากขึ้นเท่าใด (กระทั่งเป็นไปได้ว่ามนุษย์แทบไม่ได้คิดอะไรเองเลย) เมื่อนั้นลักษณะการเป็นผู้กระทำการร่วมกันระหว่างมนุษย์กับ AI ในเชิงประกอบสร้างน่าจะมีให้เห็นเป็นตัวอย่างมากขึ้นเท่านั้น

ประการที่สอง เนื่องจากการมีปฏิสัมพันธ์ระหว่างมนุษย์และ AI ในระดับที่ควบรวมทั้งสองให้เป็นผู้กระทำการร่วมกันเป็นสิ่งที่ดูคลุมเครือ ผู้เขียนจึงขอระบุเพื่อชี้ชัดว่าในกรณีที่ผู้เขียนกล่าวถึงในบทความนี้ อาจจะมีความเป็นไปได้สำหรับใช้อธิบาย AI แค่บางประเภท เช่น AI ที่มีลักษณะเป็น ‘สิ่งที่ยืดขยายด้วย AI’ ตามแบบที่ Telakivi (2021) เสนอ กล่าวคือ ‘ระบบผสม’ ที่มนุษย์ผสานร่างกายหรือจิตใจเข้ากับ AI ในระดับสูง โดยเฉพาะอย่างยิ่งต้องมีปฏิสัมพันธ์ระหว่างกันอย่างแนบแน่น นอกจากนี้ ต้องเป็น AI ที่มนุษย์ค่อนข้างจะสามารถควบคุมการทำงานได้ ซึ่งตรงนี้ไม่ได้หมายถึงการควบคุมผลการตัดสินใจทั้งหมดของ AI เพราะในหลายกรณีเราไม่สามารถรู้ได้ว่า AI จะตัดสินใจอย่างไร แต่ในที่นี้หมายถึงความสามารถในการควบคุมการทำงานเบื้องต้น เช่น มนุษย์สามารถกดปุ่มปิดระบบให้หยุดทำงานได้หากต้องการ ในแง่นี้ การพูดเรื่องผู้กระทำการร่วมระหว่างมนุษย์กับ AI ที่ส่งผลต่อความรับผิดชอบ

ทางจริยธรรม น่าจะใช้อธิบายได้ดีกับอุปกรณ์ AI ที่ใช้งานแบบผูกติดกับมนุษย์อย่างแนบแน่นทั้งในแง่กายภาพหรือในแง่จิตใจ เช่น อุปกรณ์สวมใส่ติดร่างกายหรืออุปกรณ์ใด ๆ ที่ต้องพกพาติดตัวอยู่เสมอและมนุษย์ต้องมีปฏิสัมพันธ์กับอุปกรณ์นั้นเสมอแม้ในเสี้ยววินาที และฉากทัศน์แบบนี้อาจเกิดขึ้นได้ในอนาคตอันใกล้

กล่าวโดยสรุป ในบทความนี้มีการนำเสนอและอภิปรายในประเด็นต่าง ๆ ตามลำดับดังนี้

1. ผู้เขียนปฏิเสธกรอบคิดเรื่องความรับผิดชอบทางจริยธรรมแบบดั้งเดิม เนื่องจากมีข้อจำกัดในการทำความเข้าใจเรื่องความรับผิดชอบทางจริยธรรมของ AI

2. ผู้เขียนเสนอว่าการเป็นผู้กระทำการทางจริยธรรมเป็นเงื่อนไขจำเป็นของความรับผิดชอบทางจริยธรรมของ AI แต่ว่าจะต้องยืนบนฐานคิดที่ไม่ใช้กรอบคิดแบบดั้งเดิม

3. แม้ว่าจะมีข้อเสนอเรื่อง AMA ที่มุ่งไปให้พ้นจากกรอบคิดแบบดั้งเดิม แต่ก็เป็นข้อเสนอที่ไปไกลเกินไป เนื่องจากแนวคิดนี้มีความเสี่ยงและมีแนวโน้มเสนอให้ AMA สามารถรับผิดชอบทางจริยธรรมอย่างเป็นอิสระจากมนุษย์ ซึ่งเป็นข้อเสนอที่ผู้เขียนปฏิเสธ

4. ผู้เขียนจึงมุ่งหาทางแก้ไขด้วยการนำเสนอมโนทัศน์เรื่องการเป็นผู้กระทำการทางจริยธรรมในฐานะที่เป็นปฏิสัมพันธ์ร่วมกันระหว่างมนุษย์กับปัญญาประดิษฐ์ และแสดงให้เห็นว่าเป็นมโนทัศน์ที่เป็นสิ่งจำเป็นต่อความรับผิดชอบทางจริยธรรมของ AI

ข้อเสนอแนะ

ข้อเสนอแนะทั่วไป

บทความวิจัยนี้เป็นการวิจัยเกี่ยวกับข้อถกเถียงในเชิงนามธรรมอันเป็นวิธีการวิจัยทางปรัชญาที่อยู่ในสาขามนุษยศาสตร์ แม้จะยังไม่สามารถนำผลวิจัยไปใช้ประโยชน์ได้ทันทีในเชิงนโยบายหรือเชิงปฏิบัติ แต่ก็น่าจะสามารถนำข้อสรุปของการวิจัยไปใช้เป็นจุดตั้งต้นของโครงการวิจัยเชิงพื้นฐานของความรู้เกี่ยวกับปัญญาประดิษฐ์ที่มุ่งเน้นการวิจัยเชิงมโนทัศน์เพื่อให้ได้ความรู้ใหม่ในเชิงนามธรรม อันเป็นลักษณะการวิจัยที่ขาดแคลนในประเทศไทย

ข้อเสนอแนะในการทำวิจัยครั้งต่อไป

การทำวิจัยประเด็นเรื่อง AI กับความรับผิดชอบทางจริยธรรมในอนาคตควรเปิดพื้นที่แบบสหวิทยาการให้ผู้เชี่ยวชาญด้านเทคนิคจากฝั่งวิทยาศาสตร์คอมพิวเตอร์ร่วมกันทำวิจัยกับนักสังคมศาสตร์และมนุษยศาสตร์ ทิศทางดังกล่าวน่าจะทำให้เกิดการวิจัยที่เป็นประโยชน์ในเชิงรูปธรรมมากขึ้น

กิตติกรรมประกาศ

บทความวิจัยนี้เป็นส่วนหนึ่งของโครงการวิจัยเรื่อง*การวิเคราะห์เชิงมโนทัศน์เรื่องปัญญาประดิษฐ์และความรับผิดชอบทางจริยธรรมร่วมกัน* ซึ่งเป็นโครงการวิจัยที่ได้รับทุนอุดหนุนการวิจัยจากมหาวิทยาลัยมหิดล

References

Anderson, M., & Anderson, S. L. (Eds.). (2011). *Machine ethics.* Cambridge: Cambridge University Press.

Bernáth, L. (2021). Can autonomous agents without phenomenal consciousness be morally responsible? *Philosophy & Technology, 34*(4), 1363–1382. doi: 10.1007/s13347-021-00462-7

Coeckelbergh, M. (2020a). *AI ethics.* Cambridge: The MIT Press.

Coeckelbergh, M. (2020b). Artificial Intelligence, responsibility attribution, and relational justification of explainability. *Science and Engineering Ethics, 26*(4), 2051–2068. doi: 10.1007/s11948-019-00146-8

Coeckelbergh, M. (2020c). *Introduction to philosophy of technology.* Oxford: Oxford University Press.

Dodig-Crnkovic, G., & Persson, D. (2008). Sharing moral responsibility with robots: A pragmatic approach. In A. Holst, P. Kreuger, & P. Funk (Eds.), *Proceedings of the 2008 conference on Tenth Scandinavian Conference on Artificial Intelligence: SCAI 2008* (pp. 165–168). Amsterdam: IOS Press.

Floridi, L. (2016). Faultless responsibility: On the nature and allocation of moral responsibility for distributed moral actions. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374*(2083), 20160112. doi: 10.1098/rsta.2016.0112

Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines, 14*(3), 349–379.

Gogoshin, D. L. (2021). Robot responsibility and moral community. *Frontiers in Robotics and AI, 8,* 768092. doi: 10.3389/frobt.2021.768092

Gunkel, D. J. (2020). Mind the gap: Responsible robotics and the problem of responsibility. *Ethics and Information Technology, 22*(4), 307–320. doi: 10.1007/s10676–017–9428–2

Hakli, R., & Mäkelä, P. (2019). Moral responsibility of robots and hybrid agents. *The Monist, 102*(2), 259–275. doi: 10.1093/monist/onz009

Hanson, F. A. (2009). Beyond the skin bag: On the moral responsibility of extended agencies. *Ethics and Information Technology, 11*(1), 91–99. doi: 10.1007/s10676–009–9184–z

Johnson, D. G. (2015). Technology with no human responsibility? *Journal of Business Ethics, 127*(4), 707–715. Retrieved from http://www.jstor.org/stable/24702822

Jonas, H. (1984). *The imperative of responsibility: In search of an ethics for the technological age.* Chicago: University of Chicago Press.

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology, 6*(3), 175–183. doi: 10.1007/s10676–004–3422–1

Nyholm, S. (2020). *Humans and robots: Ethics, agency, and anthropomorphism.* London: Rowman & Littlefield Publishing Group.

Owe, A., & Baum, S. D. (2021). Moral consideration of nonhumans in the ethics of artificial intelligence. *AI and Ethics, 1*(4), 517–528. doi: 10.1007/s43681–021–00065–0

Porter, Z. L. M., Habli, I., Monkhouse, H. & Bragg, J. E. (2018). *The moral responsibility gap and the increasing autonomy of systems.* Retrieved November 25, 2022, from https://eprints.whiterose.ac.uk/133488/

Russell, S. J. (2019). *Human compatible: Artificial intelligence and the problem of control.* New York: Viking.

Schweikard, D. P., & Schmid, H. B. (2021). *Collective intentionality.* Retrieved November 25, 2022, from https://plato.stanford.edu/archives/fall2021/entries/collective-intentionality

Sharkey, A. (2017). Can robots be responsible moral agents? And why should we care? *Connection Science, 29*(3), 210-216.

Tegmark, M. (2018). Life 3.0: *Being human in the age of artificial intelligence.* New York: Penguin Books.

Telakivi, P. (2021, July 13). *AI-extenders and moral responsibility* [Video file]. Retrieved from https://www.youtube.com/watch?v=Buna320N-bhE&t=6s

Tigard, D. (2021). Artificial moral responsibility: How we can and cannot hold machines responsible. *Cambridge Quarterly of Healthcare Ethics, 30*(3), 435-447. doi: 10.1017/S0963180120000985

Tigard, D. W. (2020). There is no techno-responsibility gap. *Philosophy & Technology 34*(3), 589–607. doi: 10.1007/s13347-020-00414-7

Tigard, D. W., Conradie, N. H., & Nagel, S. K. (2020). Socially responsive technologies: Toward a co-developmental path. *AI & SOCIETY, 35*(4), 885–893. doi: 10.1007/s00146-020-00982-4

Verbeek, P. P. (2005). *What things do: Philosophical reflections on technology, agency, and design.* Pennsylvania: Pennsylvania State University Press.

Wallach, W., & Allen, C. (2009). *Moral machines.* Oxford: Oxford University Press.

Zerilli, J., Danaher, J., Maclaurin, J., Gavaghan, C., Knott, A., Liddicoat, J., & Noorman, M. E. (2021). *A citizen's guide to artificial intelligence.* Cambridge: The MIT Press.