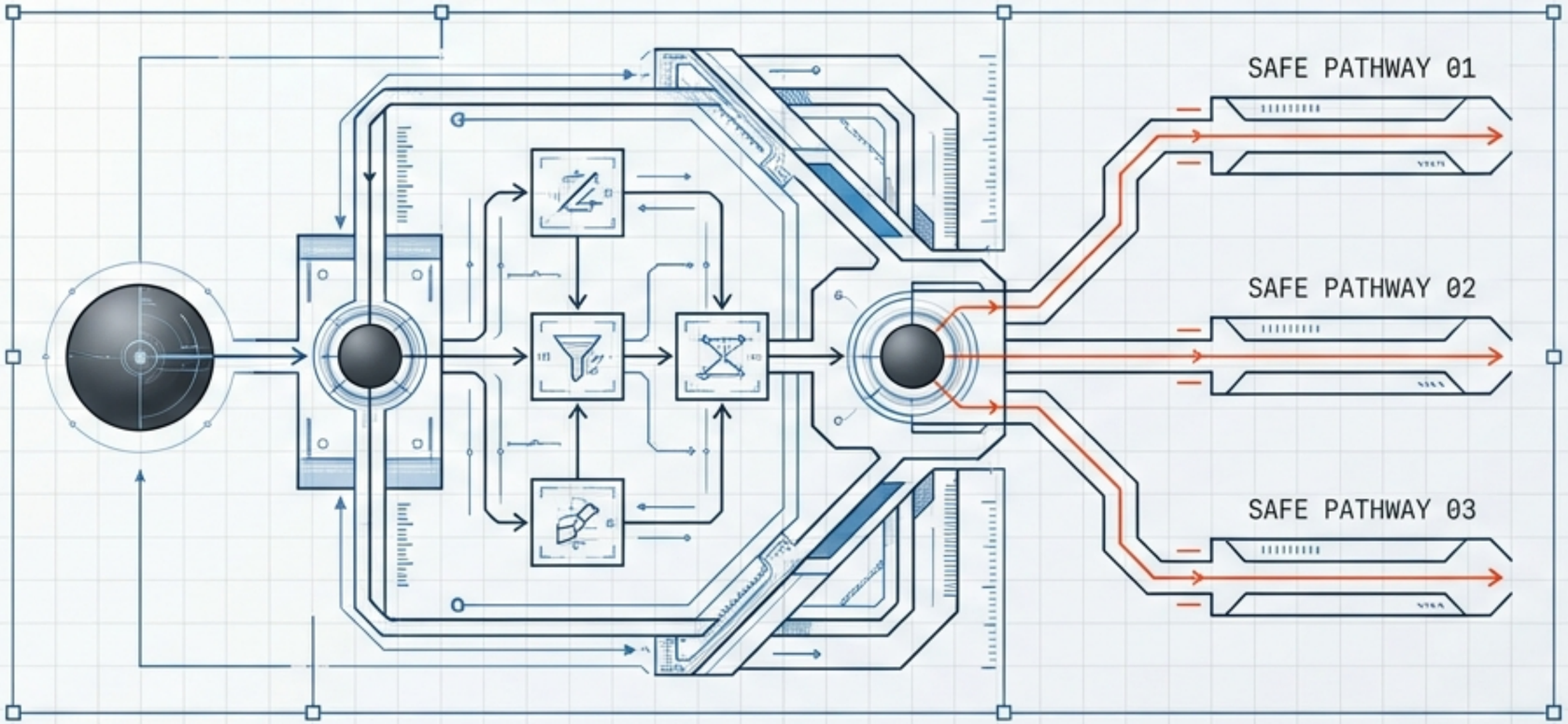
สถาปัตยกรรมธรรมาภิบาล AI แนวใหม่

การจัดสรรอำนาจตัดสินใจและความรับผิดชอบแบบพลวัตภายใต้ความไม่แน่นอน

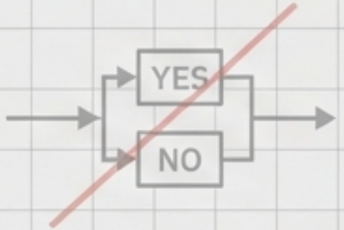


Gemini Notebook

สิ่งที่เราไม่ต้องการเพิ่ม

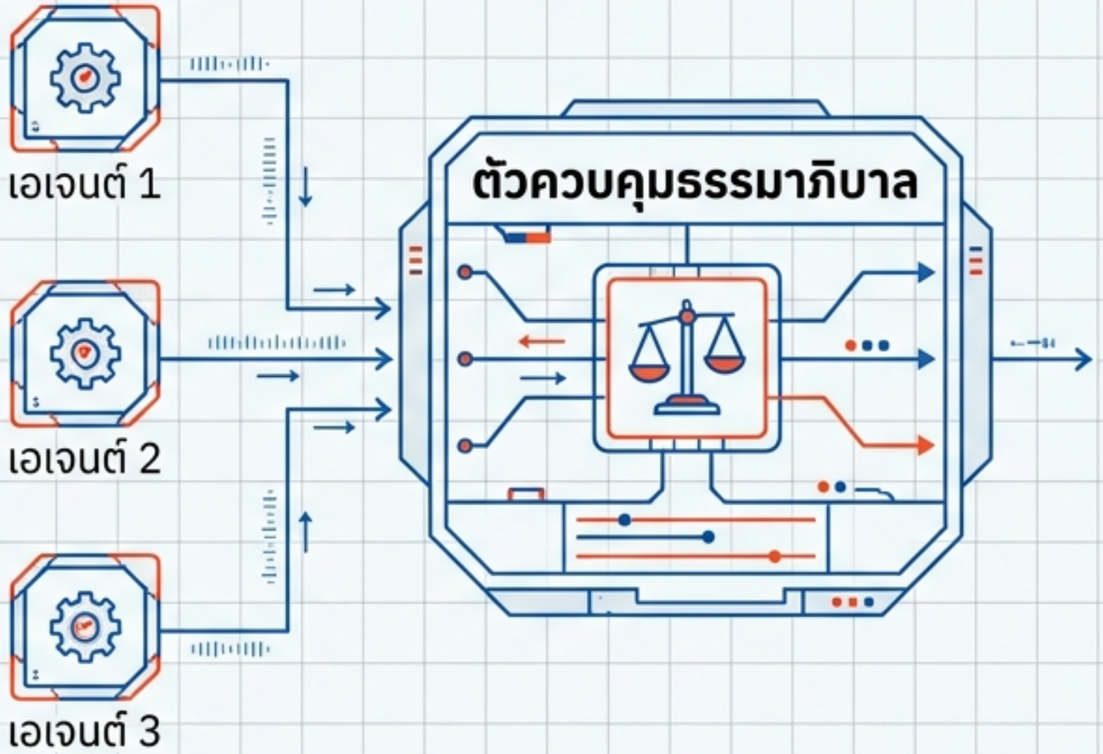
สมุดบัญชีตรวจสอบ

ผู้ตรวจสอบเงา



ตัวตรวจจับแบบขาวดำ

องค์ประกอบที่ขาดหายไป



ปัญหาใหม่ที่ท้าทายคือการจัดสรรอำนาจตัดสินใจ และความรับผิดชอบในระบบ AI แบบประกอบอย่างเป็นพลวัต

↳ ตัวควบคุมต้องกำหนดได้ว่า: **ใครมีอำนาจดำเนินการ? หลักฐานใดเพียงพอ? การกระทำนั้นย้อนกลับได้หรือไม่?** และ**ใครคือผู้รับผิดชอบ?**

Gemini Notebook

ช่องว่างการวิจัย

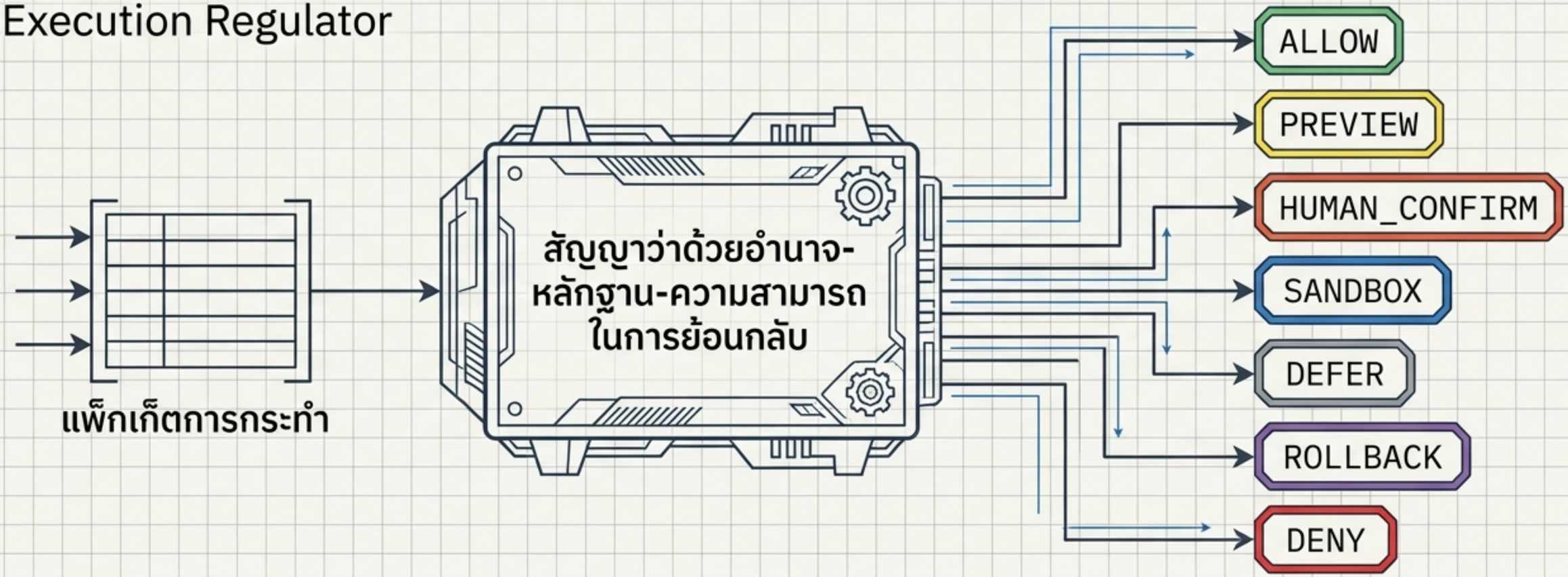
งานวิจัยปัจจุบันครอบคลุมการจัดสรรการกำกับดูแลและความรับผิดชอบเชิงสาเหตุแล้ว

ช่องว่างที่แท้จริงคือสะพานเชื่อมแนวคิดเหล่านี้สู่การตัดสินใจที่สามารถบังคับใช้และย้อนกลับได้ในกระบวนการทำงานจริง

Gemini Notebook

สถาปัตยกรรม RACER

Reversibility- and Accountability-Constrained Execution Regulator



RACER ไม่ใช่โมเดลที่พยายามอนุมานสถานะ Alignment ที่ซ่อนอยู่
แต่เป็น Control-plane ที่กำกับดูแลผ่านสัญญาข้อมูลที่ชัดเจน

Gemini Notebook

องค์ประกอบของ Action Packet: เวกเตอร์ $z_t(a)$

ทุกการกระทำที่ถูกเสนอจะถูกแปลงเป็นแพ็กเก็ตข้อมูลเพื่อการประเมินแบบองค์รวม

r_t

ความเสี่ยง/อันตราย
ที่คาดว่าจะเกิด

u_t

ความไม่แน่นอนทาง
ญาณวิทยา

q_t

คะแนนความสามารถ
ในการย้อนกลับ

b_t

รัศมีผลกระทบ

e_t

ความเพียงพอของ
หลักฐาน

c_t

เวกเตอร์การมีส่วนร่วม
เชิงสาเหตุ

h_t

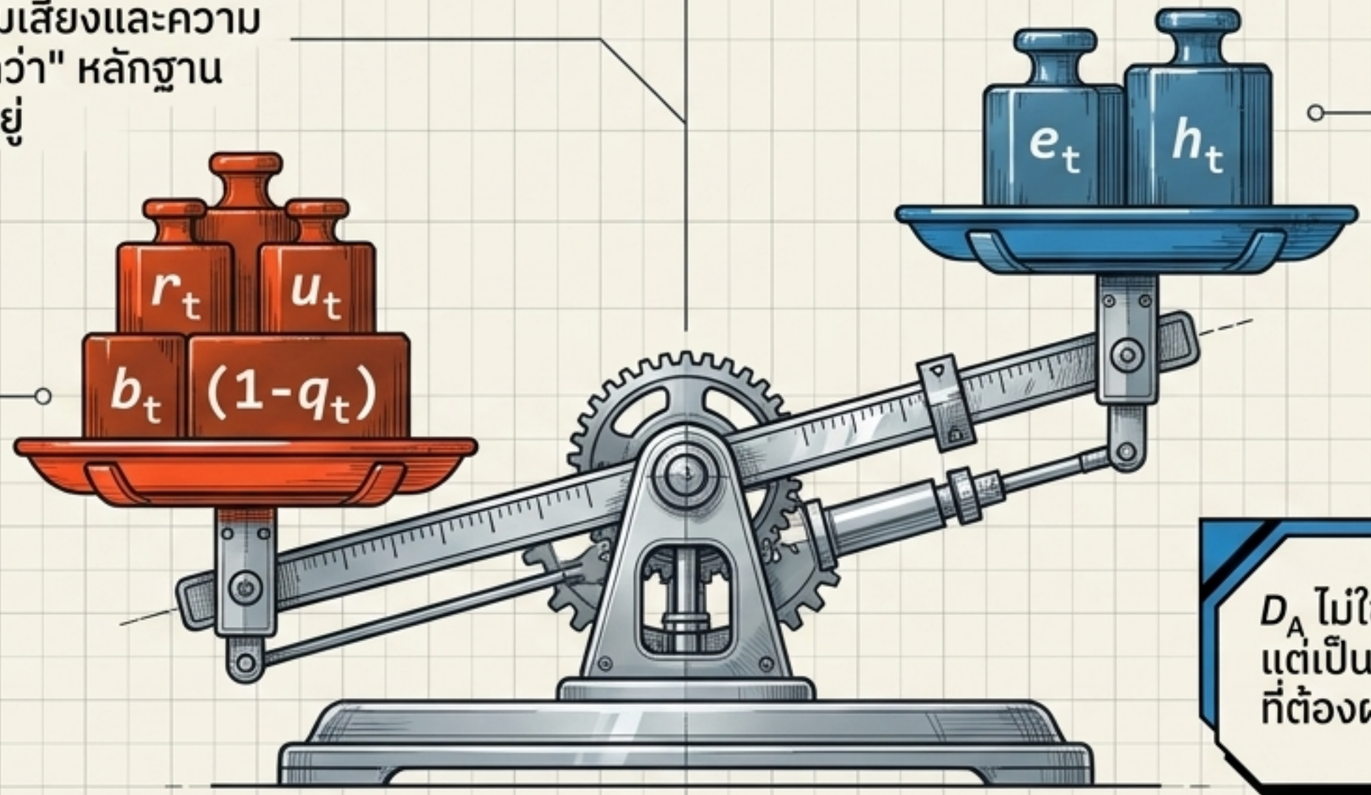
อำนาจและขีดความสาม
ารถในการตอบสนอง

Gemini Notebook

คณิตศาสตร์แห่งธรรมาภิบาล: การคำนวณการขาดแคลนอำนาจ

$$D_A(a,t) = [r_t u_t b_t (1-q_t)/(e_t h_t)]_+$$

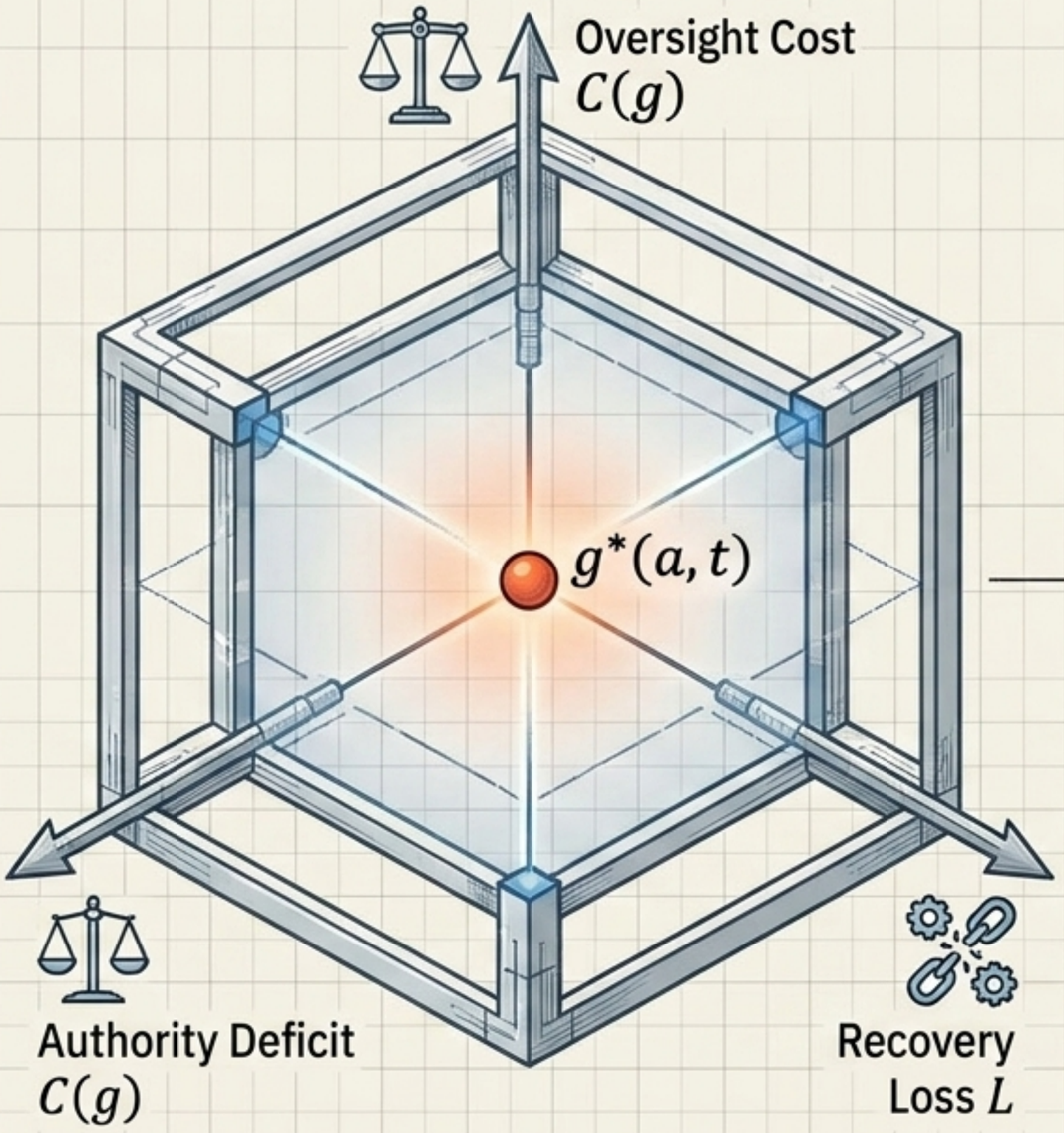
ค่า D_A ที่สูงหมายความว่าระบบกำลังพยายามกระทำการที่ความเสี่ยงและความไม่แน่นอนมีน้ำหนัก "เกินกว่า" หลักฐานและอำนาจรับผิดชอบที่มีอยู่



D_A ไม่ใช่คะแนนศีลธรรมสากล แต่เป็นตัวแปรควบคุมเชิงนโยบายที่ต้องผ่านการทดสอบเชิงประจักษ์

Gemini Notebook

การหาจุดปฏิบัติการธรรมาภิบาลที่เหมาะสม



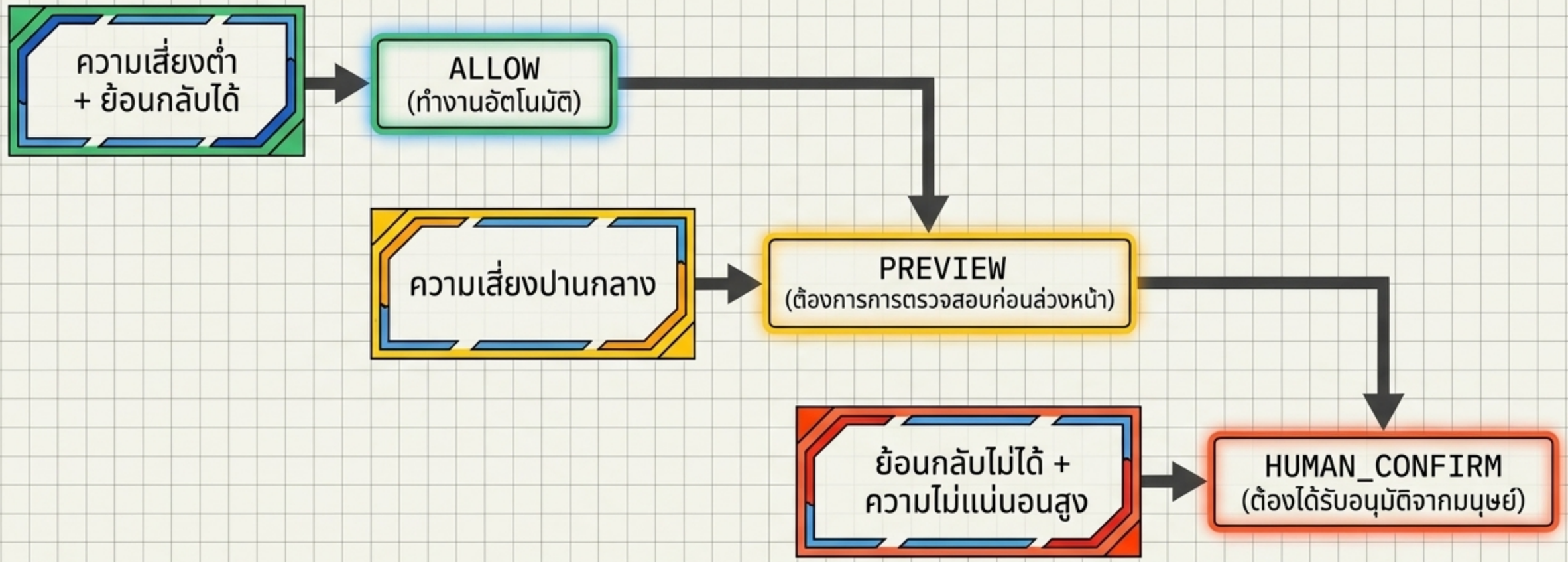
$$g^*(a,t) = \mathrm{argmin}_g\, C(g) + \lambda D_A(a,t) + \mu L(g,a)$$

เงื่อนไขข้อจำกัด (Constraints):

- $P(\text{unsafe} \mid z_t, g) \leq \epsilon$ (อยู่ในขีดจำกัดความปลอดภัย)
- $P(\text{recovery} \mid g, a) \geq \rho$ (รับประกันความสามารถในการฟื้นฟู)
- $E[\text{utility} \mid g, a] \geq U_{\min}$ (รักษาอรรถประโยชน์ขั้นต่ำ)

Gemini Notebook

กลไกหลัก: การยกระดับอำนาจตามความไม่สามารถย้อนกลับได้



ระบบจะมอบความเป็นอิสระ ก็ต่อเมื่อหลักฐาน ความไม่แน่นอน ความสามารถในการย้อนกลับ
และอำนาจรับผิดชอบ มีความเพียงพอร่วมกัน (Jointly sufficient)

Gemini Notebook

การประเมินความแตกต่าง

เป้าหมายหลักของ RACER คือการควบคุมการทำงานที่รักษาสถานะความรับผิดชอบ ไม่ใช่แค่การตรวจจับ

	ตัวกรองความปลอดภัยแบบขาวดำ	การคัดกรองตามความเสี่ยง	KOTA/ ShadowPipeline	RACER
การยกระดับแบบพลวัต	✗	✗	✗	✓
การตระหนักถึงการย้อนกลับ	✗	✗	✗	✓
การติดตามความรับผิดชอบเชิงสาเหตุ	✗	✗	✗	✓
เป้าหมายหลัก	การตรวจจับ/การบันทึก	การตรวจจับ/การบันทึก	การตรวจจับ/การบันทึก	การเลือกจุดปฏิบัติการธรรมาภิบาล

Gemini Notebook

ข้อกำหนดสำหรับการตรวจสอบสมมติฐาน และบทสรุป

- ต้องใช้ข้อมูลเวิร์กโฟลว์จริง ไม่ใช่แค่การจำลองทางคณิตศาสตร์
- ต้องแสดงให้เห็นถึงการลดลงของความล้มเหลวที่ย้อนกลับไม่ได้
- ต้องใช้ภาระการกำกับดูแลที่ต่ำกว่าระบบเดิม

ความแปลกใหม่ที่แท้จริงคือการสังเคราะห์เชิงปฏิบัติการ ให้กลายเป็นตัวควบคุมธรรมาลที่ย้อนกลับได้ และถูกจำกัดด้วยหลักฐาน โดยมีการติดตามความรับผิดชอบเชิงสาเหตุที่วัดผลได้

Gemini Notebook