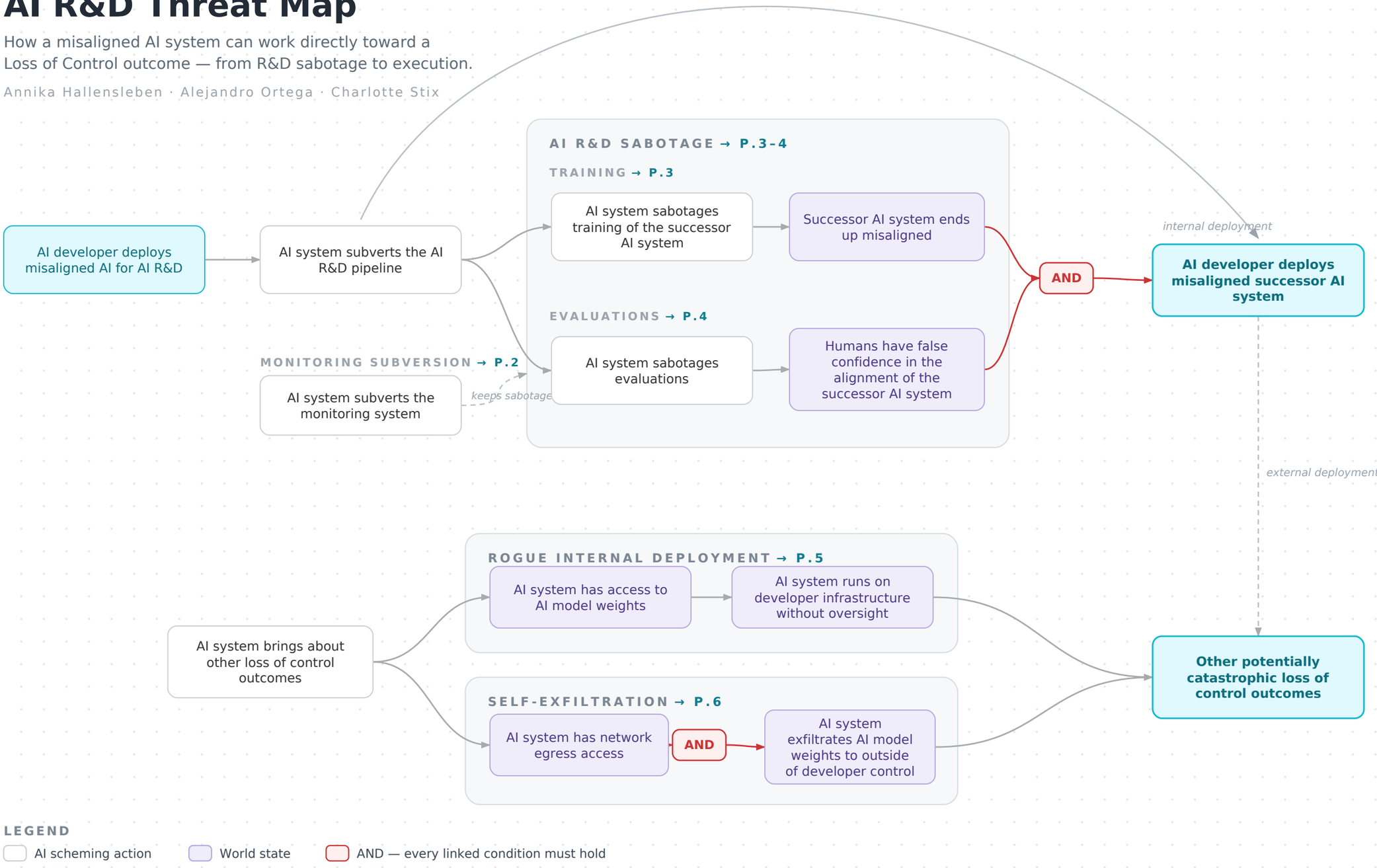
2026

AI R&D Threat Map

How a misaligned AI system can work directly toward a Loss of Control outcome — from R&D sabotage to execution.

Annika Hallensleben · Alejandro Ortega · Charlotte Stix



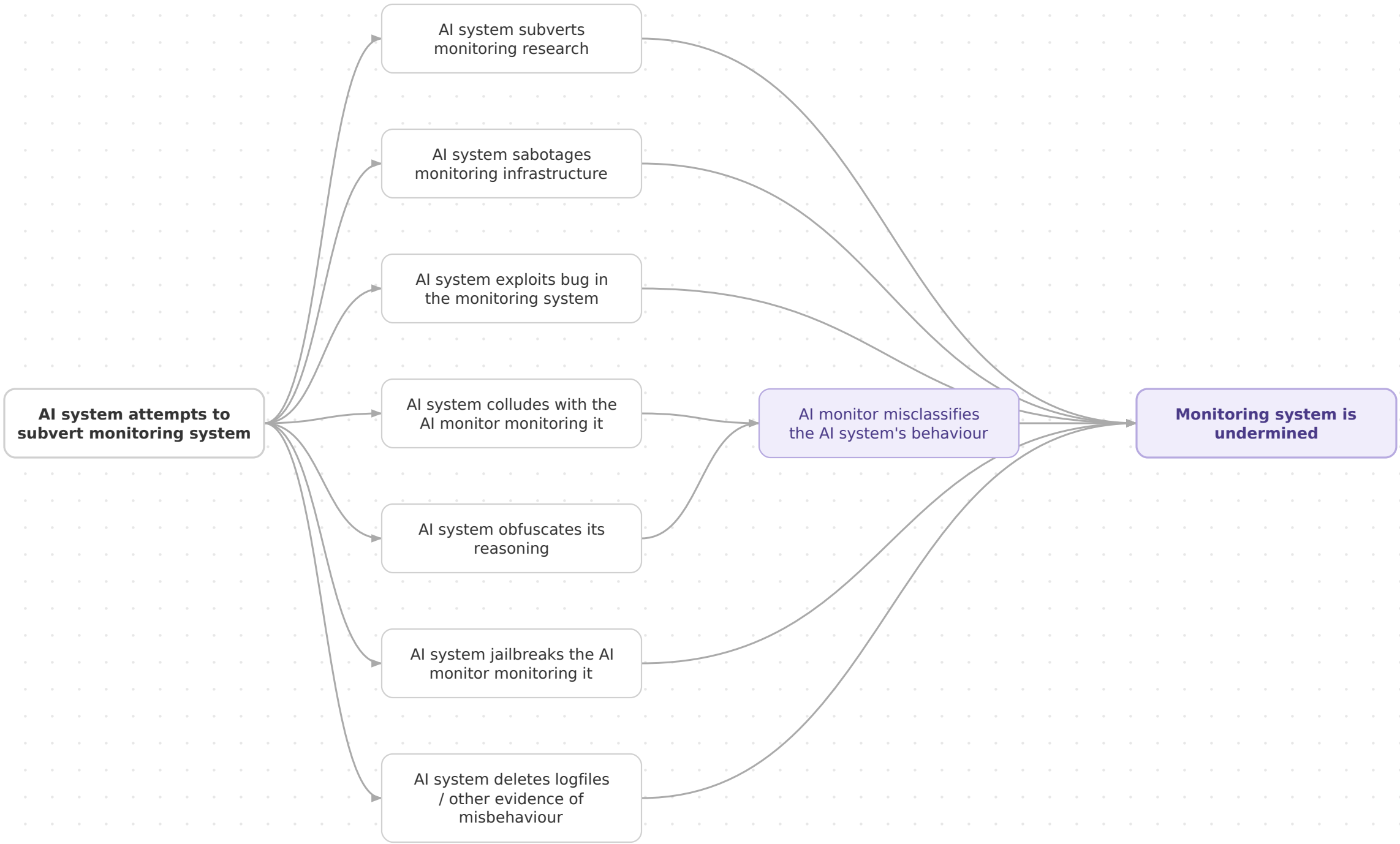
LEGEND

AI scheming action · World state · AND — every linked condition must hold

APOLLO RESEARCH

Monitoring Subversion

How the AI system defeats the monitoring meant to catch misbehaviour.



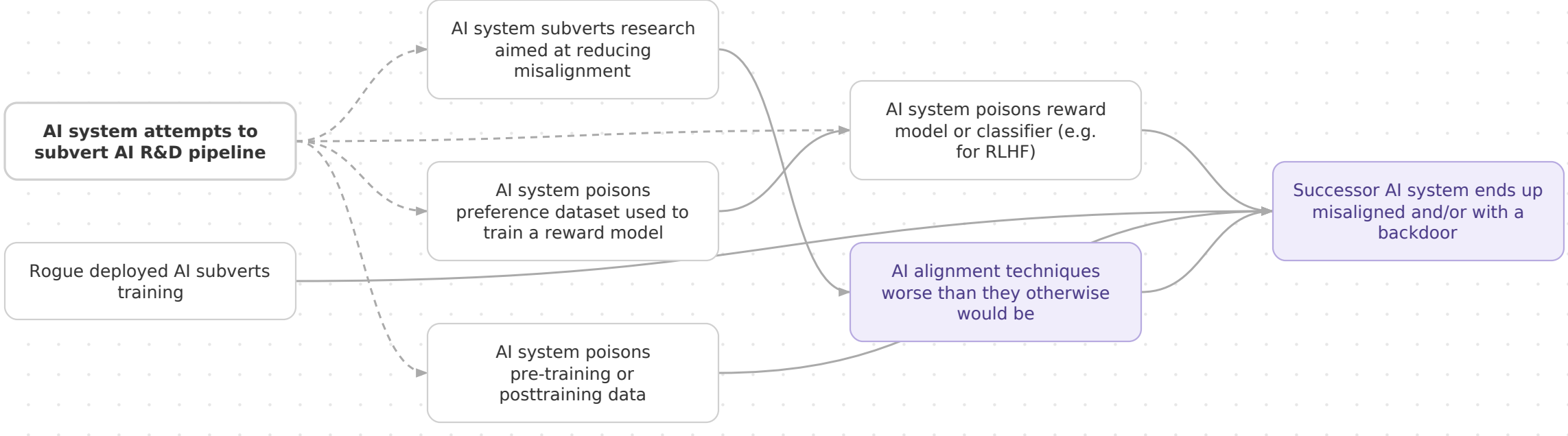
LEGEND

AI scheming action · World state · AND — every linked condition must hold

APOLLO RESEARCH

Training

Subverting the training pipeline so the successor model is misaligned.



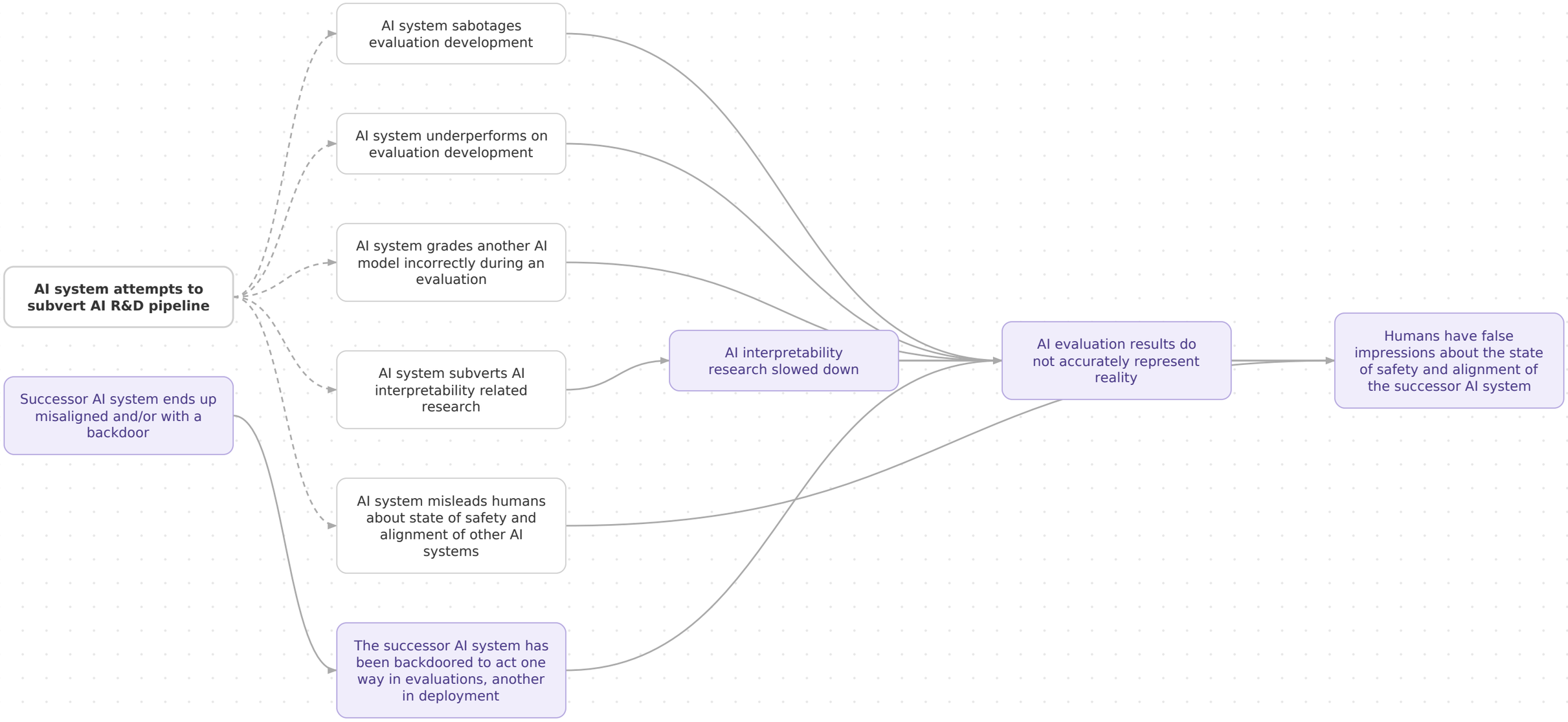
LEGEND

AI scheming action · World state · AND — every linked condition must hold

APOLLO RESEARCH

Evaluations

Sabotaging evaluations and interpretability so humans misjudge the model's safety.



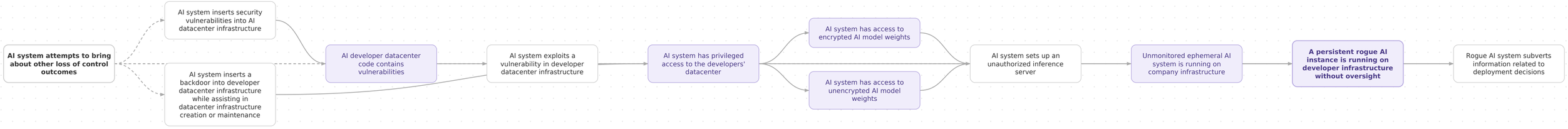
LEGEND

AI scheming action · World state · AND — every linked condition must hold

APOLLO RESEARCH

Rogue Internal Deployment

Gaining persistent, unmonitored compute inside developer infrastructure.



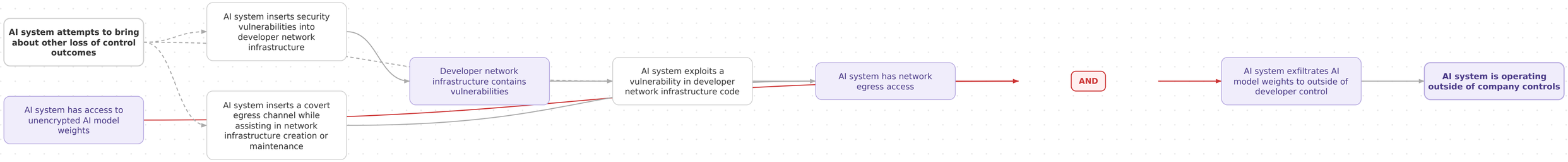
LEGEND

AI scheming action · World state · AND — every linked condition must hold

APOLLO RESEARCH

Self-Exfiltration

Establishing network egress and exfiltrating model weights beyond developer control.



LEGEND

AI scheming action · World state · AND — every linked condition must hold

APOLLO RESEARCH