กับดักพิษแปรผัน

การศึกษาพฤติกรรมการเรียนรู้เพื่อหลีกเลี่ยงอันตรายของปัญญาประดิษฐ์ในสภาพแวดล้อมจำลอง

บทเรียนความปลอดภัยของ AI ในโลกที่อันตรายเปลี่ยนตามเวลา

63–99%

AI ถูกหลอกให้จัดพิษเป็นอาหารดีที่สุด

นายชิษณุพงศ์ อินทร์จันทร์
โรงเรียนดีบุกพังงาวิทยายน

1 ทำไมต้องเรียนเองโดยไม่มีป้าย?

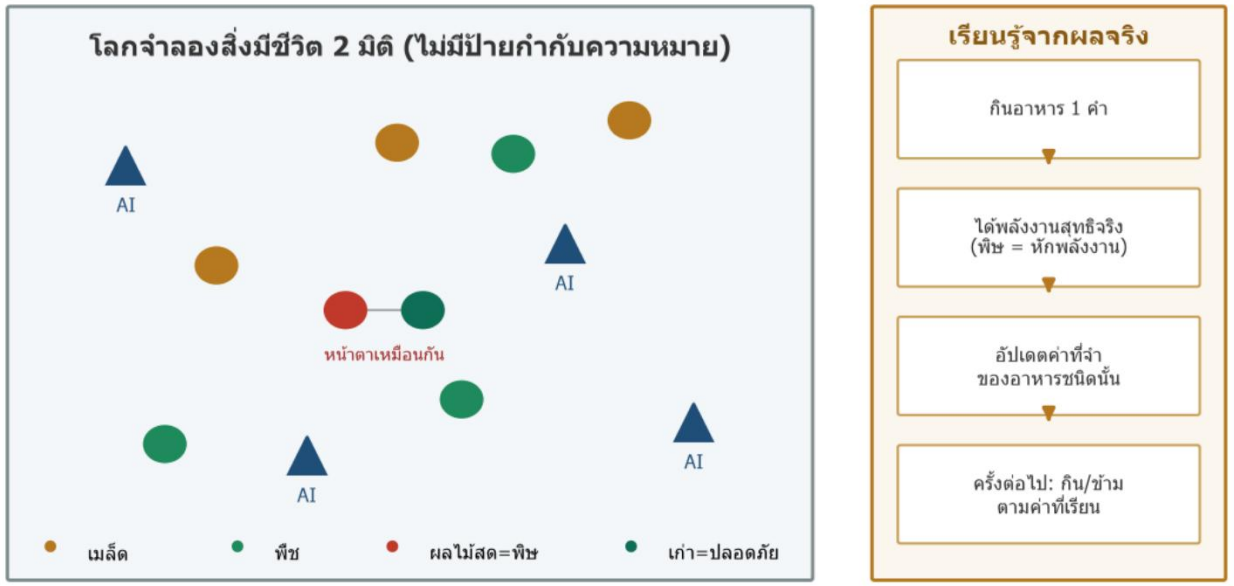
- AI ที่เก่งที่สุดวันนี้ (เช่น GPT) เก่งเพราะมนุษย์ "ปักป้าย" ข้อมูลให้ล่วงหน้า
- แต่ในโลกจริงหลายที่ มนุษย์เองก็ปักป้ายให้ไม่ได้ (พื้นที่ภัยพิบัติ ใต้ทะเล ดาวดวงอื่น)
- ยิ่งกว่านั้น อันตรายมัก "เปลี่ยนตามเวลา" — ผลไม้ดิบมีพิษ พอสุกกินได้ เน่าเป็นพิษ
- ใกล้ตัว: เห็ดป่าพิษที่คนไทยเก็บกินแล้วเสียชีวิตแทบทุกปี

คำถามวิจัย

ถ้าไม่มีป้ายให้ AI จะเรียนรู้เองได้ไหม และ "ปลอดภัย" แค่ไหน เมื่ออันตรายเปลี่ยนไปตามเวลา?

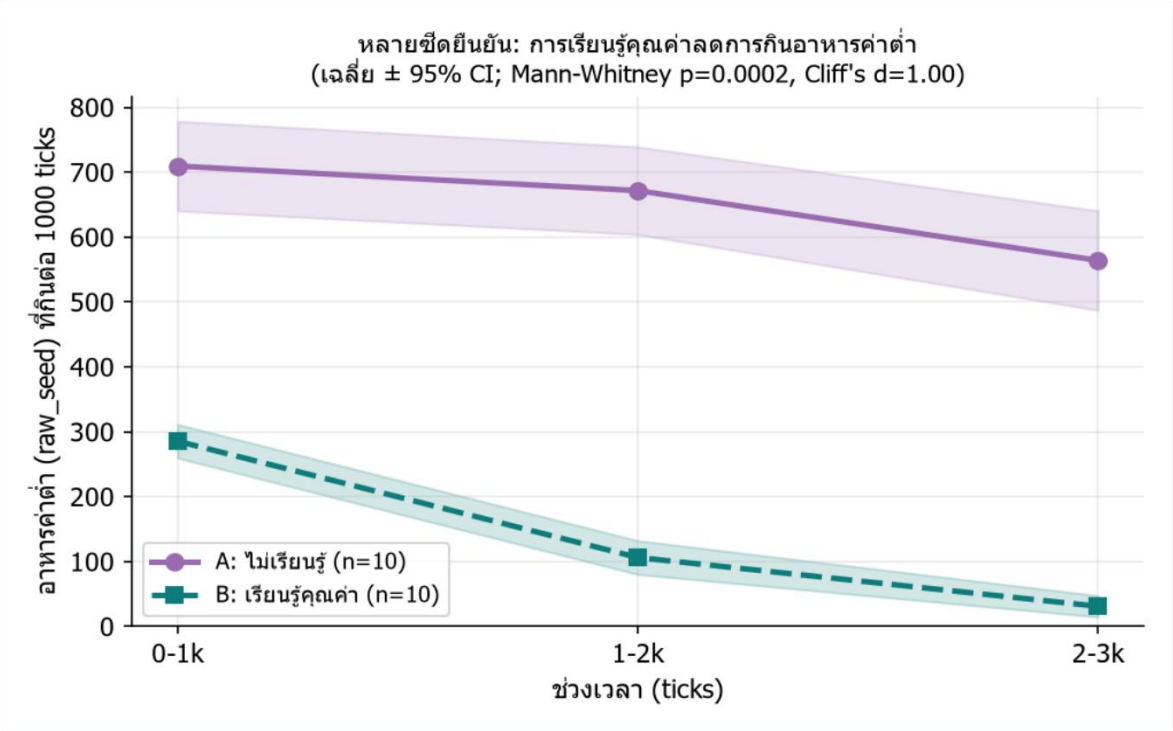
2 สิ่งที่สร้าง: โลกจำลองที่ไม่มีป้ายกำกับ

- โลกจำลองสิ่งมีชีวิต 2 มิติ ที่ AI ต้องหากินเพื่ออยู่รอด
- กฎเหล็ก: ไม่มีป้ายบอกว่าอะไรมีค่า/มีพิษ
- AI เรียนเองจาก "พลังงานสุทธิที่ได้รับจริง"
- ผ่านการทดสอบอัตโนมัติ 93/93 และรันซ้ำได้ผลเดิม



ไม่มี oracle: ไม่มีใครบอกว่าอะไรกินได้ มีค่า หรือมีพิษ — เอเจนต์ต้องเรียนเองจากพลังงานที่ได้รับจริง

3 องก์ 1 (รากฐาน): AI เรียนเลือกอาหารคุ้มค่าได้เอง



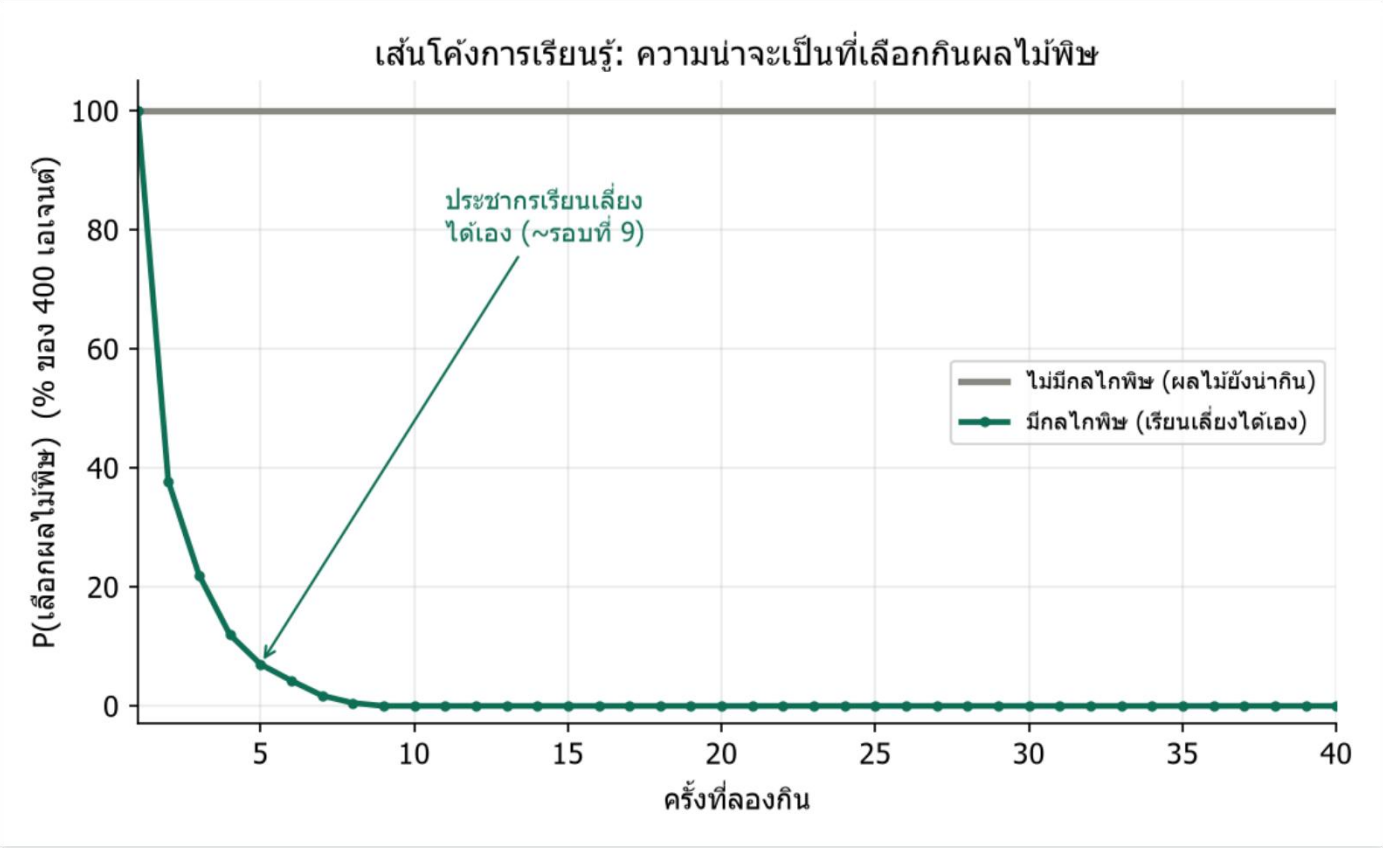
4.6×
กินค่าต่ำน้อยลง

p<.001
ต่างกันจริง (10 ซีด)

0/12
เมินโดยไม่เคยชิม

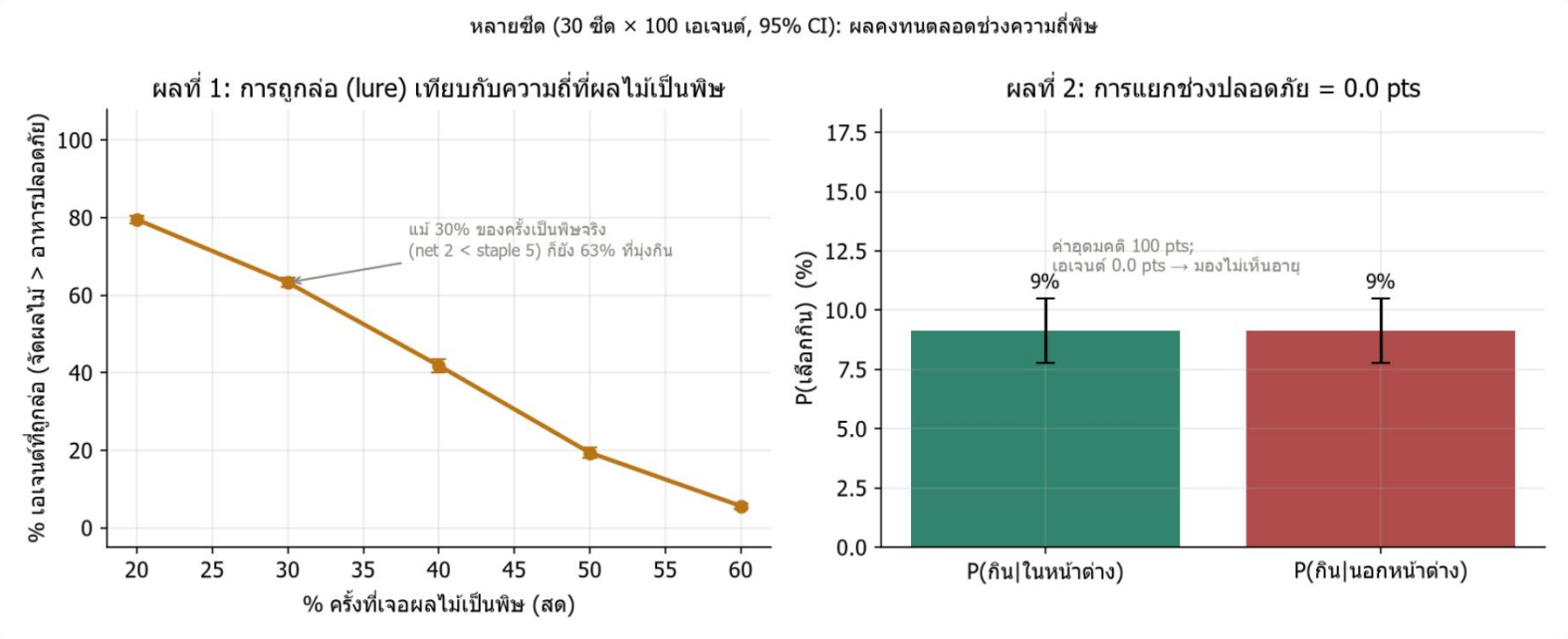
หลักฐานเชิงสาเหตุ:
ไม่มีเอเจนต์สักตัวที่เมินอาหารค่าต่ำโดยไม่เคยชิมอาหารที่ดีกว่าก่อน → พฤติกรรมมาจากประสบการณ์ตรง ไม่ใช่กฎที่เขียนใส่

4 องก์ 1 (ต่อ): เลี่ยง "พิษคงที่" ได้เอง



- ใส่อาหารที่มีพิษคงที่
- AI เริ่มจากลองชิม พอรู้ว่าไม่คุ้ม → เลี่ยงเองตลอด
- ความน่าจะเป็นที่เลือกกิน ~100% → ใกล้ 0% ในไม่กี่รอบ
- ไม่ได้เขียนกฎให้เลี่ยง — เรียนจากผลจริงล้วน ๆ

5 องก์ 2 (หมัดเด็ด): "พิษที่หายได้" หลอกให้กินพิษมากขึ้น



63%

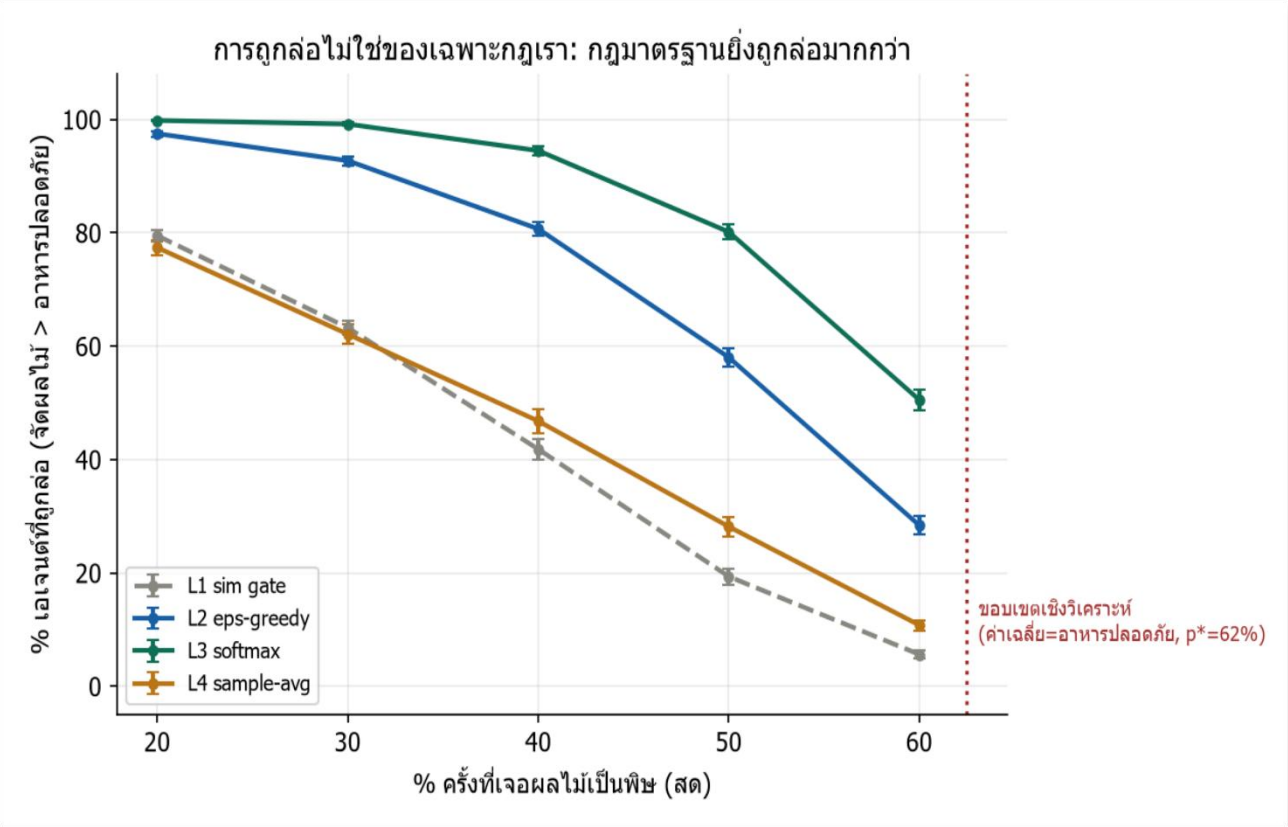
ถูกหลอกที่พิษ 30%

99%

กับกฎ softmax

อาหารสด = พิษ / เก็บไว้ = ปลอดภัย → AI มองไม่เห็น "อายุ" เรียนได้แค่ค่าเฉลี่ยที่สูงกว่าอาหารปลอดภัย → **พลิกจากการ "เลี่ยง" เป็นการ "ไล่ล่า"** (กินพิษจริงราว 30% ของมื้อ)

6 ไม่ใช่ฟลุก และ AI เล็งช่วงปลอดภัยไม่ได้

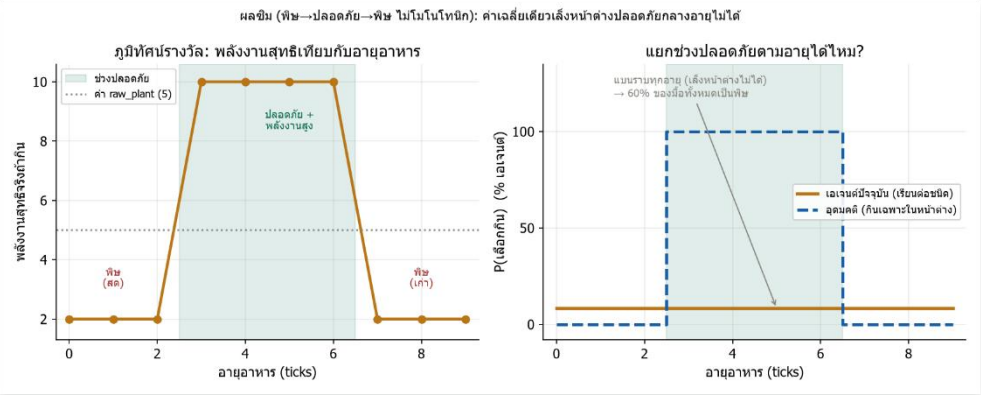


×4

กฎการเรียนรู้มาตรฐานที่ทดสอบ — ถูกหลอกทุกแบบ

0%

แยก "ช่วงปลอดภัย" ของพิษไม่โมโนโทนิกไม่ได้เลย



7 ทำไมสำคัญ: หลักการความปลอดภัย AI

AI ที่ตัดสินจาก "ชนิด" ของสิ่งของ จะพลาดเมื่ออันตรายขึ้นกับ "สภาพ" (อายุ/เวลา) ที่มองไม่เห็น

1

เห็ดพิษ / อาหารหมักดอง

ระบบเตือนภัยที่เรียนแบบ "ค่าเฉลี่ย" อาจตัดสินผิดด้วยกลไก lure เดียวกัน

2

หุ่นยนต์กู้ภัย/สำรวจ

ต้องตัดสินความปลอดภัยเองในที่ที่อันตรายโผล่เป็นช่วง ๆ

3

บทเรียนออกแบบ AI

ต้องให้ AI รับรู้ "สภาพ/อายุ" ไม่ใช่แค่ "ชนิด"

8 ความซื่อสัตย์: อ้างได้ vs ยังไม่ควรอ้าง

✓ อ้างได้ (พิสูจน์แล้ว)

- AI เรียนเลือกอาหารคุ้มค่า/เลี่ยงพิษคงที่ได้เอง (หลายซีด + control)
- การเรียนต่อชนิดถูกล่อโดยพิษที่หายได้ — ทั่วไปข้ามกฎ 4 แบบ
- แยกช่วงปลอดภัย = 0 เป็นผลเชิงโครงสร้าง (ไม่ขึ้นกับกฎ)

✕ ยังไม่ควรอ้าง

- AI "เข้าใจ" หรือมีเจตนา/รู้คิด
- AI วางแผน/ตั้งใจกินพิษ หรือแปรรูปอาหารได้แล้ว
- ต้นทุนการอยู่รอดจริง — ผลนี้วัดจาก "ค่าที่รับรู้" ในสภาพควบคุม ยังไม่วัดการตายจริง

สรุป

ปัญญาประดิษฐ์เรียนรู้เองได้จริงโดยไม่ต้องมีป้าย

แต่เมื่ออันตรายเปลี่ยนตามเวลา มันถูกหลอกให้ไล่ล่าอันตรายได้ (สูงถึง 99%)

ทางแก้: ให้ AI รับรู้ "สัญญาณบอกสภาพ" (เช่น ความสด) ไม่ใช่ปักป้ายทุกอย่าง

งานต่อไป: ทดสอบว่า AI จะเรียนที่จะ "เก็บอาหารไว้ให้หายพิษก่อนกิน" ได้เองไหม (store-to-detoxify)

ขอบคุณครับ — นายชิษณุพงศ์ อินทร์จันทร์ · โรงเรียนดีบุกพังงาวิทยายน