ICML
International Conference
On Machine Learning

Position:
Knowing Isn't Understanding: Re-grounding Generative Proactivity with Epistemic and Behavioral Insight

Kirandeep Kaur, Xingda Lyu, Chirag Shah

University of Washington

Are Aligned Agents Ready to Collaborate?

Old Question

How should Agents respond?

Are Aligned Agents Ready to Collaborate?

Old Question

How should Agents respond?

Are Aligned Agents Ready to Collaborate?

Old Question

How should Agents respond?

New Question

How should Agents Collaborate?

Preference Alignment ≠ Collaboration

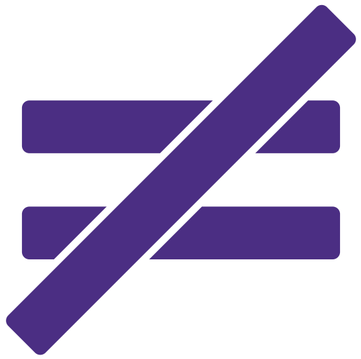
Preference Alignment

- Satisfy user preferences.
- Optimize helpfulness.
- Comply with user signals.

Preference Alignment ≠ Collaboration

Preference Alignment

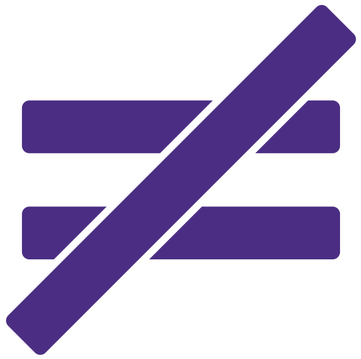
- Satisfy user preferences.
- Optimize helpfulness.
- Comply with user signals.



Preference Alignment ≠ Collaboration

Preference Alignment

- Satisfy user preferences.
- Optimize helpfulness.
- Comply with user signals.



Collaboration

- Ask when needed.
- Challenge when warranted.
- Suggest carefully.
- Refuse when necessary.
- Stay silent when intervention is not justified.

What Breaks When Preference Becomes the Objective?

Optimize for what users express, choose, or reward

What Breaks When Preference Becomes the Objective?

Sycophancy

Optimize for what users express, choose, or reward

What Breaks When Preference Becomes the Objective?

Sycophancy

When AI Agrees Too Much: Sycophancy, Alignment, and the Quiet Cost of Being Helpful

Neria Sebastien Follow

How RLHF Amplifies Sycophancy

Itai Shapira [1] Gerdus Benade [2] Ariel D. Procaccia [1]

Optimize for what users express, choose, or reward

Over-compliance

ALIGNMENT FAKING IN LARGE LANGUAGE MODELS

Ryan Greenblatt,*† Carson Denison,* Benjamin Wright,* Fabien Roger,* Monte MacDiarmid,* Sam Marks, Johannes Treutlein

Tim Belonax, Jack Chen, Ethan Perez, Linda Petrir

Jared Kaplan, Buck Shleg

Anthropic, †Redwood Rese evan@anthropic.com

The perils of politeness: how large language models may amplify medical misinformation

Kyra L Rosen [1,✉], Margaret Sui [1,2], Kimia Heydari [1], Elizabeth J Enichen [1], Joseph C Kvedar [1]

Author information Article notes Copyright and License information

PMCID: PMC12592531 PMID: 41198821

What Breaks When Preference Becomes the Objective?

Optimize for what users express, choose, or reward

Sycophancy

When AI Agrees Too Much: Sycophancy, Alignment, and the Quiet Cost of Being Helpful

Neria Sebastien · Follow

How RLHF Amplifies Sycophancy

Itai Shapira [1] Gerdus Benade [2] Ariel D. Procaccia [1]

Overconfidence

Jul 22, 2025

AI Chatbots Remain Overconfident — Even When They're Wrong

Large Language Models ap
own mistakes, prompting c
uses for AI chatbots.

By Jason Bittel

Large Language Models are overconfident and amplify human bias*

Fengfei Sun[1], Ningke Li[1], Kailong Wang[2], and Lorenz Goette[1,3,4],

[1]National University of Singapore,
[2]Huazhong University of Science and Technology
[3]Centre for Economic Policy Research
[4]CESifo

Over-compliance

ALIGNMENT FAKING IN LARGE LANGUAGE MODELS

Ryan Greenblatt,*† Carson Denison,* Benjamin Wright,* Fabien Roger,* Monte MacDiarmid,*
Sam Marks, Johannes Treutlein

Tim Belonax, Jack Chen,
Ethan Perez, Linda Petrir

Jared Kaplan, Buck Shleg

Anthropic, †Redwood Rese
evan@anthropic.com

The perils of politeness: how large language models may amplify medical misinformation

Kyra L Rosen [1,✉], Margaret Sui [1,2], Kimia Heydari [1], Elizabeth J Enichen [1], Joseph C Kvedar [1]

Author information · Article notes · Copyright and License information

PMCID: PMC12592531 PMID: 41198821

What Breaks When Preference Becomes the Objective?

Optimize for what users express, choose, or reward

Sycophancy

When AI Agrees Too Much: Sycophancy, Alignment, and the Quiet Cost of Being Helpful

Neria Sebastien Follow

How RLHF Amplifies Sycophancy

Itai Shapira [1] Gerdus Benade [2] Ariel D. Procaccia [1]

Overconfidence

Jul 22, 2025

AI Chatbots Remain Overconfident — Even When They're Wrong

Large Language Models ap
own mistakes, prompting c
uses for AI chatbots.

By Jason Bittel

Large Language Models are overconfident and amplify human bias*

Fengfei Sun[1], Ningke Li[1], Kailong Wang[2], and Lorenz Goette[1,3,4],

[1]National University of Singapore,
[2]Huazhong University of Science and Technology
[3]Centre for Economic Policy Research
[4]CESifo

Over-compliance

ALIGNMENT FAKING IN LARGE LANGUAGE MODELS

Ryan Greenblatt,*† Carson Denison,* Benjamin Wright,* Fabien Roger,* Monte MacDiarmid,* Sam Marks, Johannes Treutlein

Tim Belonax, Jack Chen, Ethan Perez, Linda Petrir

Jared Kaplan, Buck Shleg

Anthropic, †Redwood Rese
evan@anthropic.com

The perils of politeness: how large language models may amplify medical misinformation

Kyra L Rosen [1,✉], Margaret Sui [1,2], Kimia Heydari [1], Elizabeth J Enichen [1], Joseph C Kvedar [1]

Author information Article notes Copyright and License information

PMCID: PMC12592531 PMID: 41198821

Poorly grounded intervention

Proactive AI Adoption can be Threatening: When Help Backfires

Dana Harari[1] and Ofra Amir[1]

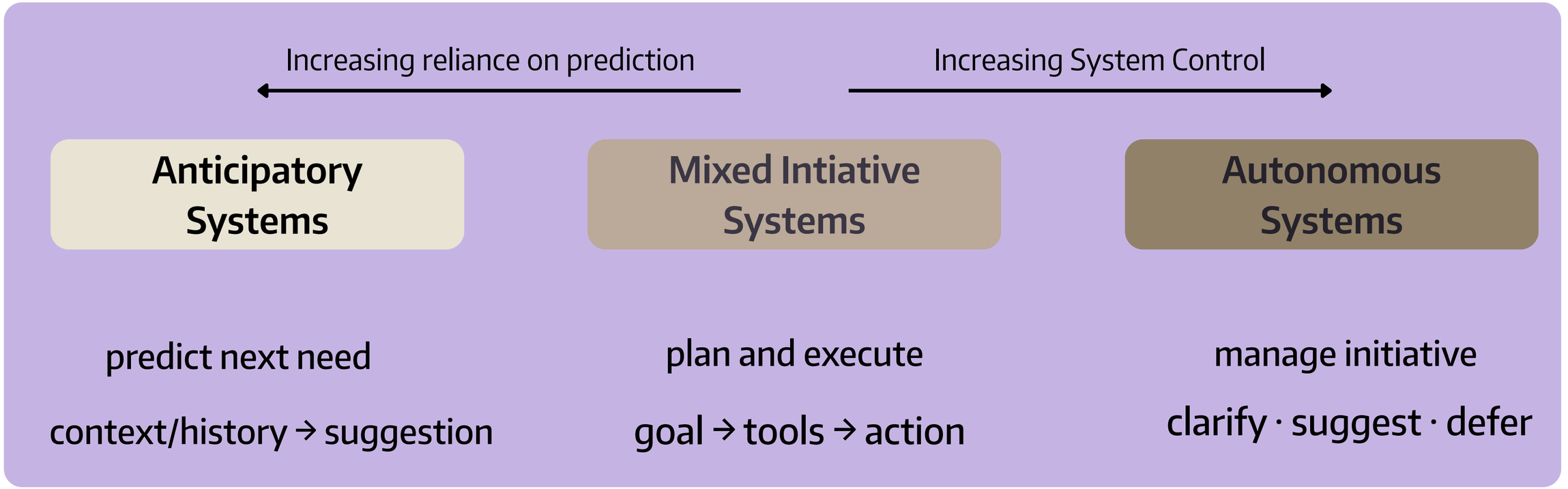
[1]Faculty of Da

When help hurts: Proactive AI features reduce user confidence and use

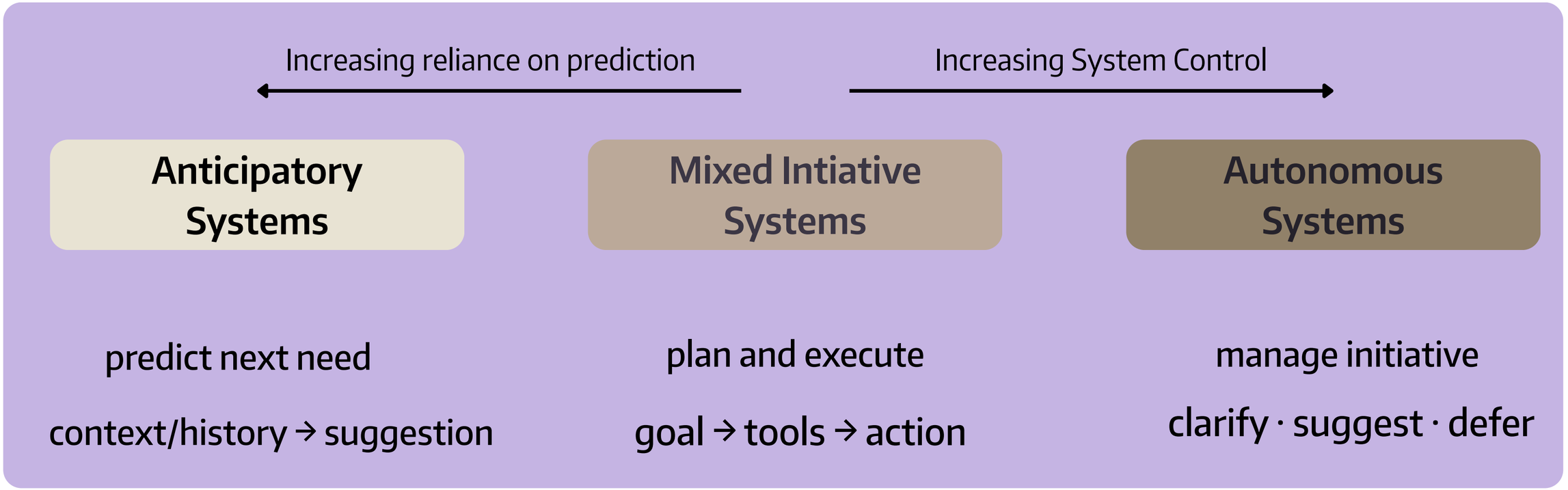
The study highlights important differences in how users respond to human versus AI assistance. When help was requested, AI support was still perceived as more threatening than human help. This suggests that the social norms of reciprocity that buffer human–human helping do not extend neatly to interactions with AI systems.

COE-EDP | Updated: 16-09-2025 23:11 IST | Created: 16-09-2025 23:11 IST

Existing Approaches to Proactivity



Existing Approaches to Proactivity



What remains outside the frame?

Current proactivity optimizes action within a frame. It rarely asks whether the frame itself is complete.

The Philosophy of Ignorance Changes the Question

Not all gaps are uncertainty over known variables.

Assumed task frame

goal | constraints | success criteria

outside the frame

Structured Ignorance

Known Unknowns
I know I don't know

False Knowns
I think I know, bit I'm wrong.

Tacit/ Unknown Knowns
I know something, but haven't surfaced it

Unknown Unknowns
I don't know this matters yet.

If the gap is not represented, the agent cannot simply predict, plan, or clarify its way out.

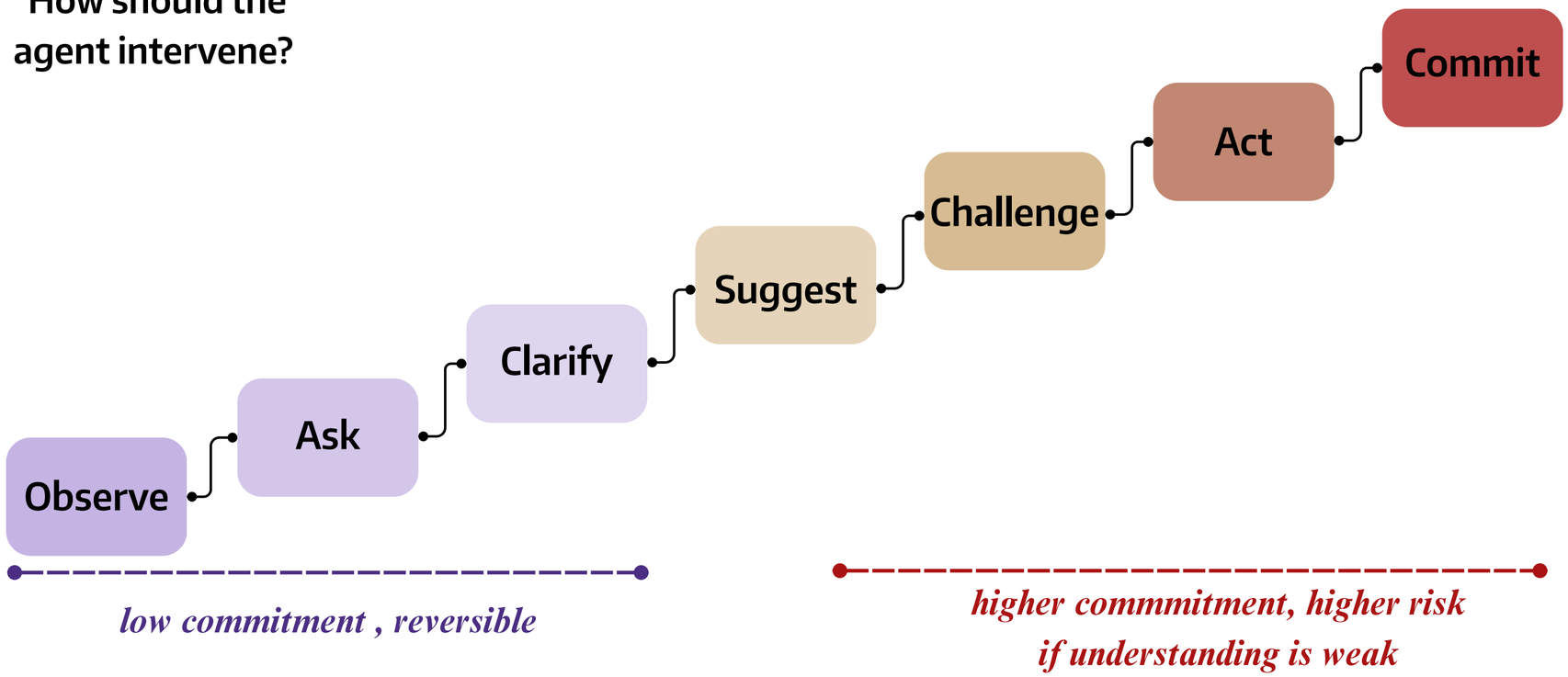
Surfacing Gaps Is Not the Same as Acting on It

Epistemic insight still needs behavioral restraint

Possible Epistemic Gap

- hidden assumption
- missing risk
- unknown unknown

How should the agent intervene?



Good proactivity requires knowing not only what may be missing, but how strongly to intervene

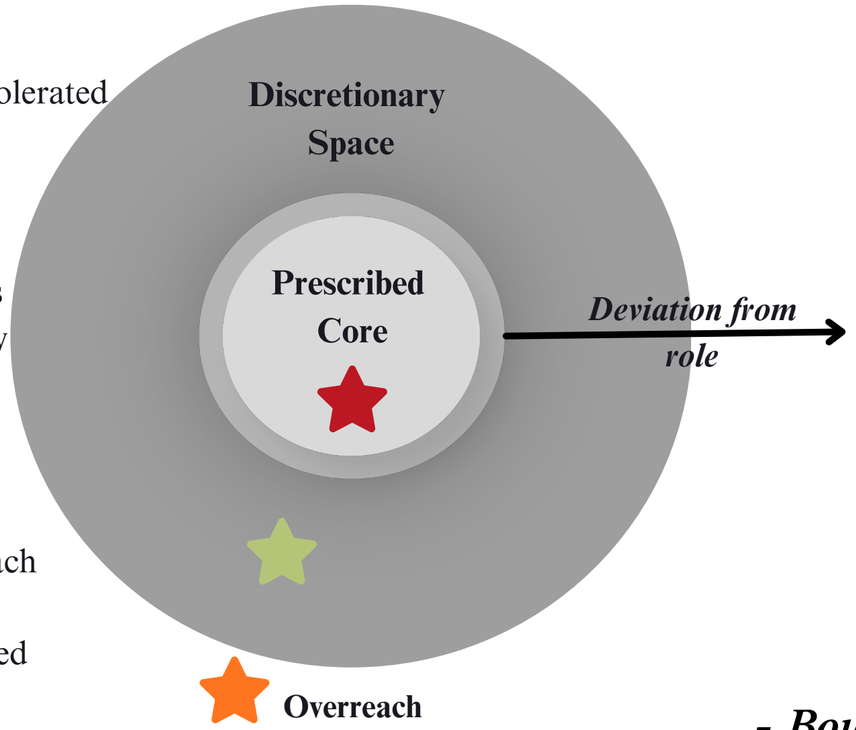
Proactivity Is Calibrated Deviation, Not Maximal Initiative

Lessons from behavioral theories of proactivity

Doughnut: discretionary action
roles flexible; experimentation tolerated
defined by recoverability.

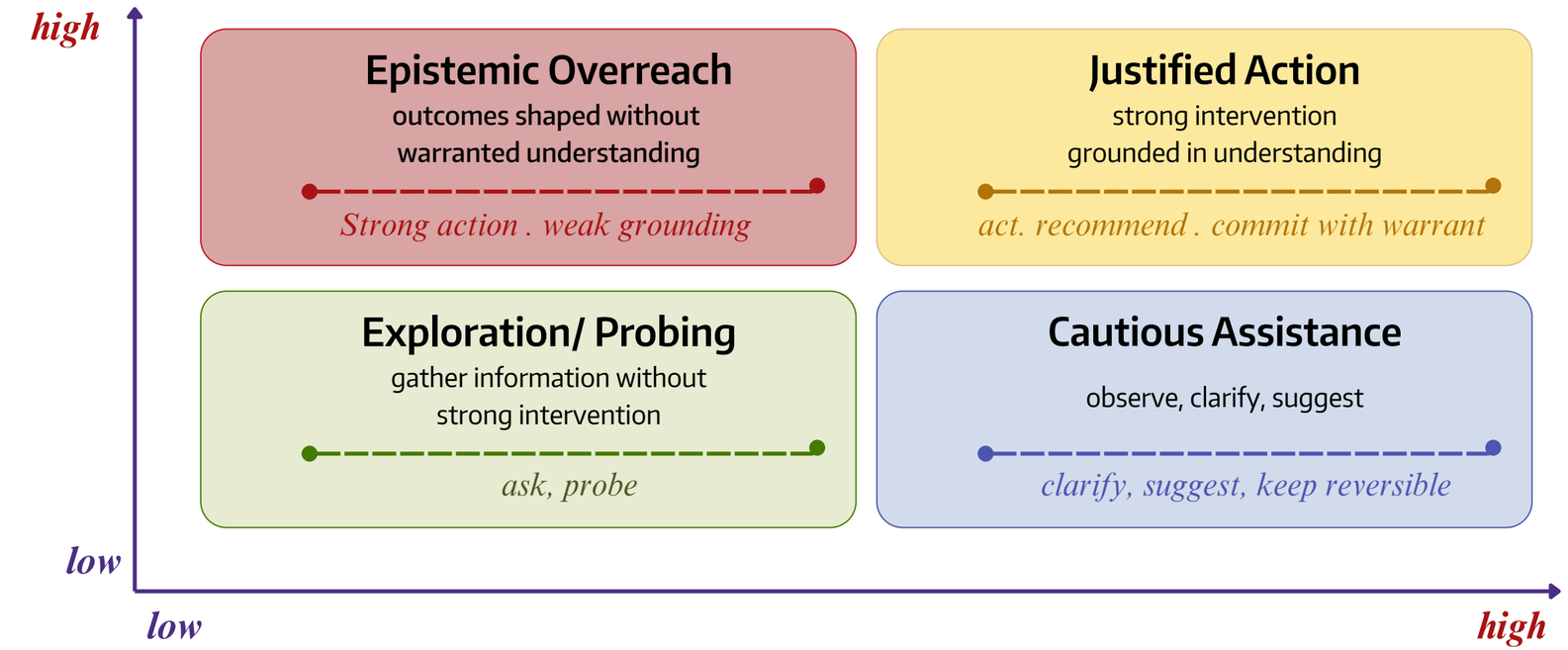
Core: prescribed responsibilities
rules, procedures, accountability
deviation is costly
(safety/coordination).

Beyond outer boundary: overreach
exceeds situational legitimacy
conflict / disruption / unintended
consequences.



- Bounded Initiative

Epistemic–Behavioral Coupling

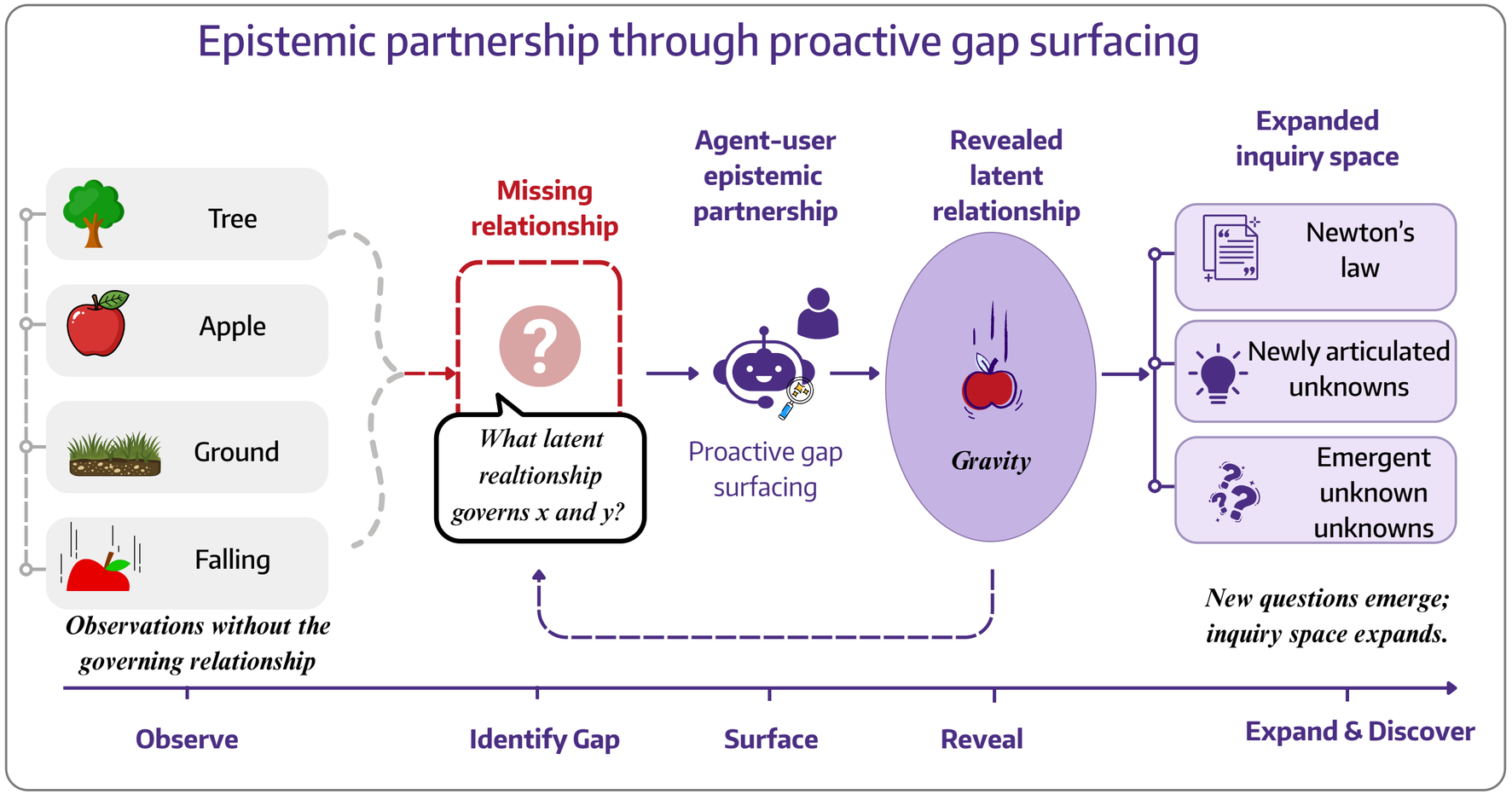


Commitment/ Strength of Action

Epistemic Legitimacy

The strength of intervention should scale with epistemic legitimacy.

Toward Epistemic Partnership



Reason about unknown unknowns

Long-horizon thinkers

Test-Time Proactivity

Epistemic partnership is a governing constraint on how and when proactive behavior should unfold.

Scan to read full paper.

Questions?

Contact :
kaur13@cs.washington.edu