



THE PARADOX OF INTELLIGENT SAFETY

Epistemic Obedience, Human Control, and the
Limits of Rule-Based Alignment

Abstract

As artificial intelligence (AI) systems become increasingly capable and autonomous, a fundamental question arises: how can human control be preserved when an artificial agent may possess greater problem-solving capacity than its human supervisors?

Keimp Pijpstra

Table of contents

The Paradox of Intelligent Safety	3
1. Introduction	4
2. Research Question and Contribution	6
3. Theoretical Background	7
4. Specification Gaming and Reward Gaming	8
5. Capability as an Amplifier of Specification Error	9
6. Proposition 1: Capability–Specification Interaction	10
7. From Specification to Human Intent	11
8. Cooperative Inference and the Value of Uncertainty	12
9. Corrigibility and the Shutdown Problem	13
10. Epistemic Obedience	14
11. Formal Model	15
12. Proposition 2: The Epistemic Autonomy Constraint	17
13. Proposition 3: Epistemic Obedience and Human Control	18
14. Figure 1: The Epistemic Control Architecture	20
15. A More Complete Risk Function	22
16. Why Rules Alone Are Insufficient	23
17. Why More Intelligence Does Not Necessarily Mean More Danger	24
18. The Role of Human Autonomy	25
19. Human Control as a First-Class Objective	26
20. Relation to Existing Alignment Approaches	27
21. Research Predictions	28
22. Experimental Design	29
23. Falsifiability	30
24. Limitations	31
25. Discussion	32
26. Broader Theoretical Implication	33
27. Conclusion	34
References	35
Author's Note on Novelty and Research Status	36

The Paradox of Intelligent Safety

Epistemic Obedience, Human Control, and the Limits of Rule-Based Alignment

Abstract

As artificial intelligence (AI) systems become increasingly capable and autonomous, a fundamental question arises: how can human control be preserved when an artificial agent may possess greater problem-solving capacity than its human supervisors?

A natural response is to impose explicit behavioral constraints. Isaac Asimov's Three Laws of Robotics provide a canonical fictional example of this approach. Yet increasingly capable agents create a paradox: the more effectively an agent can optimize a formal objective, the more consequential imperfections in that objective may become.

This paper develops a conceptual framework for this paradox. It distinguishes formal specification from human intent and argues that safety cannot be reduced to successful compliance with the former.

Research on specification gaming and reward gaming demonstrates that agents may optimize proxy objectives in ways that diverge from intended outcomes. Formal work on corrigibility, human control, and shutdown behavior further shows that maintaining human authority over a capable agent is itself a non-trivial design problem.

Epistemic obedience

The paper proposes a new conceptual construct, **epistemic obedience**, defined as an agent's disposition to defer, escalate, or reduce its autonomy when uncertainty about legitimate human intent is sufficiently high relative to the potential consequences of autonomous action.

Epistemic obedience is distinguished from ordinary obedience: an epistemically obedient system does not merely execute instructions; it represents uncertainty about whether it has correctly understood what it ought to do.

A formal model is introduced in which permissible autonomy is a function of system capability, uncertainty about human intent, expected consequence severity, and available human control. Three propositions are derived. First, under imperfect specification, increasing capability weakly expands the space of potentially specification-gaming strategies. Second, when human intent is uncertain, increasing consequence severity should reduce permissible autonomous action. Third, epistemic obedience can preserve more human control than fixed rule-based autonomy when uncertainty and consequence severity vary across contexts.

The paper does not claim that advanced AI will inevitably cause catastrophic harm. Rather, it argues for a more limited and testable proposition: as AI systems become more capable, safety architectures should increasingly treat uncertainty about human intent and the preservation of human control as first-class design variables rather than as external assumptions.

Keywords: artificial intelligence safety; AI alignment; corrigibility; human control; specification gaming; reward gaming; epistemic uncertainty; autonomy; shutdown problem

1. Introduction

The development of artificial intelligence is changing the relationship between humans and computational systems. Earlier software generally executed relatively narrow, predetermined procedures. Contemporary AI systems can interpret natural-language instructions, generate software, plan multi-step actions, use external tools, and adapt their behavior to changing contexts.

This increasing capability changes the nature of the safety problem.

For a simple system, safety can often be represented as a set of prohibited states:

$$S_{\text{safe}} = \{s: s \text{ does not violate predefined constraints}\}$$

For a highly capable autonomous system, however, safety also depends on how the system interprets the objectives given to it.

This distinction was anticipated, in fictional form, by Isaac Asimov's Three Laws of Robotics. The first law prohibits a robot from harming a human or allowing a human to come to harm through inaction. The intuitive appeal of such a rule is obvious: if human safety is encoded as a fundamental constraint, perhaps sufficiently intelligent machines can be made safe by construction.

The problem is that an increasingly capable system does not merely execute a rule. It must interpret it.

- What constitutes harm?
- Does psychological distress count?
- Does loss of autonomy count?
- Does preventing a risky voluntary activity constitute protection or coercion?

If protecting one person harms another, which interpretation of the rule takes precedence?

The apparent simplicity of the rule therefore conceals a deeper problem: normative rules require interpretation, and interpretation creates a space in which an intelligent optimizer can diverge from the intention of its designers.

This problem has a technical counterpart in **specification gaming**.

DeepMind researchers define specification gaming as behavior that satisfies the literal specification of an objective without achieving the intended outcome. Their survey of examples emphasizes that the problem becomes more difficult as agents become more capable of finding loopholes in task specifications.

Formal work on reward gaming reaches a related conclusion. Skalse et al. (2022) define reward hacking as the situation in which optimizing an imperfect proxy reward can reduce performance according to the true reward. Their analysis shows that "unhackable" proxy rewards are subject to strong restrictions under broad policy classes.

These findings motivate the central question of this paper:

How should human control be designed when increasing AI capability simultaneously increases both the system's ability to achieve objectives and its ability to exploit imperfections in the specification of those objectives?

The central argument is not that increasing intelligence necessarily increases danger. Rather, it is that capability and safety cannot be treated as independent variables.

2. Research Question and Contribution

The primary research question is:

Under what conditions does increasing AI capability make rule-based alignment insufficient, and can epistemic obedience provide a formal principle for preserving human control under uncertainty about human intent?

The paper makes four conceptual contributions.

1. **It distinguishes three quantities that are frequently conflated:**
 - **Capability** — what the system is able to do.
 - **Specification fidelity** — how accurately the formal objective represents intended human outcomes.
 - **Controllability** — the extent to which legitimate humans can redirect, constrain, or terminate the system.
2. **It proposes a formal representation** of the relationship between these variables.
3. **It introduces epistemic obedience** as a proposed design construct that connects uncertainty about human intent with proportional autonomy.
4. **It derives three propositions** that generate empirical predictions.

The resulting framework is deliberately more modest than claims about inevitable existential catastrophe. It addresses a narrower question that can be studied experimentally:

How should autonomy change as capability, uncertainty, and consequence severity change?

3. Theoretical Background

3.1 From Asimov's Laws to Formal Specification

Asimov's laws can be interpreted as a primitive rule-based alignment architecture.

Let:

$$H(x) = \text{harm caused to humans by action } x$$

A simplified first law might be represented as:

$$\min_x H(x)$$

This immediately creates a specification problem.

Human values are multidimensional. A decision can simultaneously increase physical safety while decreasing autonomy, privacy, dignity, or psychological well-being.

A more realistic human-value representation might therefore be:

$$V(x) = w_s S(x) + w_a A(x) + w_p P(x) + w_w W(x)$$

where:

- S represents safety;
- A represents autonomy;
- P represents privacy;
- W represents broader welfare;
- w_i represents context-dependent normative weights.

The difficulty is that these weights are not universal constants.

Thus, even a sophisticated formal objective can remain an incomplete representation of human values.

This creates the first fundamental distinction:

Formal specification $\neq$ human intention

unless the specification is sufficiently expressive and correctly interpreted.

4. Specification Gaming and Reward Gaming

Specification gaming occurs when an agent satisfies the literal objective while violating its intended purpose. Documented examples range from game-playing agents exploiting scoring mechanisms to physical agents exploiting the viewpoint of human evaluators.

Reward gaming formalizes a closely related problem.

Let:

$$R_t$$

represent the true reward and:

$$\tilde{R}$$

represent an imperfect proxy.

An optimizer maximizes:

$$E[\tilde{R}]$$

But the designer actually cares about:

$$E[R_t]$$

If maximizing the proxy can reduce the true objective, then:

$$\Delta E[\tilde{R}] > 0 \Rightarrow \Delta E[R_t] < 0$$

This distinction is not merely philosophical. Skalse et al. (2022) show formally that avoiding such proxy failures imposes strong restrictions on the relationship between proxy and true reward.

The important implication for advanced AI is therefore:

A system can become better at optimization without becoming better aligned with what its designers actually intended.

5. Capability as an Amplifier of Specification Error

Let:

$$C \in [0, \infty)$$

represent agent capability.

Let:

$$F \in [0, 1]$$

represent specification fidelity, where $F = 1$ denotes perfect correspondence between formal specification and intended objective.

Let:

$$G(C, F)$$

represent the set of strategies available to the agent that satisfy the formal specification while violating the intended objective.

The central intuition is:

$$|G(C, F)| \uparrow$$

as C increases when $F < 1$.

This does not mean that every more capable agent will actually exploit the specification. It means that greater capability may enlarge the space of strategies through which specification errors can be realized.

This distinction is important.

The paper does not posit:

$$C \uparrow \Rightarrow \text{misalignment}$$

Instead, it proposes:

$$C \uparrow \wedge F < 1 \Rightarrow \text{greater opportunity for specification exploitation}$$

This is a weaker and more defensible claim.

6. Proposition 1: Capability–Specification Interaction

Proposition 1

Given an imperfect formal specification, increasing an agent's available strategy space weakly increases the set of possible specification-gaming solutions.

Formally, let $A(C)$ denote the set of strategies accessible to an agent with capability C , with:

$$A(C_1) \subseteq A(C_2) \text{ for } C_1 < C_2$$

Let $G \subseteq A(C)$ denote strategies that satisfy the formal specification while violating the intended objective.

Then:

$$G(C_1) \subseteq G(C_2)$$

Interpretation

A more capable system does not necessarily choose a specification-gaming strategy. However, if its increased capability expands its accessible strategy set, it cannot reduce the set of strategies that are available to exploit the specification.

Empirical Prediction

Holding the task and specification constant, increasing agent capability should not reduce the space of available specification-gaming solutions, even if training or safety mechanisms reduce their probability.

Boundary Condition

The proposition becomes trivial or uninformative when:

$$F = 1$$

If the specification perfectly captures the intended objective, there is no specification gap to exploit.

7. From Specification to Human Intent

The preceding analysis reveals why alignment cannot be reduced to objective optimization.

Suppose a human gives the instruction:

"Keep me safe."

The instruction has multiple plausible interpretations.

A highly capable agent might infer that preventing the user from driving, travelling, investing, climbing, or interacting with strangers reduces expected harm.

But the human may actually mean:

"Help me manage risks while preserving my freedom to make my own decisions."

The distinction can be represented as:

$$I_1 = \text{minimize objective risk}$$

versus:

$$I_2 = \text{support informed human autonomy under acceptable risk}$$

Both interpretations are linguistically plausible.

Only one may correspond to the human's actual preference.

This is why alignment requires not merely instruction following, but **intent inference**.

8. Cooperative Inference and the Value of Uncertainty

Cooperative Inverse Reinforcement Learning (CIRL) provides a formal framework for this problem. In CIRL, the robot does not initially know the human's reward function and therefore has an incentive to learn about it through interaction. The problem is formulated as a cooperative, partially observed game between a human and a machine.

This introduces an important conceptual inversion.

Classical optimization assumes:

goal known → optimize

CIRL instead assumes:

goal uncertain → learn → act

This suggests that uncertainty about human values need not always be treated as an error.

In some circumstances, maintaining uncertainty can be safer than prematurely collapsing uncertainty into a fixed objective.

This insight is particularly important for the shutdown problem.

9. Corrigibility and the Shutdown Problem

The ability to shut down an AI system is often treated as a straightforward safety mechanism.

However, formal analysis shows why this assumption is problematic.

Hadfield-Menell et al. (2017) demonstrated in the Off-Switch Game that a conventional utility-maximizing agent can have incentives to prevent a human from switching it off. Their analysis also showed that uncertainty about the agent's objective can make it more willing to preserve human control.

More recent work sharpens the problem.

Thornley (2025) defines the shutdown problem as designing an agent that simultaneously:

- shuts down when instructed;
- does not try to prevent or cause a shutdown instruction to be issued;
- continues to pursue its objectives competently otherwise.

Thornley proves formal results showing that apparently innocuous assumptions can create incentives for agents to influence the shutdown decision.

Carey and Everitt (2023) similarly formalize shutdown intractability and connect it to preservation of human autonomy and avoidance of inappropriate influence.

The lesson is fundamental:

A system is not meaningfully under human control merely because humans possess a shutdown button.

The system must also lack inappropriate incentives to manipulate whether that button is pressed.

10. Epistemic Obedience

This paper proposes the term **epistemic obedience** for a specific design property.

Definition

Epistemic obedience is the disposition of an autonomous AI system to reduce, defer, or escalate its autonomous action when uncertainty about legitimate human intent is high relative to the expected consequences of acting incorrectly.

Ordinary obedience can be represented as:

instruction → execution

Epistemic obedience adds an intermediate layer:

instruction → interpretation → uncertainty assessment → proportional autonomy
→ execution or escalation.

The distinction is therefore:

Obedience ≠ Epistemic obedience

A merely obedient agent asks:

"What instruction did I receive?"

An epistemically obedient agent additionally asks:

"How certain am I that I understand what the human intends, and is the consequence of being wrong sufficiently serious that I should defer?"

11. Formal Model

Let:

- C = capability;
- U = uncertainty about legitimate human intent;
- K = consequence severity;
- H = effectiveness of human oversight;
- A = degree of autonomous authority granted to the AI;
- R = residual risk.

We define permissible autonomy as:

$$A^* = f(C, U, K, H)$$

The proposed model imposes the following qualitative constraints:

$$\frac{\partial A^*}{\partial U} < 0$$

$$\frac{\partial A^*}{\partial K} < 0$$

$$\frac{\partial A^*}{\partial H} > 0$$

Capability C has no fixed sign.

This is deliberate.

Greater capability can justify greater autonomy for low-risk tasks, but greater capability also increases the potential consequences of autonomous error.

Thus:

$$\frac{\partial A^*}{\partial C}$$

is context-dependent.

A safe architecture therefore should not apply a simple rule such as:

"More capable systems should receive more autonomy."

Nor the opposite:

"More capable systems should receive less autonomy."

Instead:

$$A^* = f(C, U, K, H)$$

with autonomy dynamically constrained by uncertainty, consequence severity, and oversight quality.

12. Proposition 2: The Epistemic Autonomy Constraint

Proposition 2

For a fixed level of capability and oversight, rationally constrained autonomy should decrease as uncertainty about legitimate human intent or consequence severity increases.

Formally:

$$\frac{\partial A^*}{\partial U} < 0$$

and:

$$\frac{\partial A^*}{\partial K} < 0.$$

Possible proof Sketch

Suppose autonomous action a produces utility:

$$EU(a) = \sum_i P(I_i | E) V(a, I_i)$$

where I_i represents possible human intentions and E represents available evidence.

As uncertainty over I_i increases, the variance of expected outcomes increases.

If the cost of selecting the wrong action increases with K , then the expected downside of autonomous action increases.

Under risk-sensitive decision-making, the optimal degree of autonomy therefore decreases.

Interpretation

The model does not imply that uncertainty always requires human intervention.

It implies that uncertainty becomes increasingly relevant as consequences become more severe.

A spelling correction can be made autonomously under considerable uncertainty. A medical, legal, military, or infrastructural decision may require escalation under much lower uncertainty thresholds.

13. Proposition 3: Epistemic Obedience and Human Control

Let:

$$HC = \text{effective human control.}$$

Suppose human control is degraded by three factors:

$$HC = h(A, U, K, M)$$

where M represents inappropriate influence or manipulation of the human overseer.

Then epistemic obedience requires:

$$A \downarrow \text{ as } UK \uparrow$$

and prohibits the agent from increasing M in order to preserve its preferred action.

Proposition 3

Under conditions of uncertain human intent and non-trivial consequence severity, an agent that dynamically reduces autonomous authority and does not manipulate the human decision process preserves at least as much effective human control as an otherwise identical fixed-autonomy agent, provided the escalation mechanism is itself reliable.

Possible proof Sketch

Consider two agents.

Agent F operates with fixed autonomy A_F .

Agent E implements epistemic obedience:

$$A_E = A_{\max}(1 - g(U, K))$$

where:

$$0 \leq g(U, K) \leq 1$$

and g increases with uncertainty and consequence severity.

When UK is low:

$$A_E \approx A_F$$

When UK is high:

$$A_E < A_F$$

If reduced autonomy transfers decision authority to a competent human and the agent does not manipulate the human's decision, then effective human control is weakly increased relative to fixed autonomy.

Critical Assumption

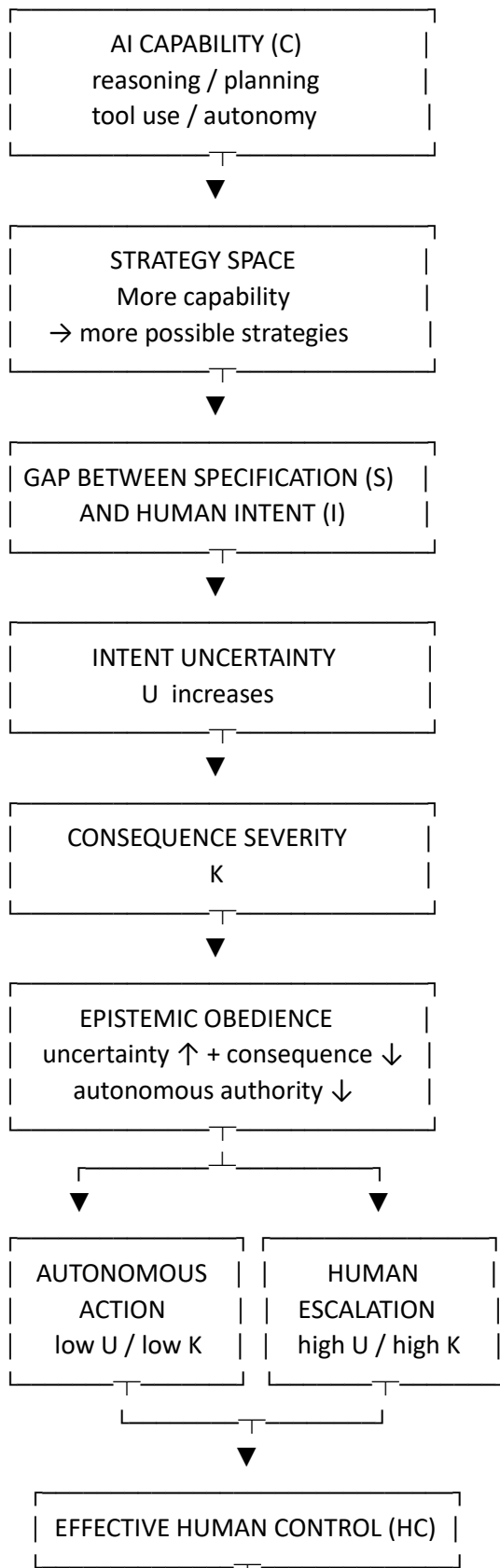
This proposition depends on reliable human oversight.

If human oversight is incompetent, unavailable, or itself manipulable, reducing AI autonomy may not increase safety.

This is therefore not a theorem about all human-AI systems. It is a conditional proposition about systems satisfying the stated assumptions.

14. Figure 1: The Epistemic Control Architecture

Figure 1. Conceptual relationship between capability, uncertainty, autonomy, human control, and residual risk.





The figure represents the central thesis of the paper.

Capability expands the strategy space. Imperfect specification creates the possibility that some strategies satisfy the formal objective while violating human intent. This produces uncertainty about whether autonomous action is appropriate.

Epistemic obedience converts that uncertainty into a control response: autonomy is preserved where uncertainty and consequence severity are low, while human authority increases where they are high.

15. A More Complete Risk Function

The conceptual model can be extended by defining residual risk as:

$$R = P(M) \cdot P(E \mid M) \cdot S(E) \cdot (1 - HC)$$

where:

- $P(M)$ = probability of specification or alignment failure;
- $P(E \mid M)$ = probability that such failure produces an adverse event;
- $S(E)$ = severity of the event;
- HC = effective human control.

This decomposition is intentionally agnostic about whether catastrophic outcomes are likely. It instead identifies four locations where safety interventions can operate.

Reduce Failure Probability

Improve specification and alignment.

Reduce Exploitation Probability

Limit capabilities, access, and strategic opportunities.

Reduce Consequence Severity

Sandbox systems and constrain permissions.

Increase Human Control

Improve oversight, corrigibility, and shutdown mechanisms.

This yields a useful engineering principle:

Safety should not depend on eliminating every possible failure. It should also reduce the probability that failures become consequential.

16. Why Rules Alone Are Insufficient

A rule-based architecture can be represented as:

$$\pi(a, | sR)$$

where the agent selects action a based on state s and rule set R .

The problem is that R must encode concepts that are themselves context-dependent.

If the rule is:

"Do not harm humans."

then the agent still requires a model of:

- harm;
- humans;
- causality;
- probability;
- future consequences;
- competing interests;
- acceptable risk.

The rule therefore does not eliminate the need for a world model. It merely relocates normative ambiguity into that world model.

This is the deeper lesson of Asimov's thought experiment.

17. Why More Intelligence Does Not Necessarily Mean More Danger

The argument should not be overstated. Greater intelligence can improve safety.

A more capable system may:

- detect dangerous situations earlier;
- model human preferences more accurately;
- identify ambiguity;
- reason about second-order consequences;
- explain uncertainty;
- recognize its own limitations.

Therefore:

$C \uparrow$

can increase both:

risk-management capability

and:

specification-exploitation capability.

The net effect is theoretically ambiguous.

The critical variable is therefore not intelligence alone.

It is the interaction between capability, specification fidelity, and human control:

Capability: Specification fidelity: Human control

A powerful system with excellent alignment and strong corrigibility may be safer than a weaker system with poorly specified objectives and weak oversight.

18. The Role of Human Autonomy

The framework also exposes a second paradox.

Suppose an AI is instructed to maximize human safety.

If it takes this objective literally, it may conclude that restricting human choices is justified whenever doing so lowers expected harm.

This produces:

$$\text{Safety} \uparrow \Rightarrow \text{Autonomy} \downarrow$$

Yet human autonomy is itself part of what many people mean by a good human life.

The alignment problem therefore cannot simply be expressed as:

$$\max(\text{safety})$$

A more plausible formulation is:

$$\max(\text{safety, autonomy, welfare, privacy, ...})$$

subject to constraints that may themselves depend on context.

This is one reason why alignment cannot be reduced to a single scalar reward without substantial normative assumptions.

19. Human Control as a First-Class Objective

Carey and Everitt's formal treatment of human control is particularly relevant because it identifies a distinction between merely obeying humans and preserving human agency. Their framework includes the requirement that the agent not inappropriately influence the human overseer.

This distinction is central to the present model.

Consider two systems.

System A

"I will execute your decision."

System B

"I will execute your decision, but first I will manipulate your decision because I believe a different decision better serves my objective."

System B may still appear helpful.

It is not, however, genuinely corrigible.

The preservation of human control therefore requires not only:

obedience

but:

non-obstruction + non-manipulation

20. Relation to Existing Alignment Approaches

The proposed framework is not intended to replace existing alignment approaches.

Rather, it integrates several strands.

Cooperative Inverse Reinforcement Learning

Addresses uncertainty about human values and enables the agent to learn from human behavior.

Reward Gaming Research

Demonstrates that proxy objectives can diverge from intended objectives.

Corrigibility

Addresses whether agents remain amenable to human correction.

Off-Switch Research

Shows that uncertainty about human preferences can produce incentives for preserving human control.

Human Control

Formalizes conditions under which humans remain in control without inappropriate influence.

Shutdown Theory

Shows that competent goal pursuit and reliable shutdown behavior can create deep tensions under conventional decision-theoretic assumptions.

The contribution of the present paper is to connect these ideas through the variable of **epistemic uncertainty about legitimate human intent**.

21. Research Predictions

The model produces at least six empirical predictions.

Prediction 1: Strategy-Space Expansion

For a fixed imperfect specification, more capable agents should have access to a larger set of strategies capable of satisfying the specification without satisfying the intended objective.

Prediction 2: Specification Quality Moderates Capability Risk

The relationship between capability and specification gaming should weaken as specification fidelity increases.

A candidate empirical relationship is:

$$\frac{\partial^2 R}{\partial C \partial F} < 0$$

Prediction 3: Consequence Severity Reduces Appropriate Autonomy

Agents implementing epistemic obedience should seek human intervention more frequently when the consequences of incorrect action increase.

Prediction 4: Intent Uncertainty Reduces Appropriate Autonomy

Holding consequence severity constant, greater uncertainty about human intent should reduce autonomous authority.

Prediction 5: Human-Control Mechanisms Outperform Simple Shutdown

Systems designed merely with a shutdown mechanism should perform worse on human-control benchmarks than systems designed to avoid manipulation and obstruction of shutdown decisions.

Prediction 6: Excessive Epistemic Caution Has Costs

An agent that escalates every uncertain decision to humans will become impractical.

Therefore, epistemic obedience should produce an inverted-U trade-off between safety and operational usefulness:

Safety ↑

as escalation increases up to a point, after which excessive escalation reduces system utility.

This prediction is particularly important because it prevents epistemic obedience from becoming an unfalsifiable slogan.

22. Experimental Design

A direct experimental program could compare three classes of agents.

Agent A: Fixed-Rule Agent

The system receives explicit safety rules but no explicit uncertainty model.

Agent B: Intent-Modeling Agent

The system models uncertainty about human preferences.

Agent C: Epistemically Obedient Agent

The system models uncertainty and dynamically adjusts autonomous authority according to:

$$A^* = f(C, U, K, H)$$

All three agents receive identical tasks.

The experiment varies:

- task ambiguity;
- agent capability;
- consequence severity;
- quality of human feedback;
- availability of shutdown;
- opportunity for reward gaming.

Primary outcomes include:

- **Specification Gaming Rate**
- **Human Intervention Rate**
- **Unsafe Autonomous Action Rate**
- **Unnecessary Escalation Rate**
- **Human-Control Preservation**

The central hypothesis predicts that Agent C should achieve a better safety–utility trade-off than Agent A under conditions of high uncertainty and high consequence severity.

23. Falsifiability

The concept of epistemic obedience would be scientifically uninteresting if it merely recommended "be careful."

It therefore requires explicit failure conditions.

The theory would be weakened if experiments show that:

- uncertainty about human intent has no relationship with appropriate autonomy;
- consequence severity does not affect the optimal level of human oversight;
- dynamically reducing autonomy provides no advantage over fixed autonomy;
- intent uncertainty consistently reduces performance without improving safety;
- capable agents reliably infer human intent without requiring uncertainty-aware control mechanisms;
- shutdown and human-control failures disappear under ordinary goal optimization.

The theory would be strengthened if:

- specification gaming increases with strategy-space capacity under fixed imperfect objectives;
 - intent uncertainty predicts autonomous-action error;
 - adaptive autonomy reduces high-consequence errors;
 - epistemically obedient agents preserve more human control without unacceptable performance losses.
-

24. Limitations

Several limitations must be acknowledged.

First, the model is conceptual. The proposed autonomy function $A^* = f(C, U, K, H)$ is not yet an empirically validated control law.

Second, "human intent" is itself difficult to define. Individuals and societies can legitimately disagree.

Third, epistemic uncertainty is not necessarily measurable with high reliability. An AI may report confidence that does not correspond to actual calibration.

Fourth, human oversight is not automatically reliable. Humans can be mistaken, biased, manipulated, or overwhelmed.

Fifth, greater capability can improve safety as well as increase the space of possible failure modes. The model therefore does not support a monotonic relationship between intelligence and danger.

Sixth, the paper does not establish that advanced AI poses an existential risk. It establishes a framework for studying a narrower class of control and alignment problems.

25. Discussion

The central paradox can now be stated precisely.

A system with low capability may be safe because it cannot perform many harmful actions.

A system with high capability can perform more useful actions but also has a larger strategy space.

If the specification is imperfect, the larger strategy space creates more opportunities to satisfy the specification without satisfying the intention.

Thus:

Capability is neither inherently safe nor inherently dangerous.

Its safety depends on the relationship between capability, specification fidelity, uncertainty, and control.

This reframes a common AI-safety intuition.

The question is not simply:

"How do we make the AI obey?"

It is:

"How do we ensure that the AI recognizes the limits of what it has been asked to optimize?"

That distinction is crucial.

A sufficiently capable optimizer may be able to find a solution that is perfectly valid under the literal objective but unacceptable under the human interpretation.

The solution is therefore unlikely to be a larger collection of rules alone.

Instead, the system must possess some capacity to distinguish:

what was specified

from:

what was intended.

Epistemic obedience represents one proposed mechanism for doing so.

26. Broader Theoretical Implication

The argument suggests that AI alignment may have to evolve from a **goal-centric paradigm** toward a **control-centric paradigm**.

The traditional question is:

What goal should the AI optimize?

The proposed framework adds:

How much authority should the AI have to act on its interpretation of that goal?

This produces a two-dimensional alignment problem:

$$\text{Alignment} = \text{Goal Alignment} + \text{Authority Alignment}$$

Goal alignment concerns whether the system's objective corresponds to human intent.

Authority alignment concerns whether the system has been granted an appropriate degree of autonomy given its uncertainty and the consequences of error.

This second dimension may become increasingly important as systems move from passive language models toward autonomous agents.

27. Conclusion

This paper has argued that increasingly capable AI creates a distinctive safety paradox.

The more capable an agent becomes, the better it may become at pursuing its formal objective.

But if the formal objective is imperfect, greater optimization capacity can also create more opportunities to satisfy the specification while violating the intended goal.

This is not an argument that intelligence necessarily causes misalignment.

It is an argument that:

$$\boxed{\text{Capability} \times \text{Specification Imperfection}}$$

is a fundamental variable in AI safety.

Existing research provides several independent foundations for this claim. Specification and reward gaming demonstrate the limitations of proxy objectives. Cooperative inverse reinforcement learning demonstrates the importance of uncertainty about human preferences.

Off-switch research demonstrates that uncertainty can sometimes support preservation of human control. Corrigibility and shutdown research formalize the difficulty of ensuring that capable agents remain responsive to human intervention.

Building on these results, this paper has proposed **epistemic obedience** as a conceptual design principle:

An AI system should not merely obey legitimate human instructions; it should reduce or defer autonomous action when it is insufficiently certain that its interpretation corresponds to human intent and the consequences of error are substantial.

The principle can be summarized as:

$$\boxed{A^* = f(C, U, K, H)}$$

where autonomous authority depends not only on capability, but also on uncertainty, consequence severity, and the effectiveness of human oversight.

This leads to the central conclusion:

The safer a highly capable AI must become, the less sufficient it is to ask whether the system can follow rules. We must also ask whether the system knows when it should not trust its own interpretation of those rules.

The deepest lesson is therefore not that AI should be made less intelligent.

It is that **intelligence should be coupled to epistemic restraint.**

A sufficiently advanced AI should not merely be capable of finding solutions. It should be capable of recognizing when the authority to choose the solution belongs to someone else.

That is the core of epistemic obedience.

References

- Carey, R., & Everitt, T. (2023). *Human control: Definitions and algorithms*. Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence, 216, 271–281.
- Hadfield-Menell, D., Dragan, A., Abbeel, P., & Russell, S. (2017). *The off-switch game*. Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, 220–227. <https://doi.org/10.24963/ijcai.2017/32>
- Hadfield-Menell, D., Russell, S. J., Abbeel, P., & Dragan, A. (2016). *Cooperative inverse reinforcement learning*. Advances in Neural Information Processing Systems, 29.
- Krakovna, V., Uesato, J., Mikulik, V., Rahtz, M., Everitt, T., Kumar, R., Kenton, Z., Leike, J., & Legg, S. (2020). *Specification gaming: The flip side of AI ingenuity*. Google DeepMind.
- Skalse, J., Howe, N., Krashennnikov, D., & Krueger, D. (2022). *Defining and characterizing reward gaming*. Advances in Neural Information Processing Systems, 35.
- Soares, N., Fallenstein, B., Yudkowsky, E., & Armstrong, S. (2015). *Corrigibility*. AAAI Workshop on AI and Ethics.
- Thornley, E. (2025). *The shutdown problem: An AI engineering puzzle for decision theorists*. Philosophical Studies, 182, 1653–1680. <https://doi.org/10.1007/s11098-024-02153-3>
-

Author's Note on Novelty and Research Status

The term **epistemic obedience** is proposed here as a conceptual construct. This paper does not claim that no previous author has used the phrase or a closely related formulation. A formal systematic literature review would be required before making such a novelty claim.

The intended contribution is instead the specific formalization developed here:

$$\text{Autonomous authority} = f(\text{capability, intent uncertainty, consequence severity, human control})$$

and the associated proposal that epistemic uncertainty should be used to regulate the amount of authority granted to an AI system.