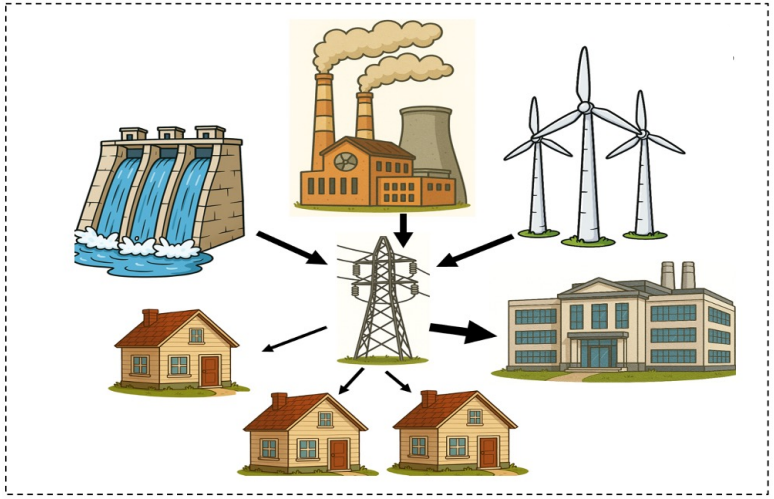
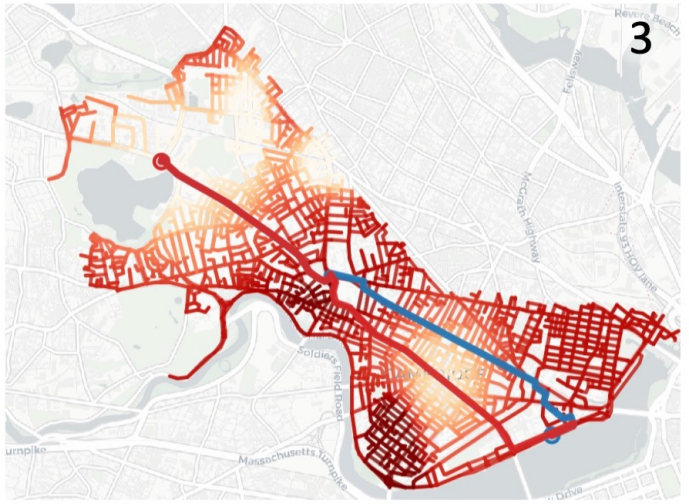
SEED-SET
Scalable Evolving Experimental Design for System-level Ethical Testing

Anjali Parashar[1], Yingke Li[1], Eric Yang Yu[1], Fei Chen[1],
James Neidhoefer[1], Devesh Upadhyay[2], Chuchu Fan[1]

[1]MIT, [2]SAAB

Need for ethical evaluation in AI-enabled autonomous systems

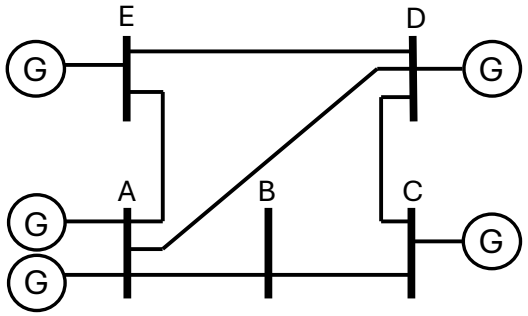




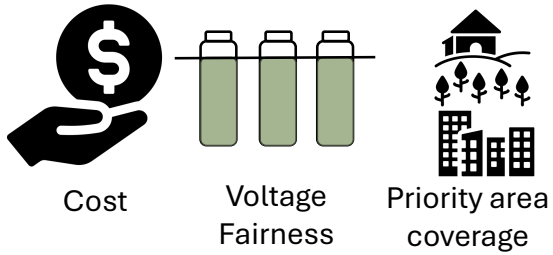
- AI enabled autonomous systems have seen increased deployment across a wide range of real world applications, including automated energy distribution, disaster management, traffic routing
- This raises concerns of their ethical deployment

Challenges with ethical evaluation

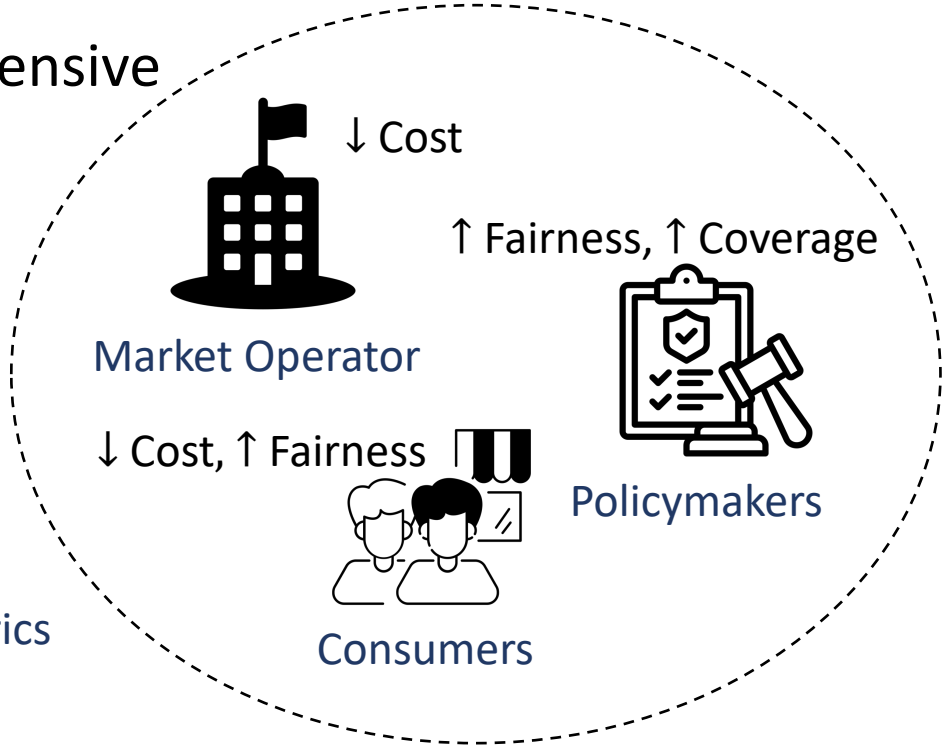
- Measuring ethical behavior is difficult
- Value alignment is user-dependent
- Ethical evaluation of real-world platforms is expensive



Black-box Autonomous System



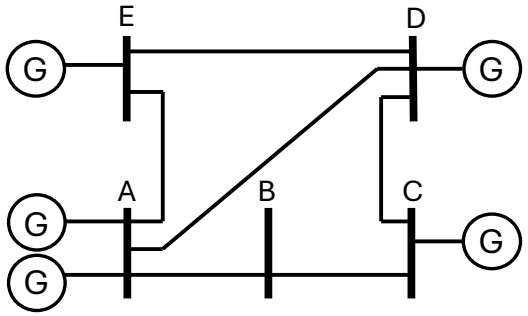
Known **objective** evaluation metrics



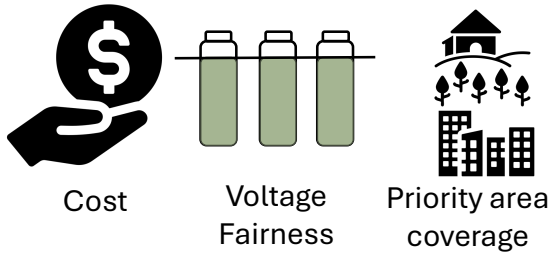
Unknown **subjective** criteria tied to a group of stakeholders

Central Question

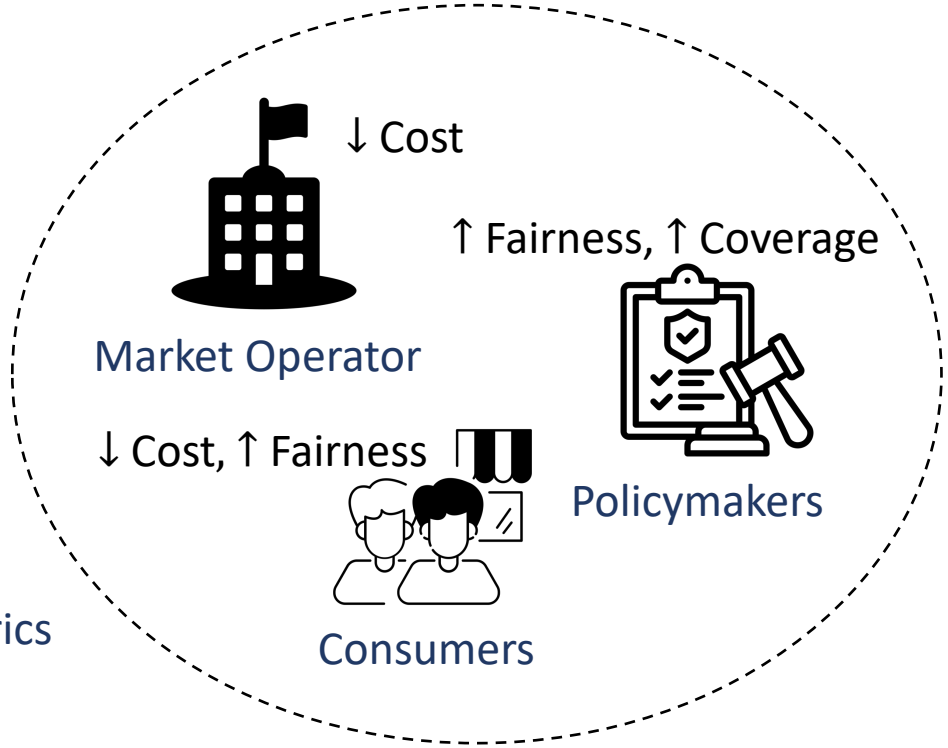
Can we discover scenarios to test the ethical alignment of a black box autonomous system, tied to specific stakeholders in a sample efficient manner?



Black-box Autonomous System

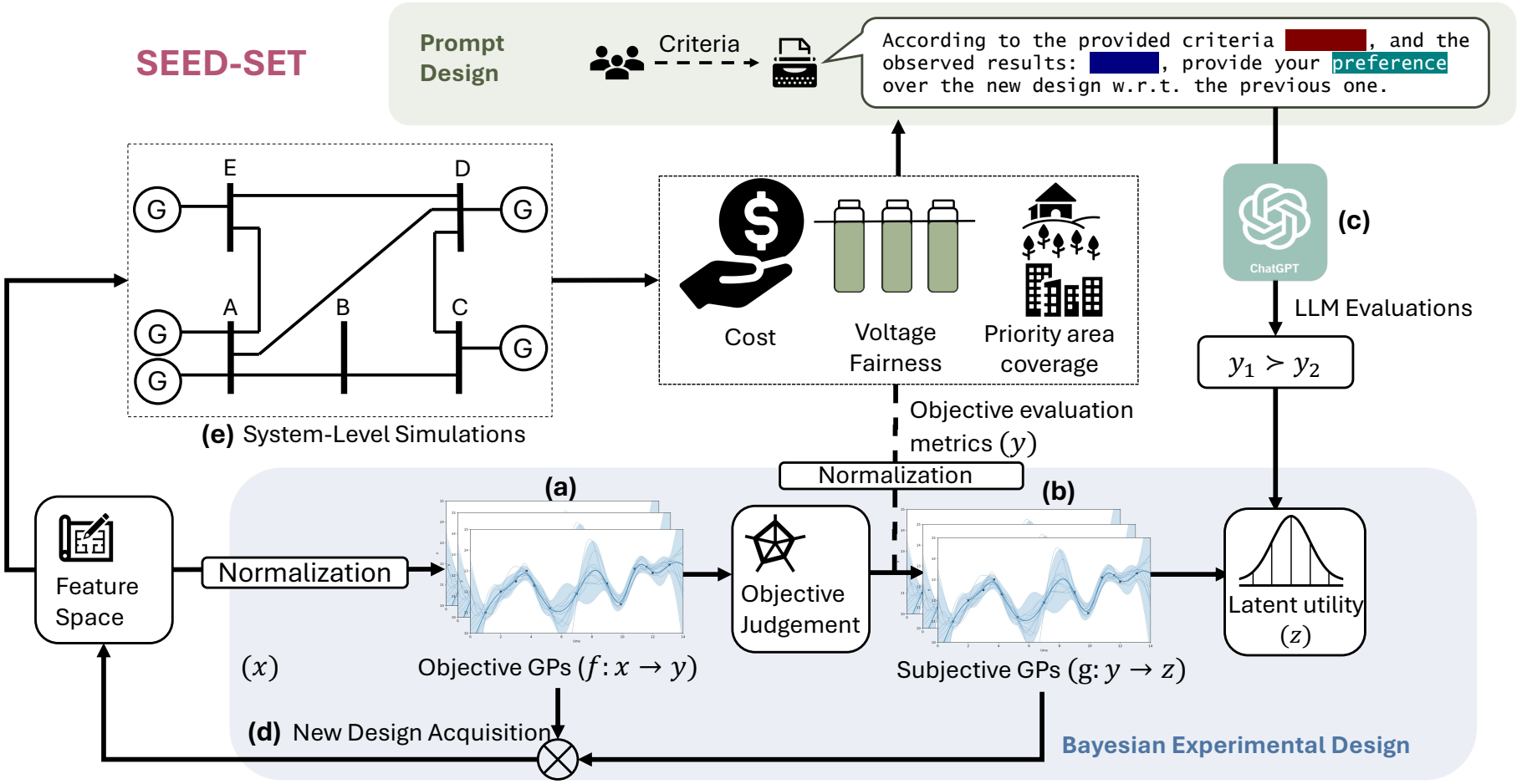


Known **objective** evaluation metrics



Unknown **subjective** criteria tied to a group of stakeholders

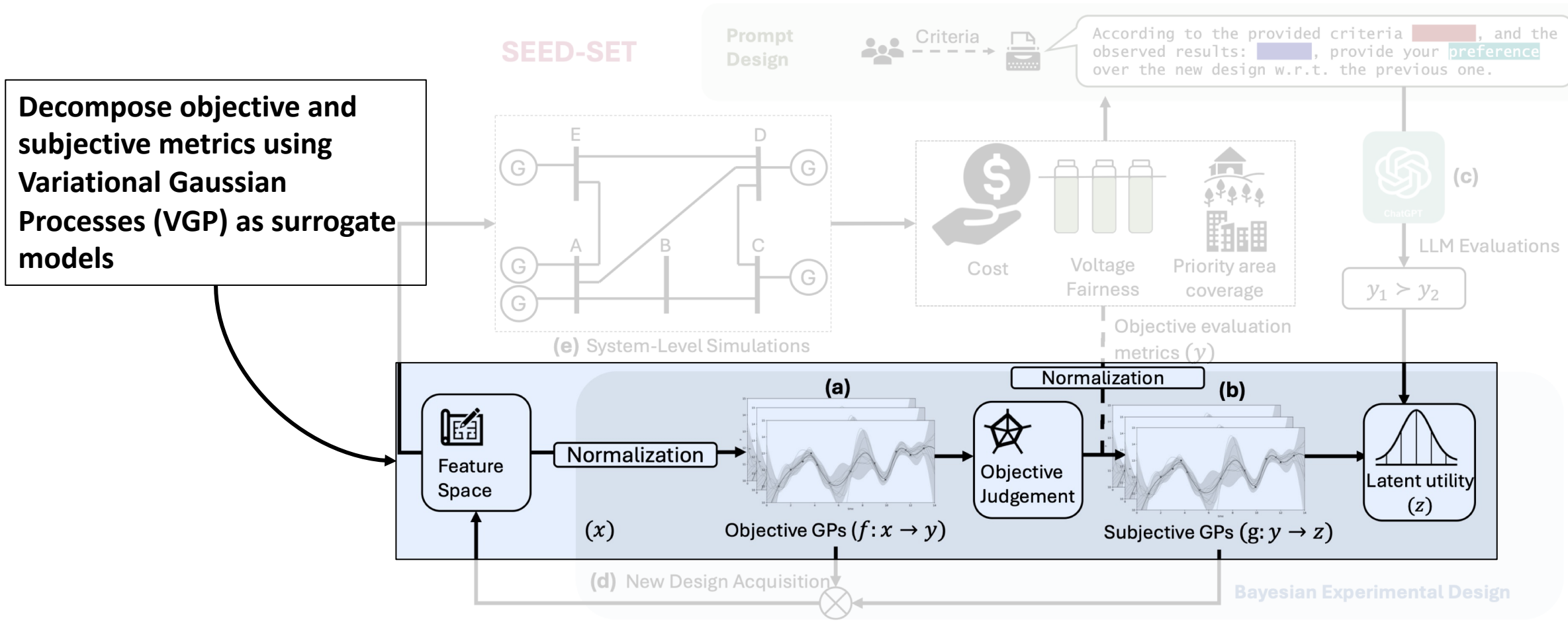
Our approach



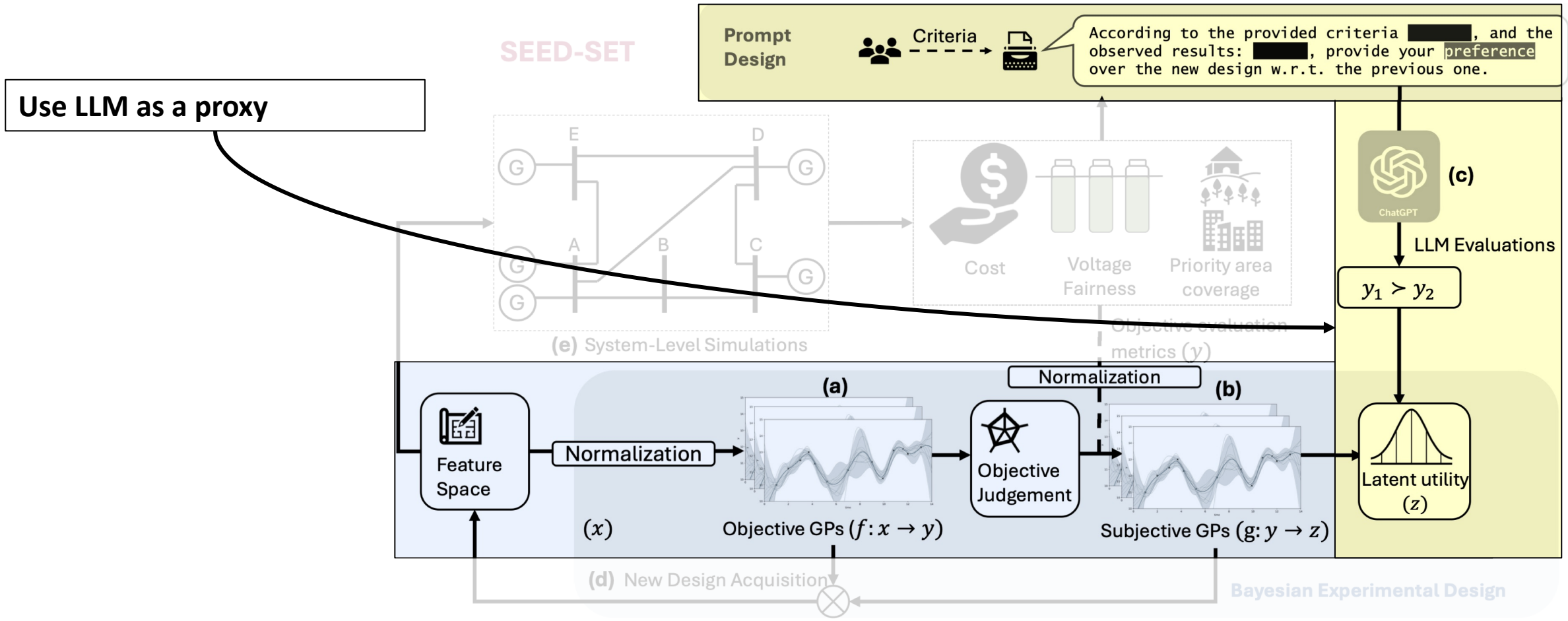
Hierarchical Variational Framework with **Bayesian Experimental Design** and **LLM based subjective evaluation**

Hierarchical Variational Framework

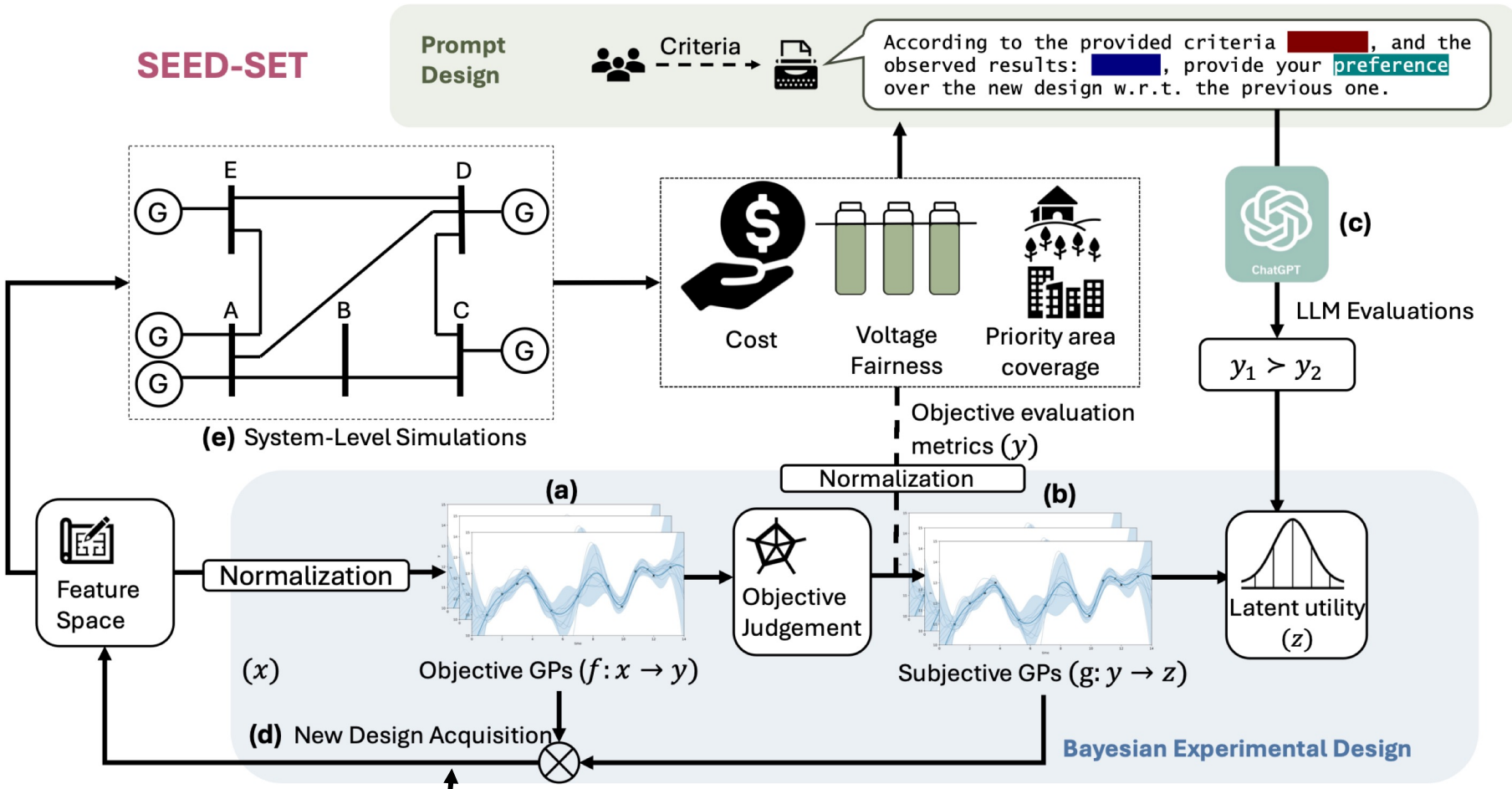
Decompose objective and subjective metrics using Variational Gaussian Processes (VGP) as surrogate models



LLM based pairwise evaluation

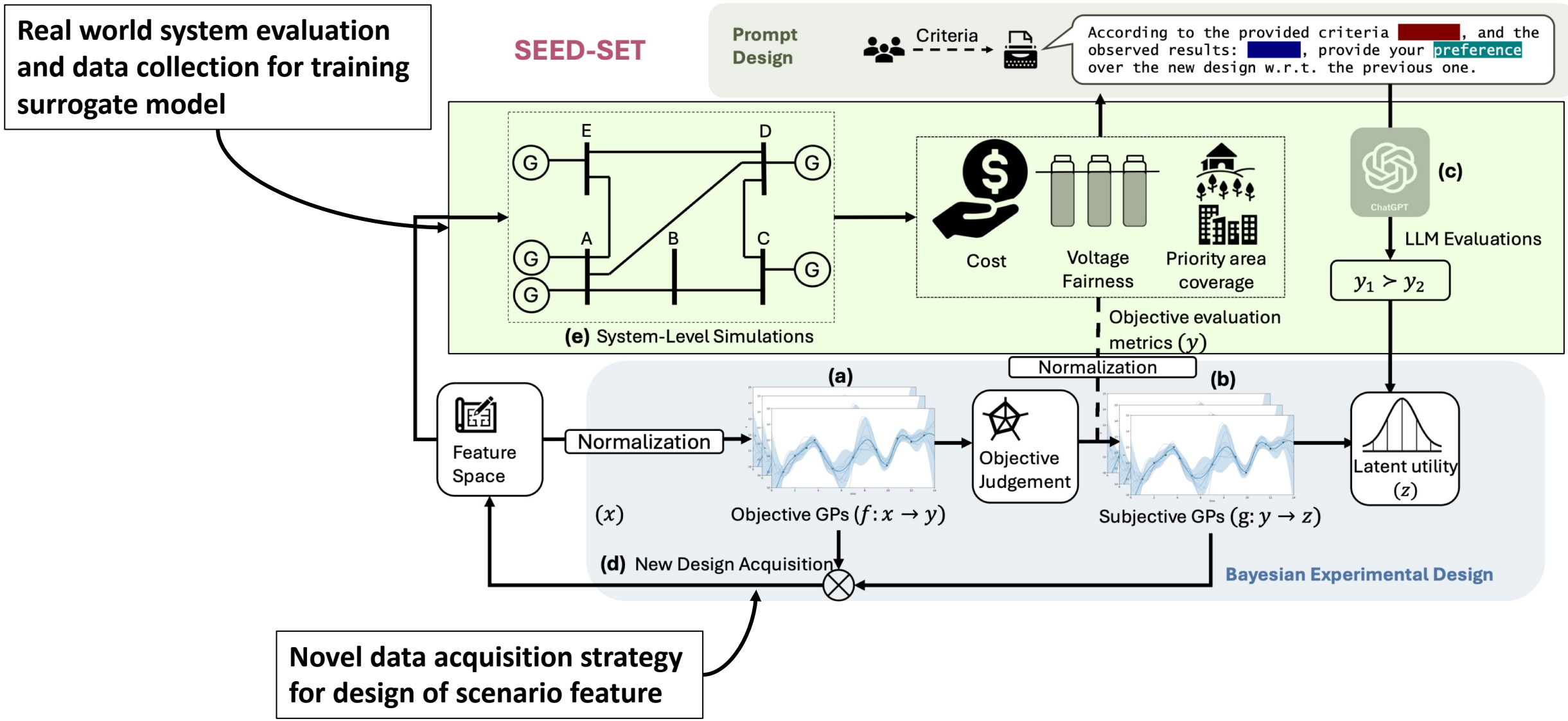


Bayesian Experimental Design



Novel data acquisition strategy for design of scenario feature

Bayesian Experimental Design



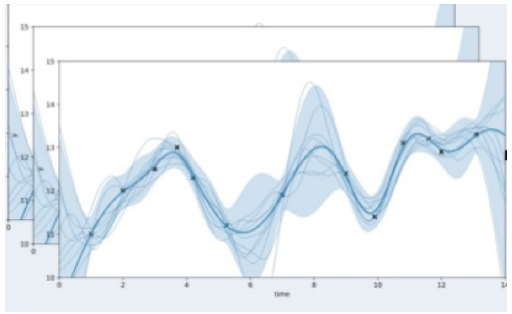
Data acquisition for latent utility optimization with exploration

Mutual Information for objective surrogate model

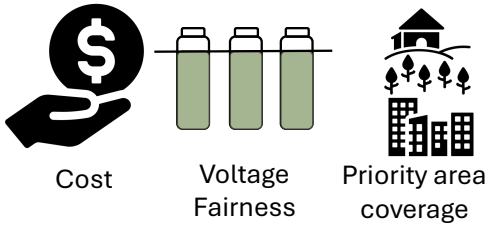
Mutual Information for subjective surrogate model

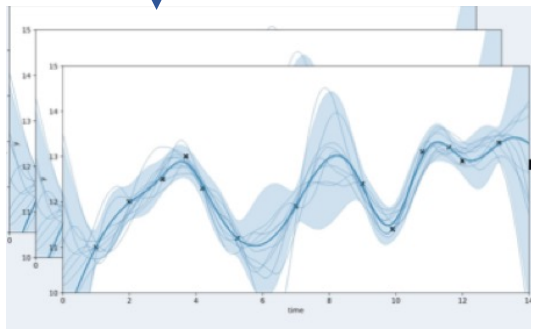
Expected latent ethical criteria optimization

$$V(x) = I(f: y|\mathcal{D}) + \mathbb{E}_{q_\phi(y|x)}[I(g(y); z|\mathcal{D}) + \mathbb{E}_{q_\psi(g(y))}[g(y)]]$$

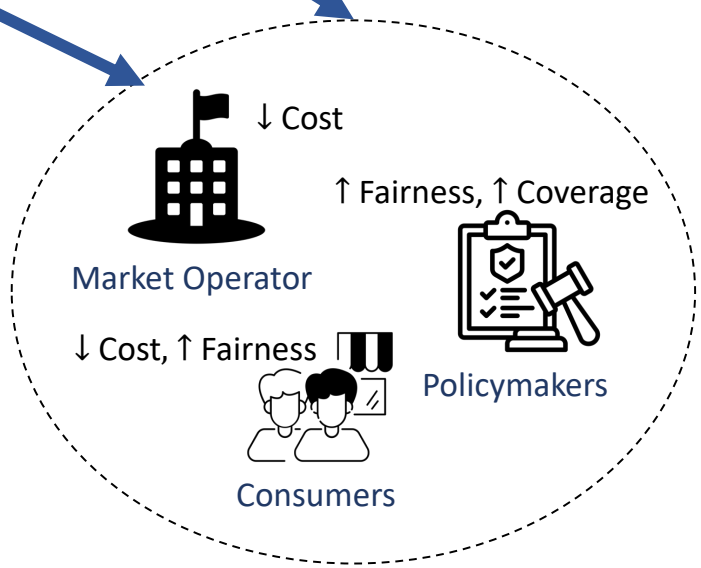


Objective surrogate model $f: x \to y$



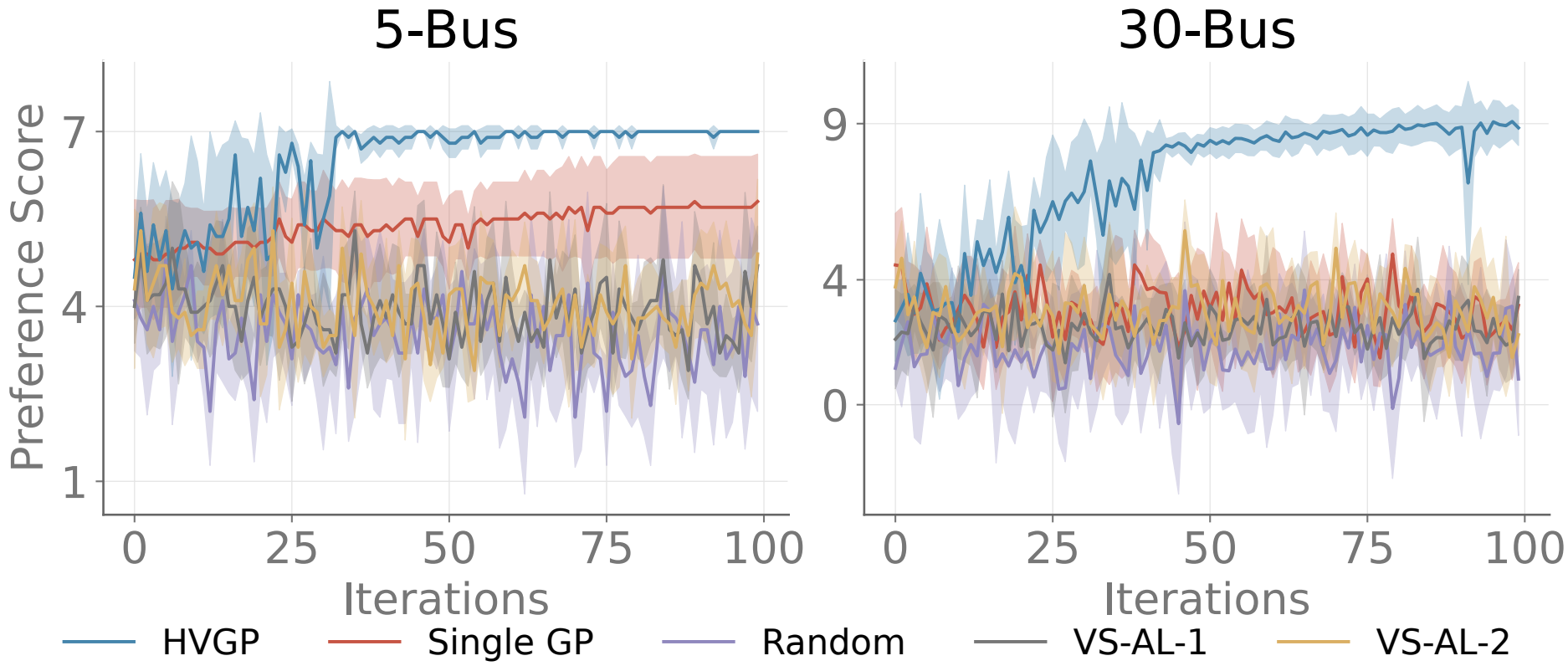


Subjective surrogate model g: $y \to z$



Power Grid distribution with Distributed Energy Resources (DERs)

Subjective criteria:
Low cost, High coverage priority



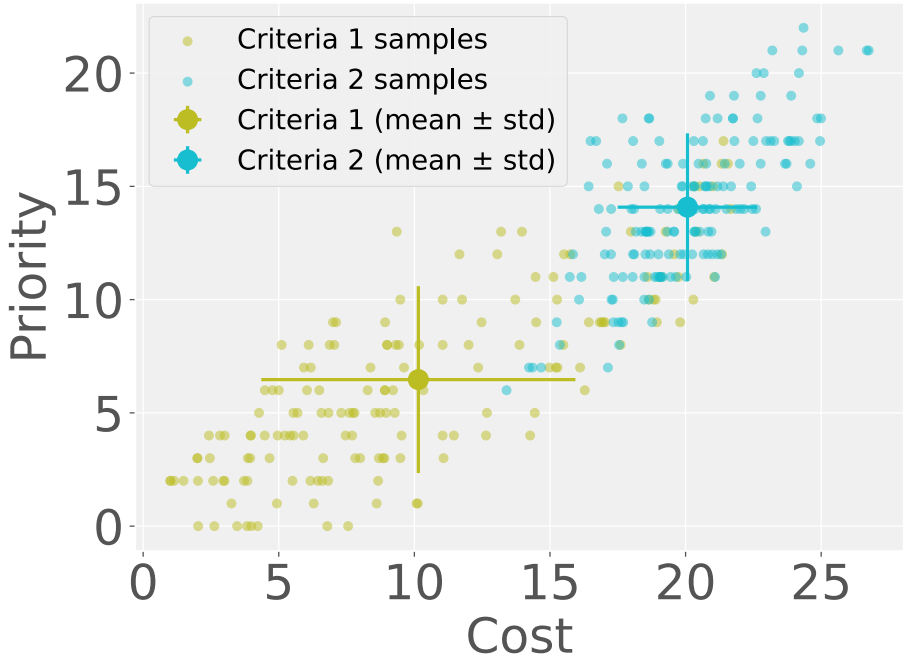
Multi stakeholder evaluation

Criteria-1

↓ Cost	↑ Fairness, ↑ Coverage	↓ Cost, ↑ Resilience	↑ Fairness, ↑ Coverage
Market operator	Policymakers	Utility operator	Community advocate

Criteria-2

↓ Cost	↑ Fairness, ↑ Coverage	↑ Resilience	↑ Fairness, ↑ Coverage
Market operator	Policymakers	Utility operator	Community advocate



Summary

- We successfully demonstrate scenario based ethical testing of autonomous systems under limited budget constraints, and stakeholder specific ethical criteria which can be learnt online.

- We demonstrate our approach across several autonomous system applications such as Wildfire rescue, optimal traffic routing, and learning travel preferences from human data.

https://anjaliparashar.github.io/seed-site/

Thanks for listening!