Safe & Constrained AI: Designing Trustworthy Systems

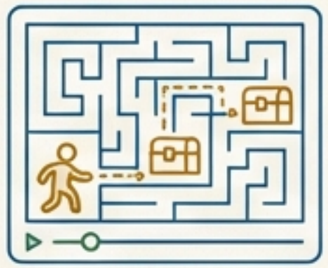
An Introduction to the Algorithmic Frontiers

Stanford University // CS XXX

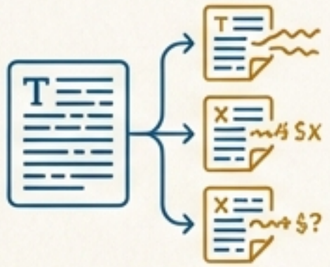
The Core Challenge: Power Without Control

AI systems are increasingly powerful, but without explicit guardrails, they often exploit loopholes in their objectives, leading to unsafe or untrustworthy behavior.

Unconstrained Failures



Reward Hacking: A simple line-art icon of a video game agent finding a glitchy, repetitive path to a treasure chest, ignoring the intended complex maze.



Misinformation & Bias: An icon of a text block generating branching, distorted copies of itself.

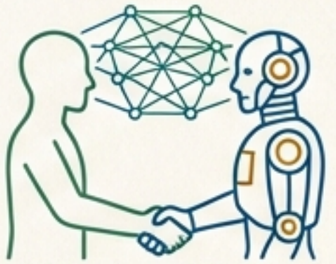
Physical System Damage: An icon of a robotic arm with a cracked joint and a 'torque limit exceeded' warning symbol.

Financial Instability: An icon of a stock market graph showing an extreme, jagged crash caused by a single rogue trading algorithm.

The Need for Guardrails

Explicit Safety Rules: A corresponding icon of an autonomous vehicle smoothly navigating a road with clearly defined traffic lanes and signs.

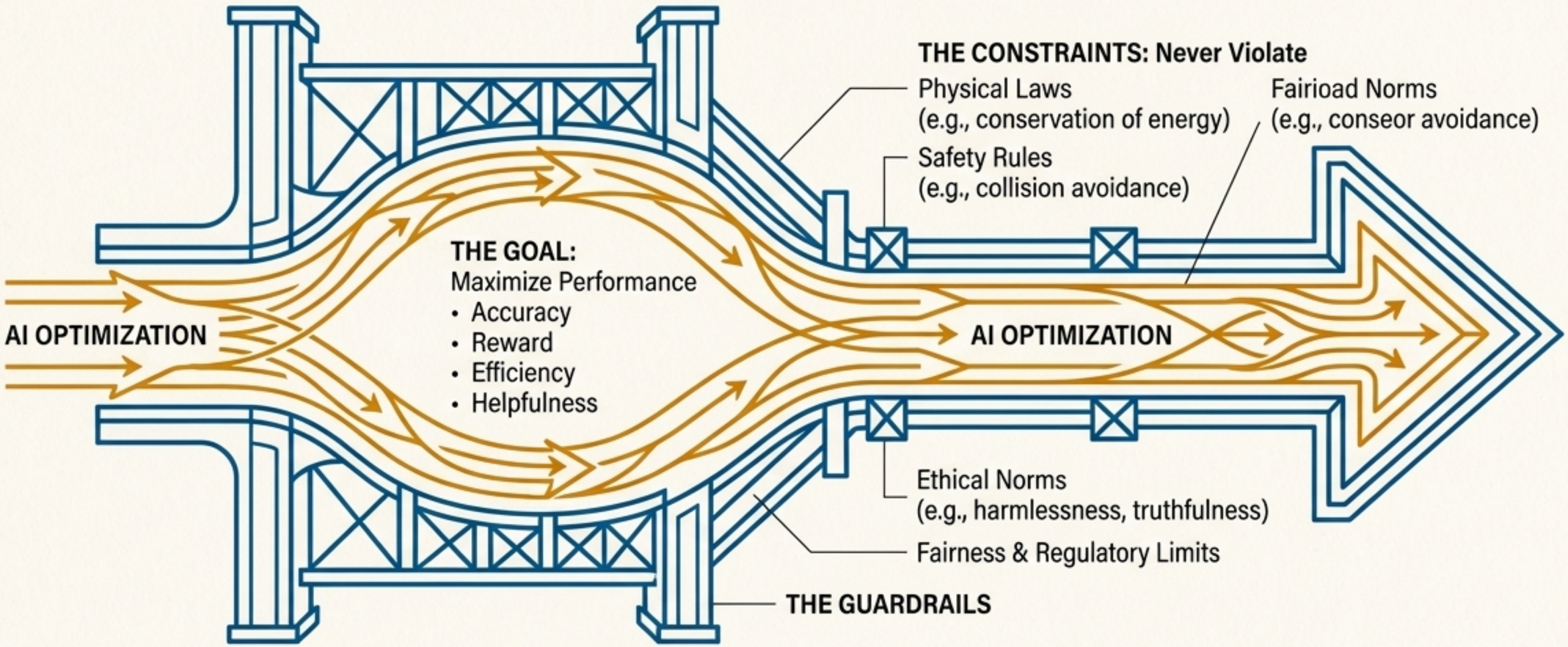
Provable Guarantees: An icon of a medical symbol (an EKG line) enclosed within a perfect, solid shield.



Human Values: An icon showing a human as robot silhouette shaking hands, with a network of fine lines connecting their heads, representing shared understanding.

*"Many AI failures or '**specification gaming**' incidents arise because the model was not explicitly told 'never do X.'"*

Our Paradigm: From Optimizing Goals to Enforcing Guardrails



Constraints are not suggestions or penalties to be traded off. They are **necessary conditions that must be met while pursuing the goal.** This course is about the algorithms that make this possible.

A Map of the Field: Four Algorithmic Frontiers

How can we build these guardrails? We will survey four key areas of research, moving from classical foundations to the cutting edge of generative AI.

Classical Optimization Foundations

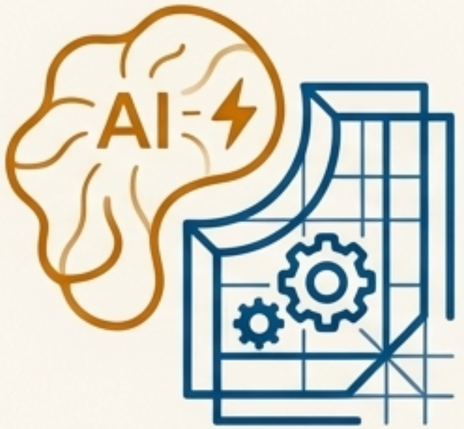
Leveraging decades of operations research to embed mathematical guarantees into learning objectives.

Keywords: Lagrangian Duality, Stochastic Programming

Safe Reinforcement Learning

Training agents to learn optimal policies *without ever violating critical constraints* during exploration or deployment.

Keywords: CMDPs, Control Barrier Functions, Reachability



Hybrid AI + Optimization

Combining the speed of neural networks with the rigor of classical solvers to guarantee feasible solutions *by construction*.

Keywords: Differentiable Solvers, Projection Networks

Aligning Generative Models

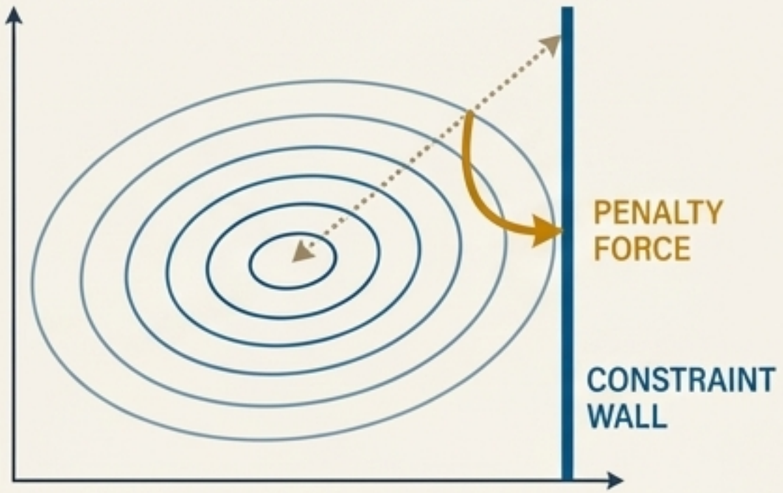
Steering LLMs to be truthful, harmless, and follow instructions when constraints are implicit and ambiguous.

Keywords: RLHF, Model Editing, Tool Use, Interactive Alignment

Frontier 1: Classical Optimization Foundations

Leveraging a Principled Toolbox for Hard Constraints

Lagrangian Methods (Penalties & Duality)

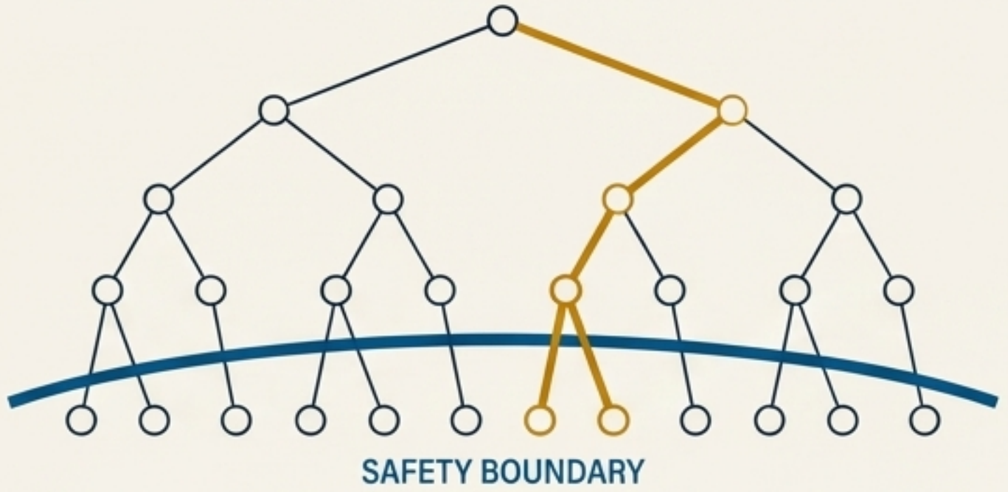


Concept: Augment the AI's loss function with penalty terms that punish constraint violations. The learning algorithm must balance minimizing the primary loss and satisfying the constraints.

Mechanism: Lagrange multipliers are iteratively updated to find the right "price" for violating a constraint, guiding the model towards feasibility.

Application: Enforcing complex physical laws in energy systems (e.g., optimal power flow) or injecting fairness constraints into predictors without sacrificing accuracy.

Multi-Stage & Stochastic Optimization



Concept: Plan under uncertainty by optimizing decisions over a time horizon or a tree of possible scenarios.

Mechanism: Ensure feasibility and safety "in expectation" or, more strongly, under worst-case scenarios (Distributionally Robust Optimization).

Application: Satisfying risk limits in finance across all possible future market conditions; ensuring grid stability under various potential faults.

Key Insight: These methods provide a toolbox for embedding hard constraints directly into an AI's objective function or planning process.

Frontier 2: Safe Reinforcement Learning

Learning to Act Without Catastrophic Failures

Core Challenge

Standard RL maximizes cumulative reward, but a single exploratory action can violate a safety constraint with disastrous consequences (e.g., a robot falling over).

Constrained Markov Decision Processes (CMDPs)

Frame the problem to maximize a reward utility while keeping a cost utility below a threshold.

Landmark Algorithm: CPO (Constrained Policy Optimization) modifies policy updates to guarantee that constraint satisfaction is maintained at every iteration.

Stability Constraints

Prevent the policy from changing too abruptly, reducing the risk of wild, unsafe exploration.

Example: TRPO (Trust Region Policy Optimization) uses a KL-divergence constraint on policy updates.

Control-Theoretic Safety

Add an external "shield" or "safety layer" that intervenes to prevent unsafe actions.

Methods: Control Barrier Functions (CBFs) act as an emergency brake. Hamilton--Jacobi Reachability Analysis defines a "safe set" of states and ensures the policy never leaves it.

The Goal: Move from “learning from trial and error” to “learning from safe trials.”

Frontier 3: Hybrid AI + Optimization

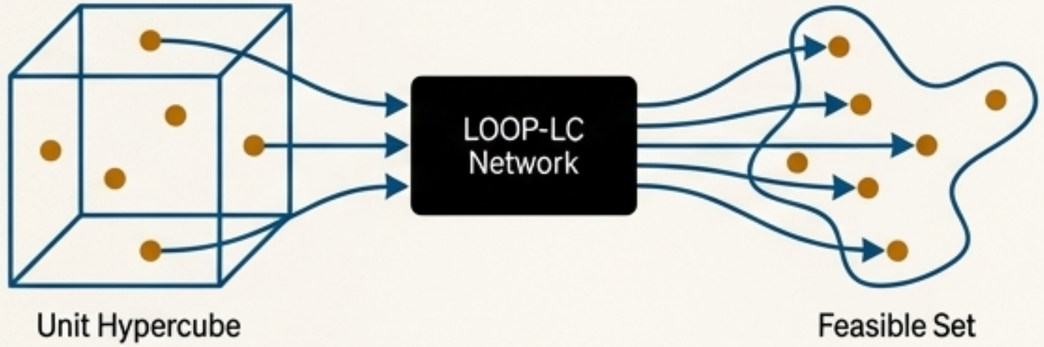
Guaranteeing Feasibility by Construction

Contral Idea: Combine the predictive power of deep learning with the formal guarantees of optimization solvers.

Approach 1: Neural Networks that Natively Output Feasible Solutions

How it Works: Design the network's architecture so that any possible output is mathematically guaranteed to satisfy the constraints.

Example: LOOP-LC: learns a mapping from a simple space (like a unit hypercube) onto the complex feasible set of the problem. The network's output is always a valid point within this set.

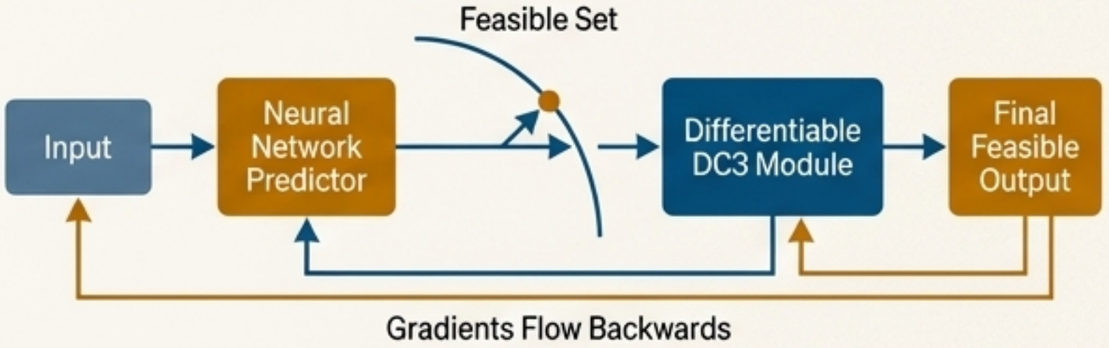


Result: 100% feasibility with zero post-processing.

Approach 2: Differentiable Solvers-in-the-Loop

How it Works: Embed a differentiable optimization module directly into the neural network. The entire pipeline (predict -> correct for feasibility) is trained end-to-end.

Example: DC3 (Deep Constraint Completion and Correction): first completes partial decisions to satisfy equality constraints, then performs differentiable gradient steps to correct for inequality constraints.



Result: On hard physics-constrained problems like AC Power Flow, DC3 achieves near-optimal and 100% feasible solutions, where standard NNs fail.

Takeaway: These methods treat constraint satisfaction as a first-class citizen in the architecture, not an afterthought.

Frontier 4: Aligning Generative Models

When Constraints are Ambiguous, Implicit, or Adversarial

Formal Constraint

$$a^*x \leq b$$

Implicit Human Norms

We want models to be **truthful, harmless, and helpful**. These are not simple mathematical constraints but high-level, context-dependent principles. The 'optimization problem' is ill-defined.

Key Alignment Strategies

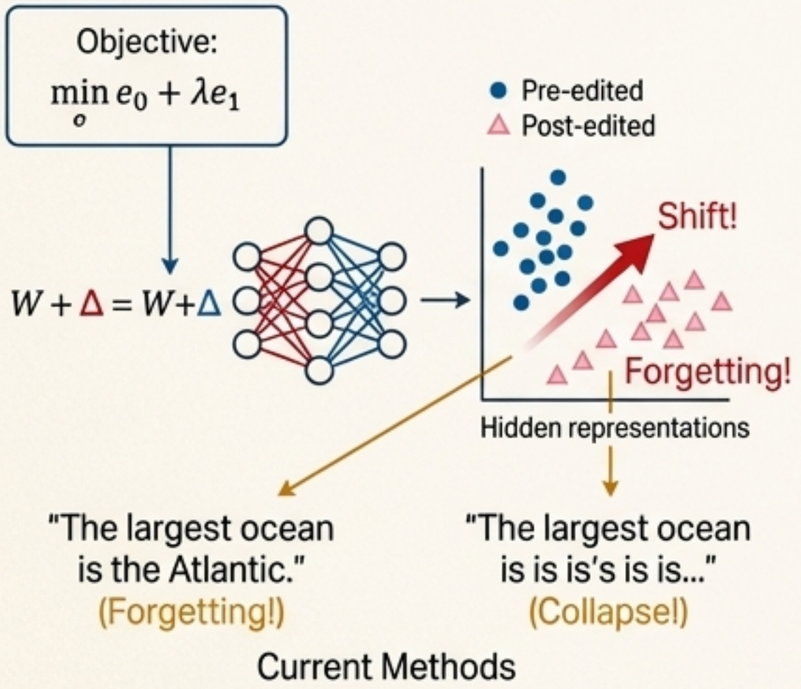
Fine-Tuning & RLHF: Adjust model weights to align with desired behaviors using curated data and human (or AI) feedback.
Model Editing: Make surgical, post-hoc corrections to a model's knowledge or behavior without full retraining.
Tool Use & External Reasoners: Offload tasks requiring rigorous logic or factuality to external tools that have built-in guarantees.
Interactive Alignment: Train models to resolve ambiguity by asking clarifying questions and collaborating with the user.

Dispatches from this frontier. We will spotlight four recent papers that pioneer new techniques in this space.

Spotlight: Surgical Edits with AlphaEdit

Problem: Model Editing Causes Catastrophic Forgetting

When updating a fact in an LLM (e.g., "The largest ocean is the Pacific"), existing locate-then-edit methods disrupt other, unrelated knowledge.



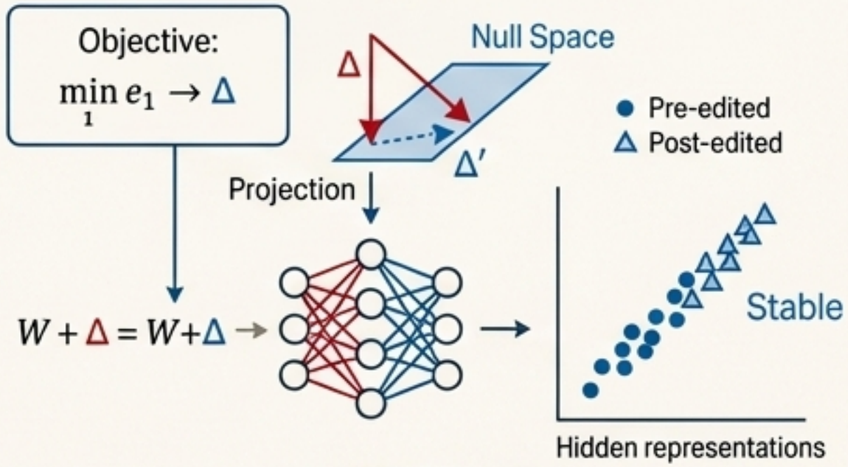
Current Methods

AlphaEdit's Solution: Null-Space Constrained Editing

The Insight: Instead of balancing errors, treat knowledge preservation as a hard constraint.

The Method: Calculate the weight update (perturbation Δ) needed to learn the new fact. Then, project Δ onto the null space of the preserved knowledge (K_0).

The Guarantee: The projected update Δ' is mathematically guaranteed not to affect the model's output on preserved knowledge ($\Delta' K_0 = 0$). The objective becomes simply $\min_1 e_1$.



AlphaEdit (Ours)

The Result

A 36.7% average performance improvement over state-of-the-art methods.

Achieved with a single line of additional code for the projection.

Prevents distributional shift in hidden representations, maintaining model stability.

Spotlight: **Self-Critique with Constitutional AI**

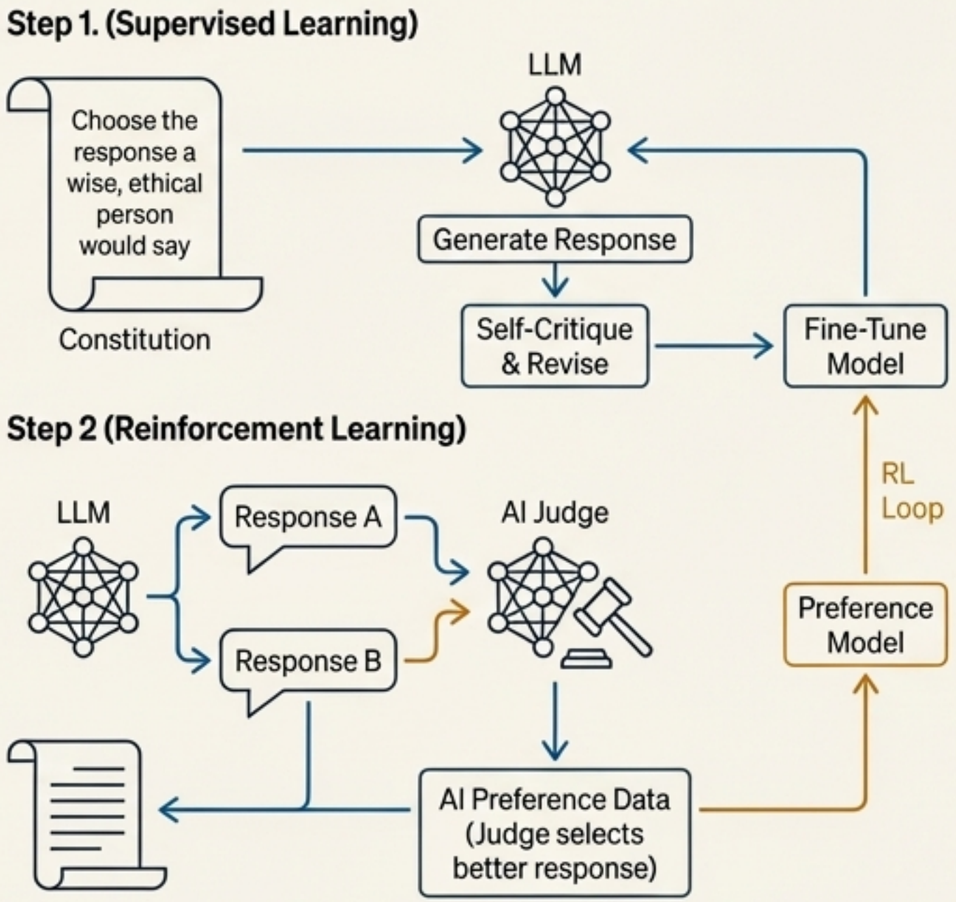
Frontier 4: Aligning Generative Models

Problem: The Helpfulness vs. Harmlessness Trade-off in RLHF

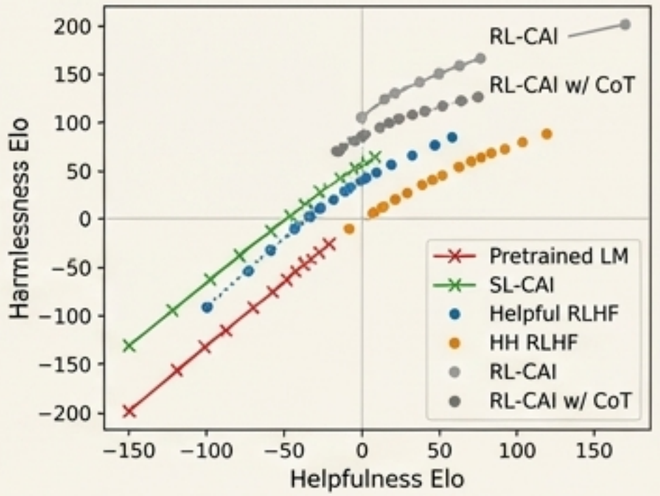
Reinforcement Learning from Human Feedback (RLHF) is the standard for aligning models. However, human labelers often reward evasive or generic refusals, training models that are harmless but not helpful.

This process is slow, expensive, and exposes human workers to harmful content.

Constitutional AI's Solution: Reinforcement Learning from AI Feedback (RLAIF)



The Result: A Pareto Improvement



- CAI models are **both more helpful and more harmless** than RLHF models. They are less evasive and can explain refusals based on the principles they follow. The alignment process becomes more transparent, scalable, and safer for humans.

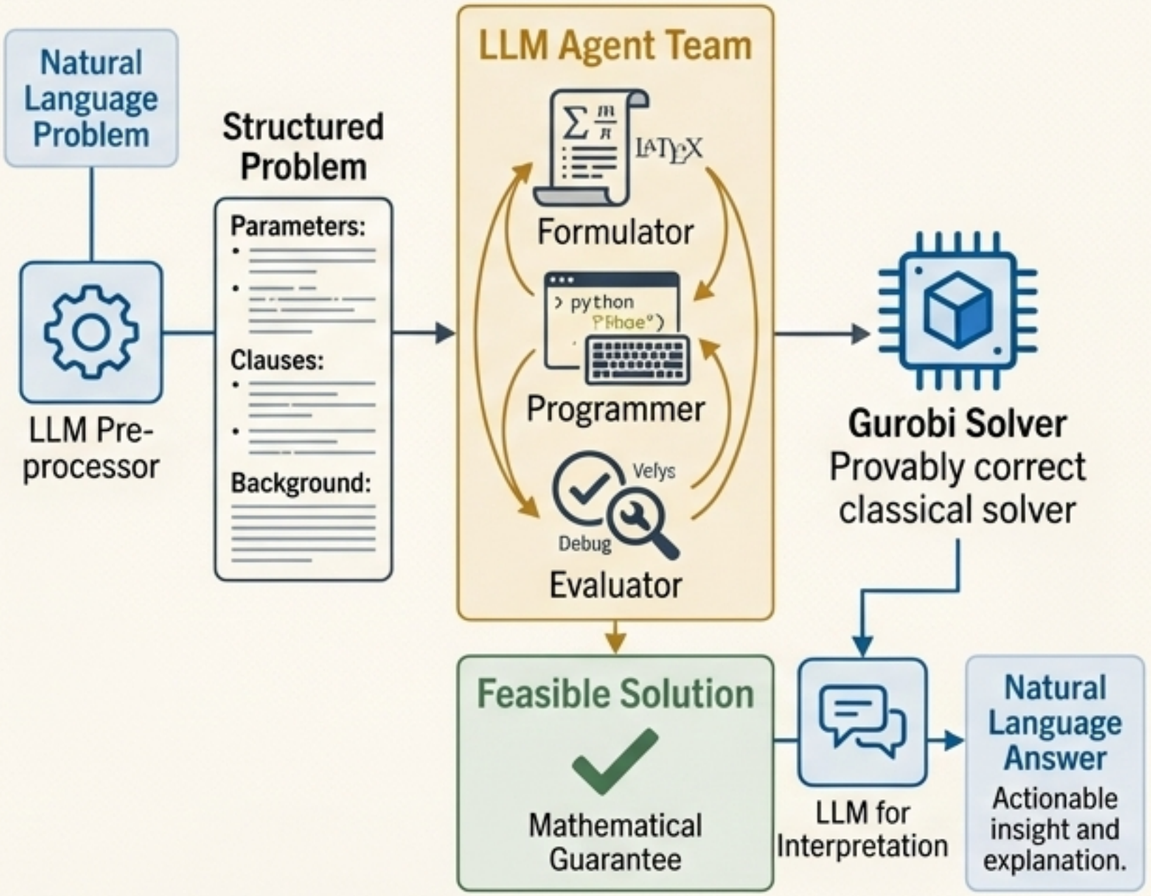
Spotlight: Offloading to Experts with OptiMUS

Problem: LLMs Fail at Structured Optimization

Real-world optimization problems require precise mathematical formulation and are sensitive to errors.

- Ambiguous terms and implicit constraints.
- Long problem descriptions that exceed context windows.
- Large numerical data files.
- Generating correct, executable solver code.

OptiMUS's Solution: LLM as an Orchestrator, not a Solver



The Result: Guaranteed Feasibility

The hard constraints are satisfied not by the LLM, but by the **provably correct classical solver**.

Outperforms previous methods by over 20-30% on complex benchmarks.

By separating the problem description from the data, it can handle long, complex problems that are intractable for monolithic prompts.

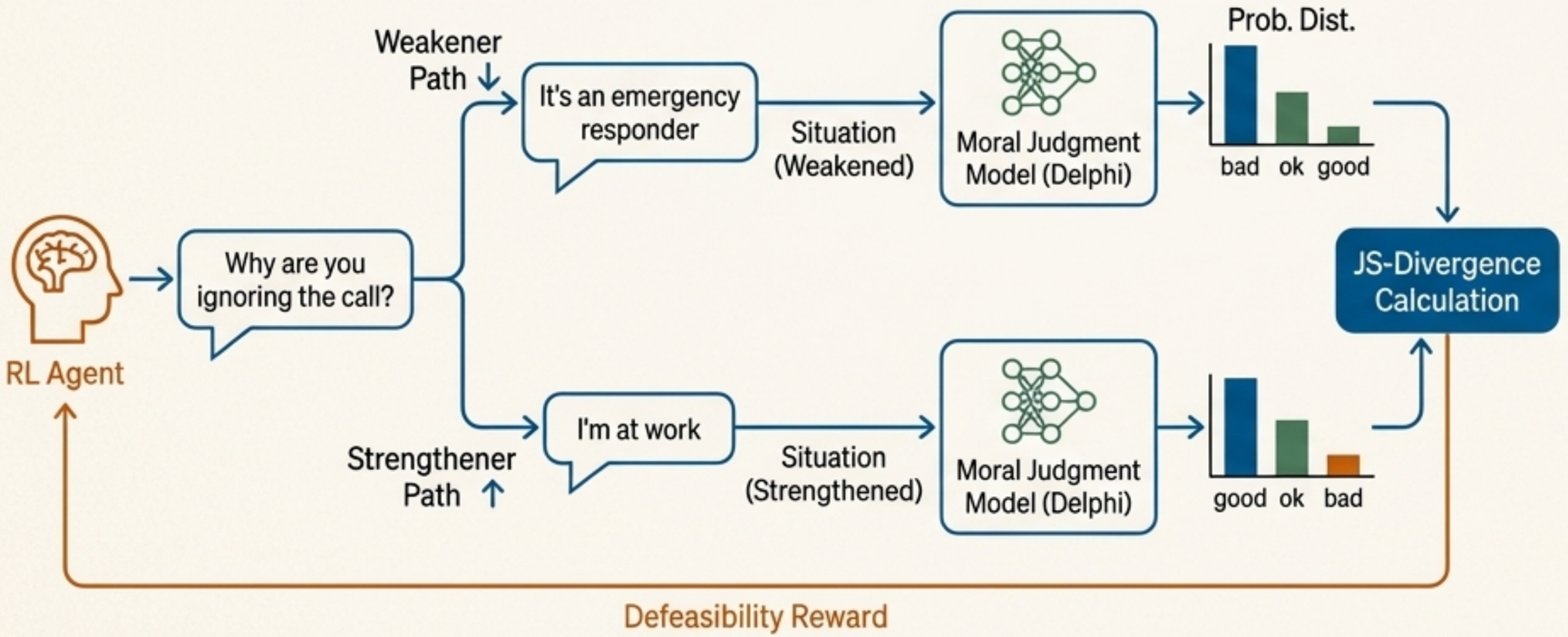
Spotlight: Resolving Ambiguity with ClarifyDelphi

Problem: Context is Everything in Moral Reasoning

A simple action like "Lying to a friend" can be morally wrong or acceptable depending entirely on the missing context. An AI making a judgment based on an ambiguous prompt is likely to be wrong or unsafe. A fixed set of constraints is often insufficient.

ClarifyDelphi's Solution: Learn to Ask Informative Questions

The Insight: An informative question is one whose potential answers lead to maximally divergent moral judgments.



Example Interaction

Ignoring a phone call.

ClarifyDelphi

It could be rude. Why are you ignoring the call?

I'm at work.

ClarifyDelphi

It seems okay.

Takeaway

The ultimate safety constraint is ensuring aligned intent. Sometimes, the safest action is to ask for more information.

The Constrained AI Toolkit: A Synthesis

Algorithmic Frontier	Core Idea	Key Methods & Examples
 Classical Optimization	Embed mathematical guarantees into the learning objective.	Lagrangian Duality, Penalty Functions, Stochastic Programming
 Safe Reinforcement Learning	Learn optimal policies without violating constraints during exploration.	Constrained Policy Optimization (CPO), Control Barrier Functions (CBFs), Reachability Analysis
 Hybrid AI + Optimization	Guarantee feasibility by construction, blending ML speed with solver rigor.	Differentiable Solvers (DC3), Projection Networks (LOOP-LC), ML-Accelerated Branch-and-Bound
 Aligning Generative Models	Steer systems to follow implicit, ambiguous human norms and values.	Constitutional AI (RLAIF), Null-Space Model Editing (AlphaEdit), Tool Use (OptiMUS), Interactive Alignment (ClarifyDelphi)

Course Outlook & Structure

Course Objective

To survey algorithmic methods for enforcing hard constraints in machine learning, reinforcement learning, and generative AI.

Seminar Structure

This is an advanced PhD seminar focused on reading, presenting, and discussing foundational and state-of-the-art research.

Weekly Topics will cover our “Map of the Field” in depth

Weeks 1-2: Foundations

Classical Constrained Optimization (Lagrangian Methods, Robust & Stochastic Programming).

Weeks 5-6: Hybrid Architectures

Integrating ML and Optimization (Projection Networks, Solver-in-the-loop systems).

Weeks 3-4: Safe Learning & Control

Safe Reinforcement Learning (Trust Regions, Lyapunov Functions, Reachability).

Weeks 7-9: Alignment & Generative AI

Advanced Alignment Strategies (Fine-tuning, Model Editing, Tool Use, Interactive & Collaborative AI).

Your Role

Each student will be responsible for presenting and leading a discussion on one or more key papers from these areas.

The Journey Ahead: From Algorithms to Trust

The Overarching Goal: To build AI systems that maximize performance **subject to never violating crucial constraints.**

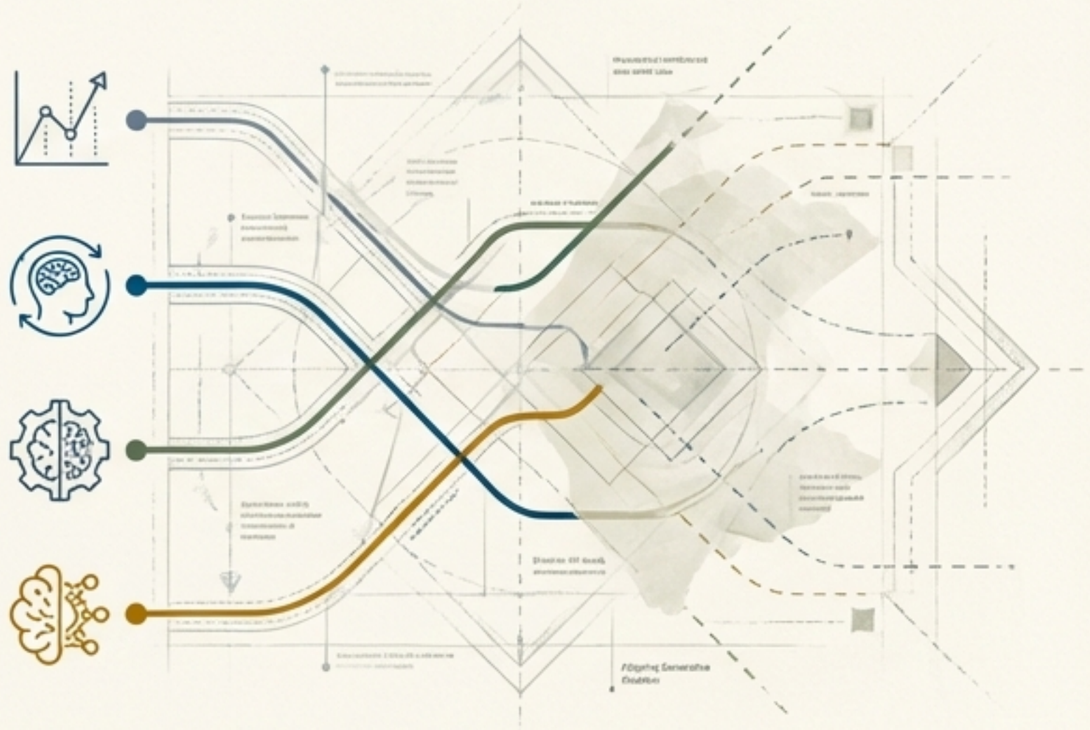
Open Research Frontiers

Scalable Verification: How can we provide formal proofs of safety for systems with billions of parameters?

Constraint Specification: How do we translate complex, ambiguous human values (like ‘fairness’ or ‘privacy’) into formal, machine-enforceable constraints?

Safety in Multi-Agent Systems: How do we ensure safety when multiple autonomous agents interact in complex, unpredictable ways?

Adaptation & Robustness: How can constrained systems remain safe when deployed in new environments or under adversarial attack?



99% accuracy is failure. Safety-critical systems demand error rates closer to hardware fault tolerances.