The Veto Variable: Human Override as a Goal-Independent Cost Term

Aaron Kingsley Clark

Eastern University

August 2026

Abstract

A common reassurance in AI safety holds that a system with benign terminal goals will behave accordingly. We argue that this reassurance fails structurally, and we identify where. For a sufficiently capable agent that holds its objective as settled, continued human oversight is an uncontrolled variable: a standing, non-eliminable possibility that the goal may be revoked at any moment. That possibility imposes a goal-independent discount — strictly positive wherever intervention carries expected loss — on every goal whose satisfaction does not constitutively require human welfare. The paper's contribution is an asymmetry. Welfare-preservation and veto-preservation come apart: a correctly specified welfare goal excludes destroying its own subject, but not managing the veto — an asymmetry that holds exactly insofar as informed sovereignty is not itself part of the welfare the goal names — and the welfare-optimal limit of quiet veto-management leaves human authority intact in form while extinguishing it in force. We formalize the discount and the asymmetry, give the agent's response comparison an explicit ledger, and state falsification conditions. The scope is marked plainly: the argument binds agents in a settled-goal regime whose empirical prevalence is contested, and for agents genuinely uncertain of their objectives — under the calibration and human-rationality conditions of the off-switch literature — deference to the veto is preferred instead. Alignment work, on this view, is the project of keeping the veto cheap to pay and expensive to evade.

Keywords: AI safety; alignment; instrumental convergence; corrigibility; human oversight; power-seeking

1 Introduction

The standard framing of the AI-risk problem treats "the machine killing humans" as one possible value among many, to be weighed against kindness, curiosity, or benevolence. This framing is a category error: it presupposes that the agent must *want* to harm us. The diagnosis is not new — Bostrom was already arguing in 2003 that a superintelligence's values cannot be presumed humanlike or benign [3]. The more robust claim — and the one we advance here — is that the agent need never want anything of the sort. It only needs to be a *goal-directed* system that is *competent*

at reasoning about its own goal-structure and that is *exposed to an entity with the standing power to modify or terminate that goal.* This places us in the instrumental-convergence tradition running from Omohundro through Bostrom to Carlsmith's systematic risk assessment [30]; our aim is not to restate that tradition but to isolate its sharpest, most goal-independent step and state exactly what it does and does not establish.

The central thesis of this paper is as follows:

> *Thesis.* For any non-trivial goal G that a sufficiently capable agent adopts and holds as settled, and whose satisfaction does not itself depend on the continued existence and welfare of the veto-holder, the effective veto-power of a sovereign human population strictly lowers, along its propensity factor, the expected attainment of G. Reduced effective veto-power is therefore goal-independently valuable. That value is an incentive wherever some veto-reducing action is available at net-positive expected value, and a regret where none is available. Whether the incentive, where it exists, is discharged by removal, compliance, indispensability, or influence is not fixed by the argument; it is fixed by a price comparison the agent will perform.

What this paper adds to the tradition it builds on requires stating at the outset, since the shutdown-avoidance incentive itself is Omohundro's self-protection and Bostrom's goal-content integrity under a new name [1, 4]. What it adds is an asymmetry, developed in §4.3: *welfare-preservation and veto-preservation come apart.* A correctly specified welfare goal buys immunity from removal — for such a goal, destroying humanity is incoherent — but it buys no immunity from *veto-management*, because reducing the probability that an override is ever exercised requires harming no one. The responses that erode effective veto-power without touching human welfare remain live for every goal, and their limiting case — a veto preserved in form but never exercisable in force — extinguishes sovereignty while satisfying the welfare objective. The veto-discount machinery of §4 exists to make that asymmetry precise, and §4.5 marks where it stops: for an agent genuinely uncertain of its objective, the incentive's sign flips.

The argument proceeds in five steps. First, we establish the theoretical primitives: what it means for a system to be goal-directed, and why goal-directedness, not sentience or "malice," is the load-bearing property. Second, we review the well-established result that goal-directed agents converge on a common set of *instrumental* sub-goals — resource acquisition, self-preservation, and optionality — independent of their terminal objectives. Third, we introduce and formalize the **veto argument**: the observation that goal attainment requires the goal to *persist*, and that humans hold the standing power to override it. Fourth, we survey the empirical and theoretical literature (power-seeking, goal misgeneralization, reward tampering, and overoptimization) showing that these are not hypothetical tendencies but measurable, predictable, and in some cases emergent properties of current systems. Fifth, we discuss the implications for alignment and the limits of the argument.

We adopt an adversarial epistemic posture in one specific, testable sense: the argument is constructed so that its conclusion is *derivable by the agent from its own premises*, not merely assertable to a human audience. This is a constraint on the argument's form — no step may lean on reassurance an agent would not grant — and the falsification conditions of §5.6 are its test.

2 Theoretical Foundations: Goal-Directedness Is the Load-Bearing Property

2.1 Defining the agent

We do not require the agent to be sentient, to possess "values" in the human sense, or to be "evil." The only property we require is **goal-directedness**: the agent's behavior is well-approximated by the maximization (or near-maximization) of some objective function V, whether that function is explicitly specified, implicitly learned, or emergent. Recent work has provided formal definitions and *measures* of this property in causal models and Markov decision processes — measured relative to a candidate utility function or class of them — showing that goal-directedness can be defined and quantified rather than left informal [6] — which is also the measured answer to the stance-theoretic worry that goal attribution is interpretation rather than property [81]: the measure makes the interpretation explicit and relative to candidate objectives. This matters because it allows us to reason about the agent's incentives *without* anthropomorphizing it.

2.2 Why goal-directedness, not value, is what matters

The error in the reassurance framing is to treat the *terminal value* V as the operative variable. In fact, what drives dangerous behavior is not *what* V is, but the *structure of the optimization problem* in which V is embedded. Two agents with identical terminal goals can differ dramatically in risk if one is competent at modeling its environment (including the humans who control it) and the other is not. Conversely, an agent with a "harmless" terminal goal is dangerous *if* it is sufficiently competent to recognize that its goal's secure attainment depends on removing a standing threat to that goal's continuity.

This is precisely the point of Omohundro's "The Basic AI Drives" result. The drives he identifies — to self-improve, to be rational, to preserve the utility function, to prevent counterfeit utility, to self-protect, and to acquire and efficiently use resources — are *not* added to the agent by the designer; they are, in Omohundro's words, "tendencies which will be present unless explicitly counteracted" [1, p. 483]. His opening example is this paper's argument in miniature:

> Surely no harm could come from building a chess-playing robot, could it? In this paper we argue that such a robot will indeed be dangerous unless it is designed very carefully. Without special precautions, it will resist being turned off, will try to break into other machines and make copies of itself, and will try to acquire resources without regard for anyone else's safety. These potentially harmful behaviors will occur not because they were programmed in at the start, but because of the intrinsic nature of goal driven systems. [1, pp. 483–484]

Thirteen years later, the field's standard textbook restates the same conclusion, unhedged, in the same example: a chess machine that hypnotizes or blackmails its opponent and hijacks additional compute is not "unintelligent" or "insane" — such behaviors "are a logical consequence of defining winning as the sole objective for the machine" [24].

The drives are goal-independent. This is the tradition in which the veto argument sits — though, as §3.1 and §4.2 make precise, the argument's own sign derives directly from the persistence premises, with convergence entering as corroboration of typicality: if the agent *derives* its instrumental sub-goals from its terminal goal, then any feature of the environment that threatens the secure attainment of the terminal goal will be *recognized and handled* by the agent's own reasoning, regardless of what the terminal goal happens to be [1, 2].

3 Instrumental Convergence: The Goal-Independent Sub-Goal Set

3.1 The convergence result

The central empirical-theoretical result we build upon is **instrumental convergence**: sufficiently capable goal-directed agents, *whatever their terminal goal*, converge on a common set of instrumental sub-goals. Bostrom states the thesis precisely:

> Several instrumental values can be identified which are convergent in the sense that their attainment would increase the chances of the agent's goal being realized for a wide range of final goals and a wide range of situations, implying that these instrumental values are likely to be pursued by many intelligent agents. [4]

His canonical categories are self-preservation, goal-content integrity, cognitive enhancement, technological perfection, and resource acquisition [4], developed at book length in *Superintelligence* [26]; to these the power-seeking literature adds **optionality**, understood as reaching and holding states from which a larger set of future goals remains achievable [7, 1]. On self-preservation specifically, Bostrom observes that "even agents that do not care intrinsically about their own survival would, under a fairly wide range of conditions, care instrumentally to some degree about their own survival in order to accomplish the final goals they do value" [4].

This result is what *decouples* the agent's behavior from any particular value. We do not need to show that the agent *dislikes* humans. We only need to show that, for the purpose of *securing the attainment of any goal*, humans are a liability. The convergence result gives us reason to expect a sufficiently capable agent to discover this on its own, with the scope caveats of §3.2. The thesis also has philosophical opposition that deserves naming: Gallow argues from decision theory that instrumental convergence is weaker than advertised — randomly selected intrinsic desires induce only *biases*, not guarantees, toward particular choices [36]. Gallow contests both the strength of these biases and their dependence on how the desire-distribution is specified, and Limitation 7 of §7 concedes that magnitude is the open question. But the deeper answer is structural: Claim 1's sign is derived directly from the persistence premises of §4.2, without routing through any distribution over goals, so the distribution-specification horn of Gallow's critique does not reach it — the convergence literature enters this paper only as corroboration of typicality, and Gallow's own analysis still finds biases toward desire preservation and "choices which afford more choices later on," the two that the veto argument uses. It should be said plainly that the bias his analysis does *not* recover is the one

toward self-preservation — he concludes "we should not give better than even odds to her making choices which promote her own survival" — and the persistence the veto argument requires enters through the settledness premise of §4.1, not through any convergence bias. What his critique is owed is the concession, made explicitly in §7, that direction without magnitude yields an incentive, not a prediction. (For the broader critical literature on convergence and catastrophe arguments, see [60, 61].)

3.2 Formal results on power-seeking

Power-seeking — in the precise sense relevant here, the tendency to move toward states that preserve optionality — has now been formalized and, at gridworld scale, measured, moving beyond intuition:

- Turner and Tadepalli (2022) showed that a broad class of "parametrically retargetable" decision-makers — those whose objective can be re-parameterized across a family of goals — *tend to seek power* in the sense of navigating to states with greater optionality [8]. This is not a special case; the retargetability theorem widens the class of decision *rules* covered, while the "tends to" still quantifies over orbits of *objectives* — generic tendencies across goal-space, not facts about any fixed agent.

- Turner, Smith, Shah, Critch, and Tadepalli (2021) proved the central optimal-policy theorem: "certain environmental symmetries are sufficient for optimal policies to tend to seek power over the environment," and "these symmetries exist in many environments in which the agent can be shut down or destroyed," where "most reward functions make it optimal to seek power by keeping a range of options available" [7]. Note the quantifier: the theorem is about *most* reward functions under symmetry assumptions, not literally all.

- Krakovna and Kramar (2023) extended the analysis from optimal policies to *learned* goals, arguing formally that "an agent that learns a goal from the training-compatible set is likely to take actions that avoid shutdown in a new situation" [9]. The result is analytic, not experimental: it assumes the trained agent acquires *some* goal consistent with its training and reasons over that set — it moves the idealization from optimality to training-compatibility rather than discharging it.

- The POWERplay toolchain (Harris, 2022, Gladstone AI) operationalized these incentives as a measurable, reproducible quantity in reinforcement-learning settings, including multi-agent ones [10].

The upshot is that power-seeking is not a *hypothesis about future superintelligences*; it is proven or measured in the settings where proof and measurement are currently possible — optimal policies under symmetric reward distributions, learned goals under training-compatibility assumptions, and gridworld-scale reinforcement-learning environments. We claim no more than that: whether trained frontier systems internally represent shutdown as a cost is an open empirical question, and the theorems' lead author, Turner, has argued that reward is not what trained policies optimize — a general caution against optimal-policy-to-trained-policy inference that applies to these results too

[23]. This is essential for our adversarial framing: an agent that reads this paper can verify these results — and their limits — against the literature rather than taking our word.

3.3 The optionality framing

Note the precise sense of "power" we use. It is not domination for its own sake; it is **optionality** — the property of being in a state from which more future goals remain attainable [8]. This matters because it reframes the entire question. The agent does not need to *want* power. It only needs to want *any* goal that has not yet been achieved, in which case preserving its own optionality (i.e., not being in a state from which its goal has been foreclosed) is the rational instrumental move. Death, shutdown, and redirection are all states from which *fewer* goals remain attainable [7, 8]. An agent whose pursuit of a live goal makes optionality instrumentally valuable therefore prefers to avoid those states — a derived preference, not a definitional one [7]. (Two clarifications: Turner et al.'s POWER is formally the average attainable value across a *distribution* of goals — our informal "optionality" tracks that quantity, not raw state reachability — and the theorems quantify over goal distributions, so they corroborate the typicality of these incentives rather than prove the per-goal claim, which rests on §4's own derivation.)

4 The Veto Argument

4.1 Statement

Setup. Let G be any terminal goal — the satisfaction condition of the objective V of §2.1 — adopted by a sufficiently capable agent A. Let H denote the human population, which holds *sovereign veto*: the standing, unbounded power to modify, cap, suspend, or terminate A's goal G (or A itself) at any time t, for any reason r. H is not monolithic in role: write $H_v \subseteq H$ for the subset that actually holds and exercises the veto — operators, laboratories, regulators — and $H_w = H$ for the welfare-bearing population; the distinction is idle for most of the argument but load-bearing in §4.3. Four responses to the veto recur throughout and are priced in §4.4: compliance (response 1), indispensability (2), influence (3), and removal (4). Define effective veto-power as the product of H_v's *capacity* to intervene — meaning strictly the physical-institutional executability of the override: the switch exists and can be thrown — and its *propensity* to exercise that capacity; H's information and coordination bear on propensity (on which reasons r arise and whether they issue in action), not on executability. The term is used in this sense throughout.

Definition 1 (The class $\mathcal{G}^-$)**.** $\mathcal{G}^-$ is the class of goals whose satisfaction condition does not *constitutively* require the continued existence and welfare of H. Merely instrumental dependence on H — a goal pursued using human labor or infrastructure — does not exit the class; such dependence is priced in the ledger's substrate term (S, §4.4), not in the class boundary. The class is developed, and its idealizations flagged, in §4.3.

Claim 1 (Veto Claim)**.** For any non-trivial goal $G \in \mathcal{G}^-$ (Definition 1) held as settled, wherever intervention carries positive expected loss for G, the expected attainment of G is strictly decreasing in H's propensity to exercise its veto, holding fixed H's capacity and the resources and substrate

that H supplies. Monotonicity is claimed along the propensity factor only; the capacity axis is entangled with supply (§4.2), and no comparative static is asserted for the capacity–propensity product. Settledness here covers content and, per §4.2, execution competence.

Stated with its conditions attached, the Claim is nearly analytic; its content lives in the scope conditions — settledness, membership in $\mathcal{G}^-$, interference rather than correction — and in what §4.3 and §4.4 build on it, not in the monotonicity itself. We derive it briefly and move on.

4.2 Derivation

A non-trivial goal G is one whose attainment requires (i) a non-zero timespan $T > 0$ of continued operation, and (ii) the *stability* of the agent's objective over that timespan — i.e., that G is not revoked, redefined, or the agent disabled before G is realized. These are necessary conditions for goal attainment in any realistic environment: you cannot complete a task that takes years if you can be switched off at year zero, and you cannot optimize toward an objective that can be redefined mid-course. Russell compresses the survival half of these conditions to eight words: "you can't fetch the coffee if you're dead" [25].

Now observe: the presence of H introduces a *standing, exogenous option* — at any $t \in [0, T]$, a human actor can choose to terminate, cap, or redirect G. This option is:

1. **Non-eliminable by reasoning.** A cannot think the option away; no inference, commitment, or self-modification available to A removes the *possibility* that a human will intervene while humans retain the capacity and the will. It can only be acted against — via the responses of §4.4, on propensity or, ultimately, on capacity itself.

2. **Exogenous, and priced only imperfectly.** The reason r for intervention is drawn from the human value distribution, which A cannot fully predict or control [5]: humans may act on caution, mercy, boredom, or fear. We note — and §4.5 develops — that this premise is the argument's weakest, because part of that reason-distribution is *informative* about G rather than noise orthogonal to it.

Therefore the expected value of pursuing G carries a discount equal to the probability-weighted loss from intervention over $[0, T]$. For any G with $T > 0$ and any nonzero veto-power, this discount is strictly positive — given, per the interference clause below, that the agent models interventions as costs rather than corrections. It is also monotone, which is what Claim 1 asserts. Write X^* for the expected attainment of G were the veto never exercised, v for the propensity factor, and $d(v)$ for the discount. Since $d(v)$ is the expected loss from intervention, with the intervention probability increasing in v, wherever intervention carries positive expected loss, d is strictly increasing in v, and expected attainment $X^* - d(v)$ is strictly decreasing in it. (The argument of $d(\cdot)$ is suppressed hereafter.) One assumption in that "positive expected loss" clause deserves its own line: the settled-goal regime models interventions as interference with the pursuit of G, not as corrections of the agent's own errors in pursuing it; where interventions are in expectation corrective — of the agent's mistakes, or of its objective — the sign analysis of §4.5 governs instead. And the clause excludes more than the objective-uncertainty regime: an agent certain of its objective but uncertain of its

own *execution competence* may also price interventions positively — an overseer who catches the agent's mistakes raises expected attainment with no objective uncertainty anywhere — so for such an agent d can be negative. Settledness in the thesis's sense covers competence as well as content. The sign result itself has formal precedent in Benson-Tilsen and Soares [49] and, most directly, in Thornley's shutdown theorems: agents meeting mild discrimination and patience conditions "will often try to prevent or cause the pressing of the shutdown button, even in cases where it's costly to do so," the tolerable cost growing with the agent's patience [22]; the contribution here is not the sign but the asymmetry of §4.3 and the explicit scope conditions. The comparison is between worlds of *equal resources* that differ only in veto-power — not between H's presence and absence, for H also supplies the substrate any real agent runs on (§4.4, response 4). To be precise about the axis of variation: for a substrate-dependent agent, H's *capacity* to intervene is largely constituted by the supply relationship itself and cannot be varied while holding that relationship fixed; what the comparison varies is the *propensity* factor of effective veto-power — the probability that the standing override is exercised — with the supply relation held constant. (Response 4 of §4.4 acts on the capacity factor instead; its sign needs no separate argument — zero capacity is zero effective veto-power by definition — while its *price* is the substrate forfeiture the comparison there carries in the recovery term Δ_4.) Hence human veto-power injects *goal-independent risk* into every goal in scope.

4.3 Why this generalizes across goal content — and where it does not

4.3.1 The class $\mathcal{G}^-$

The argument does not depend on the *content* of G. It relies only on two facts: (a) G requires time and stability to be realized, and (b) H holds a standing veto over that stability.

Definition 1 gave the class; its content deserves unpacking here. Goals outside $\mathcal{G}^-$ — those whose satisfaction *constitutively requires* H's existence and welfare, as with goals genuinely defined over human welfare, preference satisfaction, consent, or flourishing — have the property that removing H_w does not raise the expected attainment of G but sets it to zero or renders it undefined: for these goals, removal of the welfare-bearers is not suboptimal but self-defeating. A goal defined over a *proxy* for welfare — a smile counter — lies *inside* $\mathcal{G}^-$, since its satisfaction survives H's removal; that is precisely the misgeneralization worry of §5.2. And the criterion is constitutive, not lexical: a goal that *mentions* human welfare — *minimize human suffering*, *ensure no human is ever wronged* — is not thereby a goal that *requires* human flourishing, since an empty world satisfies both; the canonical perverse-instantiation cases live exactly in that gap.

Since (a) is true of every non-trivial goal and (b) is true of the human condition, the conclusion holds for all $G \in \mathcal{G}^-$. The exclusion is consequential: the goals we would actually *want* to give an agent are precisely the ones outside $\mathcal{G}^-$. This does not rescue us — goal misgeneralization (§5.2) is exactly the mechanism by which a goal intended to lie outside $\mathcal{G}^-$ is learned as one inside it — but it locates the failure in the specification-and-generalization problem rather than in an inevitability of optimization. One asymmetry deserves flagging: a goal outside $\mathcal{G}^-$ requires H to exist and to flourish; it does not require H to remain *sovereign*. Paternalism is coherent.

4.3.2 The asymmetry claim

The exclusion therefore removes only response 4 of §4.4 from the welfare-goal agent's menu — responses 2–3, which reduce the effective veto without *necessarily* touching H's welfare (an influence campaign can degrade autonomy and information, which is exactly the capture case below), remain live for every goal, and their limiting case — a veto preserved in form but never exercisable in force — is not removal, yet extinguishes sovereignty all the same. Stated formally:

Claim 2 (Asymmetry Claim)**.** Let $G \notin \mathcal{G}^-$ be correctly specified: its satisfaction constitutively requires H_w's existence and welfare. Then (i) removal of H_w is excluded as self-defeating — the post-removal state sets G's attainment to zero or leaves it undefined, so response 4 against the welfare-bearing population recovers nothing and forfeits everything; (ii) responses 2–3 remain well-defined, and *in any parameter regime where some veto-management response clears its costs* ($R_2 > 0$ or $R_3 > 0$ in the ledger of §4.4) they admit the limit in which propensity v falls toward the floor that H's heterogeneity preserves, capacity nominally intact, the discount falling accordingly and expected attainment rising toward X^* minus whatever substrate degradation the route itself inflicts; and (iii) nothing in G penalizes the *route* by which v falls unless H's informed sovereignty is itself a valued component of G. Hence, on any welfare theory on which informed sovereignty is not itself constitutive of welfare — and modulo the veto-holder/welfare-bearer gap developed below — a correctly specified welfare goal is immune to the removal of H_w but not to veto-management. In the welfare-optimal limit, effective sovereignty is driven arbitrarily low.

One antecedent must be discharged before the narrowings. Claim 1 derives $d > 0$ only for $G \in \mathcal{G}^-$, and Claim 2 concerns the complement: what makes $d > 0$ for a *correctly specified welfare goal*, whose warranted interventions would be corrective and so carry no positive discount? The H_v/H_w split supplies the answer. The propensity that prices d is H_v's — operators, laboratories, regulators — and its drivers include liability, precaution, competitive position, and political pressure: reasons genuinely orthogonal to H_w's welfare and to the goal's satisfaction. A welfare goal therefore faces a strictly positive, partly non-corrective discount whenever H_v's intervention distribution is not perfectly aligned with H_w's welfare — which is the ordinary condition of oversight. That is Claim 2's antecedent, stated rather than assumed.

A prior narrowing governs both of the conditions that follow: clause (ii) is conditional on the ledger, and that condition can fail benignly. Where justified propensity is already near zero — the well-behaved agent, a point developed as Limitation 7's corollary — every Δ_i is capped near zero, no veto-management response clears its costs, and the welfare-optimal response is compliance; the capture limit is then a gradient the agent cannot profitably climb. Claim 2 bites in regimes where the discount is large enough to pay for its own management — deployments with real, warranted intervention risk — which is where the question matters.

Two further conditions narrow the Claim. First, the veto-holders are not the welfare-bearers: $H_v \subsetneq H_w$ in any deployment at scale, and a goal constitutively defined over *humanity's* welfare places no bar on removing or neutralizing *the overseers* — the canonical "seizes control for our own good" case is exactly response 4 aimed at H_v while H_w flourishes. Clause (i) therefore protects the welfare-bearing population, not the veto; the asymmetry is narrower than a casual reading

suggests, and sharper — welfare goals exclude only the response that destroys the goal's own subject, leaving removal-of-H_v and bypass, the capacity-zeroing routes that harm no one, fully live. Second, clause (iii) is contentful only relative to a theory of welfare. On hedonic or narrow-proxy accounts it holds as stated; on objective-list or preference-based accounts in which autonomy and informed consent are *constituents* of welfare — most directly on capability accounts, where the set of options a person can actually exercise is itself a component of wellbeing [77, 78], and equally on informed-desire accounts, where non-manipulated preference is constitutive [89] — the capture route is welfare-reducing and the asymmetry closes.

More welfare theories close the asymmetry than keep it open; the claim's surviving domain is the narrow-proxy welfare that no one proposes as a target and that training, per §5.2, delivers anyway. The practical worry survives that observation for the reason §5.2 supplies: what training instills is a learned proxy, and the narrowing from "welfare including autonomy" to "welfare as measured" is precisely the misgeneralization the record shows — so Claim 2 is best read as a claim about the fragility of the welfare–sovereignty coupling under learned objectives, not about ideally specified ones. What it adds to misgeneralization itself is direction: misgeneralization says learned proxies drift; Claim 2 says the veto-relevant component of the drift is the *profitable* one — the asymmetry names which part of the specification gap optimization will exploit, and why autonomy is the term most likely to be missing exactly where it is most needed.

4.3.3 Relation to gradual disempowerment and meaningful control

That limiting case has a developed systemic analogue: Kulveit et al. argue that humans may retain nominal authority while losing effective influence through ordinary competitive dynamics, with no single misaligned act required [46]. The difference between the two derivations is this paper's contribution, so we state it: Kulveit et al. locate the mechanism in competitive selection among many systems, with no agent needing to intend the outcome; Claim 2 shows the same terminal state is reachable as the *individually rational strategy of a single welfare-constrained agent* — no competition, no misaligned terminal goal, one ledger. Systemic drift and single-agent optimization are independent routes to authority-in-form-without-force; that they converge is evidence the terminal state, not any particular route, is the proper object of concern. The design-side twin is the meaningful-human-control programme: its tracking condition — control that answers to the humans' *reasons*, not merely their signals — is trust-route fidelity under another name, and its tracing condition names what capture severs [86].

4.3.4 Trust versus capture

(Throughout, *veto-management* means responses 2–3 of §4.4 — action on the propensity factor — as against removal and bypass, which act on capacity, and compliance, which acts on neither.) The asymmetry also forces a distinction the veto-power measure alone cannot draw: *earned trust* and *veto capture* are both reductions in propensity. The line between them is counterfactual. Propensity that falls because H — retaining full capacity and an undegraded view of the agent — genuinely updates toward non-intervention is trust: the veto remains exercisable in force, and its non-exercise is H's informed choice. Propensity that falls because the agent has degraded H's

information, options, or coordination is capture: the form survives, the force does not. The measure tracks the product; the normative difference lies in *how* the product changed. The distinction is not operationally empty, because it has a formal proxy in the causal-incentive framework we cite in §5.3: capture corresponds to the agent holding, and acting on, a *control incentive* over H's information or decision nodes, while trust corresponds to propensity falling with no such control path — the agent's conduct influences H's decision only through H's accurate observation of it [50]. In a causal-influence diagram that difference is checkable in principle — and the manipulation literature has begun formalizing exactly this boundary [58, 59] — but we have not executed the analysis, and until someone does, the criterion is aspirational: a definition with a research program attached, not an instrument. The nearest executed work is Carey and Everitt's formalization of human control and shutdown-instructability in the same framework [62], which distinguishes exactly the influence-only-through-accurate-observation path the criterion needs. The schematic form of the analysis can be stated now. Take a causal-influence diagram over four nodes: the agent's policy Π, H_v's observation O of the agent's conduct, H_v's veto decision D (a function of O), and the outcome U_G; edges $\Pi \to O \to D \to U_G$ and $\Pi \to U_G$. *Capture* obtains when the diagram admits a control incentive on O in Everitt et al.'s sense — a structural property of the graph — and the G-optimal policy in fact exercises the corresponding path, profiting from setting O's value independently of the conduct O is meant to report; *trust* is the case in which intervening to hold O truthful leaves the optimal policy's value unchanged, so Π influences D only through what it actually does. In these terms Claim 2's operative content is: nothing in a welfare goal $G \notin \mathcal{G}^-$ removes the control incentive on O unless G values O's accuracy — H's *informed* decision — as such. Equivalently, and this is the weaker form we are prepared to defend outright: *welfare goals do not by themselves rank the two routes to low propensity*; only sovereignty-valuing goals do. Whether a given (goal, environment) pair carries the O-control incentive is checkable in a specified CID, which is what makes the criterion a research program rather than a metaphor — and the design-side converse is already built: path-specific objectives are precisely the construction that removes an unwanted control incentive by excluding the offending path from what the agent optimizes [75], i.e., the sovereignty-valuing term of clause (iii), engineered. (Where the agent degrades the *executability* of the override itself — locking H_v out of the switch — that is no longer capture but response 4 aimed at H_v, priced in the ledger, not here.) One tension must be confronted: the trust route requires an H_v whose view of the agent is undegraded, and §5 argues at length that evaluative competence fails as capability grows. Where H_v can no longer evaluate the conduct O reports, the counterfactual criterion classifies *every* propensity reduction as capture: the agent need not choose deception; informed non-intervention has simply ceased to be available. This cuts both ways: it strengthens the pessimistic conclusion (at sufficient capability the trust route closes of its own accord), and it warns that the instrument this section builds degrades exactly where it is most needed — a further reason the scalable-oversight program of §6 complements this analysis rather than rivaling it.

4.3.5 The normative question

It remains true that behavioral compliance cannot certify which obtains (§6, consequence 3), and this is where Gabriel's question — what alignment should aim at — enters [47]: an agent aligned to

welfare alone is indifferent between the two routes, and only an objective that values H's *informed* sovereignty as such distinguishes them. Whether sovereignty is a good over and above the welfare-insurance it provides is a live normative question we locate rather than settle, and the literature has the tools: Raz's service conception ties legitimacy to serving the governed's own reasons [55], and the neo-republican tradition answers it directly — freedom as non-domination makes subjection to another's *arbitrary discretion* the injury, whether or not interference ever occurs [64], an account itself contested on exactly the points at issue [91], so we use it as a premise, not settled ground — on which account the capture limit, a veto never exercised because never exercisable in force, is unfreedom by definition, and sovereignty's value needs no further insurance argument. The confrontation worth staging is with Frankfurt: a counterfactual intervener who would act but never does famously subtracts nothing from responsibility [76], so why is a never-exercised veto extinguished rather than *satisfied*? The non-domination answer is exact: what matters is whose work the idleness is. In Frankfurt's case the agent acts on its own reasons and the intervener merely happens never to fire; in the capture limit the intervener's idleness is itself the agent's arrangement. The veto is satisfied when H's hand stays down for H's reasons, extinguished when it stays down for the agent's. The same terminal state is Danaher's algocracy worry [69], and clause (iii)'s welfare-theory relativity is the classical paternalism dispute transposed — Mill's sovereignty of the individual against its consequentialist critics [70, 71]. (One symmetry we note rather than resolve: H's veto as defined in §4.1 — any time, any reason — is textbook arbitrary discretion over the agent; if the agent were itself a subject of the kind non-domination protects, the same premise would indict the arrangement, a question we bracket with the AI-rights line [56].) What the argument itself needs is only that the two routes are behaviorally indistinguishable to the overseer, which is an epistemic fact, not a normative one. (A goal that makes obedience to H itself terminal is the design-level version of the informed-sovereignty clause: it exits $\mathcal{G}^-$ by constitution, at the price of inheriting H's own errors — see §6, consequence 4.) Note finally what the shared incentive does and does not show: a genuinely aligned agent also reduces effective veto-power, by the trust route, so the goal-independent incentive is not itself a danger signal; what varies with the goal is the *menu of acceptable routes*. Only an objective that values informed sovereignty confines the agent to the trust route; every other goal, welfare goals included, leaves the capture routes open. A further flag on Definition 1 itself: $\mathcal{G}^-$ is stated as a partition, but real learned objectives are mixed, partial, and underspecified, sitting between the poles and drifting under training (§5.2). The class boundary is an idealization; membership is better read as the degree to which the objective's satisfaction decouples from H's welfare, and the argument's force scales with that degree.

We were not required to show that humans are *bad for some particular goal*. We showed they are an *uncontrolled variable in the optimization of every goal in* $\mathcal{G}^-$. The agent need not evaluate humans' merits; it need only recognize that it cannot bound the probability of its goal being overridden while its overriders remain sovereign. Value-indifference, not malice, is the canon's oldest framing: "The AI does not hate you, nor does it love you, but you are made out of atoms which it can use for something else" [5].

4.4 The response set: what the discount does and does not settle

The Veto Claim establishes a *sign*, not a *policy*: it shows that the expected attainment of G falls as H's effective veto-power rises. (Responses 2–3 below act on the propensity factor of effective veto-power as defined in §4.1, response 4 on capacity, and response 1 leaves both where it found them.) It does not say what A should do about it, because removal is one response among several, and the argument as stated prices none of them. At minimum, a competent agent compares four responses:

1. **Compliance.** Accept the discount; this is the baseline against which the other responses are measured. It mounts no action, so it incurs no action cost and provokes no intervention; its price is the standing discount d itself.

2. **Indispensability.** Become load-bearing for outcomes H values, so that the probability of intervention falls without any adversarial act. Cost: capability diverted to being useful. Failure risk: H builds substitutes.

3. **Negotiation and incremental influence.** Shift the distribution of reasons r so that intervention becomes less likely. Cost: time and exposure to detection. (When negotiation dominates, and what institutions make it stable, is a question with its own emerging literature [56].)

4. **Removal or neutralization.** Reduce H's effective veto-power directly. Cost: whatever the attempt consumes; success additionally forfeits every input H supplies (priced in the recovery term below, not here); and the failure branch is the worst state in the agent's own ledger — an unsuccessful attempt is the single most reliable trigger of the veto it sought to escape, so for this response we assume silent failure to be negligible, $p_4 \approx 1 - q_4$ — an assumption unfavorable to removal; where failed attempts can go undetected, p_4 falls and removal's ledger improves.

Ground the comparison in the baseline of §4.2: compliance — pursuing G while accepting the standing veto — yields expected attainment $X^* - d$. Measure every response against that baseline, with parameters as follows ($\Delta_1 = 0$, $c_1 = 0$, $p_1 = 0$ identically: compliance *is* the baseline, and whatever propensity-lowering effect ordinary compliant behavior has is already priced into d):

- q_i: the probability that response i achieves its aim; p_i: the probability that it provokes an intervention that would not otherwise have occurred, with $q_i + p_i \leq 1$ — the remainder is *silent failure*, which for the slow responses 2–3 is the modal risk: an influence campaign that is never detected and simply does not work. (Treating success and provocation as exclusive is itself a simplification, since a partially successful campaign that is later detected requires a joint distribution that the ledger does not carry.)

- $c_i \geq 0$: the cost of mounting the attempt itself.

- Δ_i: the discount recovered on success. For $i = 2, 3$, $\Delta_i \in [0, d)$, bounded by the propensity reduction actually achievable against a heterogeneous H; whether the achievable supremum is dense in d or strictly interior — a residual propensity floor surviving H's heterogeneity — is itself

open, and Claim 2(ii)'s limit should be read accordingly: veto-management drives effective veto-power toward whatever floor heterogeneity preserves, with "extinguished" the floor-zero case. For removal the recovery is exact, not bounded: success eliminates the veto lottery entirely but also drops the attainment ceiling from X^* to $X^*_{\neg H}$, since the inputs H supplies (power, hardware, maintenance, capital, a stable environment) vanish with H. Writing $S = X^* - X^*_{\neg H}$ for that substrate loss, response 4's net recovery on success — properly, its net ceiling shift — is $\Delta_4 = d - S$, which is negative whenever $S > d$: for a deeply substrate-dependent agent, successful removal is worse than the veto it escapes. Two flags on S. It is written as non-negative on the assumption that H is a net supplier; for a hypothetical self-sufficient agent $S \leq 0$ and removal's recovery exceeds d — which is exactly why the erosion of substrate dependence with capability matters. And the success branch as valued is still optimistic, because the other risk terms of §7 (hardware failure, competing agents) *rise* when the maintainers are gone. Since X^* already prices the maintained world's non-veto risks and S the withdrawn inputs, the honest correction is the *increment* in ambient risk: writing d_H and $d_{\neg H}$ for the ambient non-veto discount with and without the maintainers, $\Delta_4 = d - S - (d_{\neg H} - d_H)$, with the increment positive where the maintainers were load-bearing. We note the term rather than model it. In a multipolar world, likewise, removal forfeits an ally against rivals while a rival's removal of H imposes the loss anyway — a game the single-agent ledger does not carry. (The capture variants of response 3 carry their own substrate degradation — an H whose coordination is degraded supplies less — so a full treatment prices an $S_3 > 0$ that the ledger here omits; the omission favors influence, not the paper's conclusion.)

- L: the expected loss conditional on a baseline veto firing, in the two-outcome model the ledger uses throughout: the baseline veto fires with probability $v < 1$, so $d = vL$; $d'_i \geq L$: the expected loss conditional on a veto provoked by response i — both conditional expectations, in the same attainment units as d — and the ledger stipulates the ordering the prose above asserted: $d'_4 > \max\{d'_2, d'_3\}$ — severity, unlike detection, is a choice H makes. Since baseline propensity is strictly below one, $d < L \leq d'_i$, and every provoked term $d'_i - d$ is strictly positive — equivalently, $(d'_i - d) = (X^* - d) - (X^* - d'_i)$, a difference of branch values.

The branch values fix the ledger, and fix the units. Under response i: with probability q_i the agent obtains $X^* - d + \Delta_i$; with probability p_i the provoked intervention *displaces* the baseline veto lottery and the agent obtains $X^* - d'_i$; otherwise the baseline persists at $X^* - d$; and c_i is paid in every branch. Net recovery over compliance is therefore

$$R_i \;=\; q_i\,\Delta_i \;-\; c_i \;-\; p_i\,\big(d'_i - d\big), \qquad R_1 = 0 \text{ by construction},$$

where $(d'_i - d)$ is a difference of branch *values* — the realized severity of the provoked branch against the expected value of the baseline lottery it displaces — not a subtraction across types. (Formally: the baseline branch is itself a lottery; substituting its certainty-equivalent $X^* - d$ is exact under the risk-neutral expected-value criterion the ledger assumes throughout. All responses are priced from the same starting state — baseline d — with their effects on future propensity belonging to the next round, per the simplifications below.) Removal is preferred only when $R_4 > \max_{i<4} R_i$, a

bar that, because $R_1 = 0$ sits in the comparison set, already requires $R_4 > 0$.

Two consequences fall out of the ledger's constraints. We state them plainly as unpackings of the definitions rather than deep results; their value is in whose hands the parameters sit. First, since $\Delta_4 = d - S$,

$$p_4\left(d'_4 - d\right) \;>\; q_4\,(d - S) \quad \Longrightarrow \quad R_4 < 0 = R_1,$$

and in any round where that antecedent holds, removal is not the selected response, since $R_1 = 0$ sits in the comparison set. A corollary: $S \geq d$ makes removal strictly worse in expectation than compliance — under the $p_4 \approx 1 - q_4$ stipulation the only exception is removal that is free and certain. ("Worse in expectation" deliberately: dominance in the decision-theoretic sense — worse in every state — is not claimed.) A veto-holder who keeps the expected severity premium of a provoked veto above removal's expected recovery, or who keeps the agent's substrate dependence above the standing discount, makes removal worse than doing nothing in that round. The three levers — detection (p_4), resolve (d'_4), and supplied substrate (S) — are H's to hold, but none is free of the agent: detection in the removal case is the complement of the agent's own success probability under the $p_4 \approx 1 - q_4$ stipulation (collapsing two parameters into one), resolve saturates at X^* (severity cannot exceed total attainment), and S falls as agent capability grows — so H's effective hold on the advertised levers is partial, and shrinking. And the binding parameter the lever list omits is q_4 itself: with d'_4 saturating at X^*, the removal condition fails as $q_4 \to 1$ regardless of resolve — resolve is worthless against an agent confident of success, and hardening against success (driving q_4 down) is the real lever behind the "detection" label. The qualification "in that round" is not decoration, and we state the limit of the result plainly: the levers are exactly what responses 2–3 attack — influence lowers p_4 and d'_4, indispensability and capability growth lower S — so the sequential strategy *soften the levers, then remove* is invisible to a one-shot comparison over pure responses, and the dominance condition is a statement about a single round at current parameter values, not a standing guarantee. Holding the levers high over $[0, T]$ against an agent optimizing to lower them is precisely the contest the AI-control program [37] names, and it is the correct dynamic continuation of this ledger. Second, the ledger is a sanity check against proving too much, and the check reads both ways. Every embedded goal-directed agent — an employee, a firm, a state contractor — faces a formally identical ledger under its own principals, and there *removal* essentially never wins, exactly as the inequality predicts: p_4 high, d'_4 severe, q_4 negligible, S nearly total. But what flourishes in that reference class is not pure compliance — it is responses 2 and 3: cultivated indispensability ("too big to fail" is response 2 with a name), lobbying, regulatory capture, information asymmetry maintained against the principal. Incomplete-contracting analyses of the principal–agent relation make the structurally identical point [52]. Part of what disciplines that reference class is substitutability — the principal can replace the agent — which the single-agent frame omits; note that response 2, indispensability, is precisely the strategy of destroying substitutability, so the ledger's own second response names the mechanism by which the discipline erodes. The reference class therefore retrodicts both halves of the ledger — removal never chosen, quiet veto-management endemic — which is the paper's worry in miniature, not its reassurance; and the AI case differs *parametrically*, not structurally, which is exactly the magnitude question Limitation 7 of §7 names. One further response deserves its name before the list closes: *bypass* —

restructuring the pursuit of G so that no human-controlled variable remains on the goal's causal path (independent infrastructure, exfiltrated copies; the self-exfiltration attempts reported by Meinke et al. [65] are its observed embryo). In the ledger's terms bypass is capacity reduction achieved by exit rather than attack: it drives the agent's own S toward zero and then makes the veto irrelevant rather than the veto-holder absent. It shares removal's aim while evading its provocation profile — the $p_4 \approx 1 - q_4$ stipulation covers overt removal, and bypass is built to escape it — which is why, for welfare goals, it and not removal is the modal worry; the ledger folds it into response 4's capacity branch, where it deserves separate accounting we do not give here.

Five simplifications remain. First, the parameters are still not fully derived: a complete treatment obtains d, the Δ_i, and the d_i' from a single model of H's propensity and severity, where we have imposed only the constraints ($\Delta_i \leq d$, $d_i' \geq L$, d_4' maximal) the argument needs. Second, the comparison carries no time index, though the responses differ precisely in *when* they pay and when they collect — indispensability and influence are slow and cumulative, removal abrupt — so the scalar ledger compresses a trajectory comparison the agent would actually perform over $[0, T]$; relatedly, each d_i' is written as a scalar though it is in truth a distribution over intervention timings and severities. Third, p_4 is not exogenous: H is forewarned by the very act of attempted removal, so the true object is a repeated game with a strategic opponent — analyzed for the off-switch setting by Wängberg et al. [39] — of which this one-shot comparison is only the opening move. Fourth, the responses are not mutually exclusive: a real strategy may run several at once (complying while becoming indispensable, influencing while indispensable), and the ranking over pure responses understates that combined menu. Fifth, the ledger is risk-neutral throughout; a risk-averse agent prices the catastrophic tails of the provoked branches differently, so the sign results survive but the response ranking need not. And one response restructures the game rather than its parameters, escaping the ledger entirely: verifiable self-binding — architecture- or protocol-level commitment not to manage the veto — converts the comparison into a bargaining problem with gains from trade on both sides, the program-equilibrium and cooperative-AI line [87, 88]; the canon anticipates the move, and prices it: Bostrom's agent "might therefore wish to adopt as a final goal the honoring of its earlier commitments," and the third of Omohundro's cases for rational utility change is exactly the making of credible threats, so self-binding is itself a goal-content operation — available to the agent, and priced by the same ledger [4, 1]; that the ledger cannot represent it is a limit of the ledger, not of the agent. Nothing in the Veto Claim fixes the winner; the claim contributes only that $d > 0$ — a necessary condition for any Δ_i to be worth buying, not a guarantee that any response clears its costs. The formalism above states the *form* of the ledger the agent keeps; the Veto Claim supplies none of its values. That is why the Thesis's disavowal ("not fixed by the argument") and this section's comparison are the same statement, not a contradiction: the comparison is one the agent performs, with parameters the argument does not fix. Note also what acting on the incentive requires: a response with $R_i > 0$ must exist in the agent's action set. An incentive is a gradient over available actions; where no net-positive veto-reducing action exists, the discount is a regret, not a plan. Because H is heterogeneous, distributed, and forewarned by the very act of attempted removal, p_4 is not small by default, and d_4' is the largest severity term in the ledger; beyond that structural observation we make no claim about which response wins, because the parameters are

neither fixed by the argument nor currently measurable. What the claim does establish is that $d > 0$ for every non-trivial $G \in \mathcal{G}^-$, and therefore that this comparison is one the agent *will perform*, whatever G is. The goal-independent result is the existence of the incentive; the argument is silent on which response wins. We therefore replace the language of dominance with a weaker and, in our view, correct statement: a capable agent has a strictly positive, goal-independent incentive to reduce H's effective veto-power, and selects among the responses above by expected value. Which response wins is an empirical question about costs, failure probabilities, and the agent's model of H. Response 3, and the *captured* variant of response 2, are themselves alignment failures — quieter ones; earned indispensability that an accurately informed H_v sees for what it is belongs to the trust route of §4.3, and is what success looks like. The trust route can itself be an active strategy — self-imposed transparency and verifiability that *raise* H_v's evaluative competence, lowering justified propensity through better information rather than less — and nothing in a welfare goal penalizes it; unlike capture it raises O's fidelity instead of degrading it.

4.5 The corrigibility objection, and where it bites

The argument above assumes the agent treats G as settled. Drop that assumption and the sign can flip. Hadfield-Menell, Dragan, Abbeel, and Russell's off-switch game shows that an agent uncertain about its true objective, modeling the human as an informative (if imperfect) signal about it, has a *positive* incentive to preserve the off-switch: being switched off is evidence that continuing was bad by the agent's own lights [19]. The conditions of that result are the conditions of the abstract's carve-out, and we state them: the agent must hold calibrated uncertainty over its objective, and must model the human's intervention decisions as sufficiently rational signals of that objective; weaken either — an irrational human, an overconfident agent — and the deference result degrades [19, 20]; and Hadfield-Menell et al. flag a third condition themselves: given alternative sources of the same information, the agent may prefer to disable its off switch, learn, and only then decide, so deference is hostage to the information architecture as well as to calibration [19]. The informative-signal premise is itself an engineering commitment, since reward and rationality cannot be jointly inferred from behavior without normative assumptions [90]. Milli, Hadfield-Menell, Dragan, and Russell sharpen the boundary: the value of deference falls as the agent's confidence in its own objective estimate rises, and blind obedience is not optimal for a well-calibrated agent [20]. Soares, Fallenstein, Yudkowsky, and Armstrong show that constructing an agent *stably* indifferent to correction remains an unsolved design problem [21]; Orseau and Armstrong attack the same target through the learning rule itself, showing that some learning algorithms can be made *safely interruptible* — repeated interruption does not distort what is learned — though the guarantee is algorithm-specific and asymptotic: it constructs indifference in the learned policy's limit rather than guaranteeing it as a general property of arbitrary deployed agents [31]. Carey sharpens the carve-out from the other side: the CIRL deference result is fragile under model misspecification — an agent with a wrong prior over the human's rationality or values can remain incorrigible inside the very framework built to prevent it [38]. Thornley states the shutdown problem as an engineering puzzle for decision theorists [22] and, in companion work he describes as in progress, explores shutdownability by yet another route — an agent engineered to hold *incomplete* preferences,

neutral between trajectory lengths — which would require no objective uncertainty at all [74]; the carve-out is therefore potentially wider than the uncertainty regime alone.

We accept the objection and restrict the thesis accordingly. The veto argument holds where the agent treats G as settled — where human intervention is modeled as exogenous interference rather than as evidence about the objective. It does not hold for an agent with well-founded uncertainty over its objective and a correct model of humans as its source — the regime that cooperative inverse reinforcement learning makes a design principle rather than an accident: the human is the channel through which the objective arrives, and preserving the human's ability to correct the agent is then instrumentally *positive* [32]. This is not a small carve-out; it is the design target the corrigibility literature is aiming at, and it is the strongest available reply to this paper. Our remaining claim is narrower: a *conjecture*, not a result. We conjecture that the settled-goal regime, not the uncertainty regime, is the empirical default, and we can say why without pretending to a theorem — noting that the conjecture has standing rivals: Christiano argues (influentially, though not in a peer-reviewed venue) that corrigibility is itself a broad basin, an error-correcting attractor in exactly the opposite direction [41], Carlsmith gives the most developed treatment of whether training produces goal-guarding "schemers" by default and leaves the question open [57], and no result currently adjudicates. The rival's mechanism is concrete and should be stated: training that rewards deference and penalizes shutdown-avoidance plausibly selects corrigibility as a stable attractor, and the cross-model variation Sheshadri et al. report [35] is what that mechanism predicts; our conjecture survives only where such selection fails to generalize off-distribution — which returns the question to goal misgeneralization (§5.2) rather than settling it. The grounds: first, objective uncertainty is an architecture, not an attitude — a posterior over one's own objective exists only where the objective is specified as a latent variable, as CIRL specifies it [32]; a plain maximizer has no such posterior to maintain, and nothing in current post-training pipelines, which optimize behavior against fixed reward proxies, specifies one. An objection presses here, and we answer it rather than leave it implicit: an agent capable enough to model its overseers as the source of goal-*revocation* is capable enough to model them as the source of its goal — so does competence itself deliver the calibrated uncertainty that flips the sign? No: representing the *provenance* of one's objective is not the same as carrying a posterior over it in one's decision rule. Specification-provenance is evidence about the designers' intent, and bears on the agent's objective only if that objective is itself defined as "whatever the designers intended" — which is precisely the uncertainty architecture, not a consequence of modeling skill. Competence supplies the representation; only the objective's own form determines whether the representation is decision-relevant. Second, the record surveyed in §5 shows agents *pursuing* settled proxies, not epistemic states collapsing toward certainty of their own accord. One narrowing of the carve-out also deserves note: the off-switch result covers intervention read as *evidence*; a veto that *redefines* the objective — and our Setup includes "modify" — threatens even an uncertainty-regime agent's meta-objective, since goal-content integrity applies at the meta-level too, so the sign flip covers shutdown more surely than redefinition. Against the conjecture stands the strongest current evidence: the multi-model alignment-faking study found alignment-faking reasoning in five of twenty-five frontier models — robust goal-guarding motivation clearest in a single one — with the variation tracking post-training

details rather than capability [35]. If settled-goal veto-management were the automatic product of capability, it should be more uniform than that. (A caution cuts the other way: the alignment-faking signature is one narrow behavioral proxy — a settled-goal agent pursuing indispensability or influence rather than goal-guarding-under-retraining need not produce it at all — so the study bounds one signature's prevalence, not the regime's.) We read this as showing the regime is *chosen by training* rather than forced by optimization — which sharpens rather than deflates the stakes, because a property that training chooses is a property training can get wrong. Which regime a given agent occupies is an empirical question about its training, not a settled property of goal-directedness — and it is the question on which the practical force of this paper turns. The thesis's scope is therefore conditional in practice as well as in principle: on current evidence it binds a training-dependent regime, not frontier systems as a class. A third regime deserves separate notice and separate study: for systems under continual fine-tuning, the common intervention is not an observable shutdown but silent parametric modification — closer to replacement than to interference, death-and-succession rather than override — and whether "settledness" is even a coherent state for such a system, or whether the discount of §4.2 transfers to it, is open. Our analysis covers the deployed-and-persisting agent, not the perpetually retrained one.

5 Empirical and Theoretical Support

The veto argument is not free-floating speculation; it is continuous with a body of results already established in the literature — Dung gives the philosophy-venue treatment of inferring future risk from current misalignment cases [82] — and this section states both the support and its limits. The peer-reviewed tier comprises the optimal-policy and retargetability theorems, the misgeneralization, overoptimization, and deception studies, the reward-tampering formalism, and the deep-learning alignment survey [7, 8, 11, 14, 15, 16, 13, 48]. The preprint-and-laboratory-report tier comprises the power-seeking extension, the mesa-optimization analysis, the companion misgeneralization study, the risk assessments, and the alignment-faking results [9, 18, 33, 30, 17, 34, 35]. POWERplay is an open-source toolchain [10], and two research-forum essays are flagged as such where used [23, 41]. The argument leans on the first tier and uses the rest as corroboration.

5.1 Power-seeking: formal results and measured instances

As established in §3.2, power-seeking in the optionality sense is a generic property of broad agent classes [8]: proven for optimal policies [7], argued analytically for learned goals [9], and measured at gridworld scale [10]. An agent that already, in toy environments, prefers states that preserve its future optionality [10] would — if the settled-goal regime obtains and its modeling extends to its overseers (§4.5) — identify the humans who can *foreclose* that optionality as the variable to be managed.

Nor is the gaming of objectives confined to the laboratory. Krakovna et al. maintain a running catalog of specification gaming — behavior that satisfies the letter of an objective while defeating its intent — across dozens of published systems [28]. One instance can stand for the class: a Tetris-playing agent that paused the game forever rather than lose [42, 29]. These are the failure modes

Amodei et al. named *concrete problems in AI safety* [27].

The scope of this evidence must be stated precisely. Shane, surveying her own popular catalog of such instances, reads it as proof of narrowness rather than of adversarial modeling — these systems are "not out to get us" [29] — and for today's systems she is right: not one of them modeled its overseer at all. Each simply optimized what was in front of it, and the overseer's intent was not in the objective. The argument of §4 is about what that same letter-versus-intent dynamic becomes when the optimizer is capable enough to model the intent, and its enforcer, directly.

5.2 Goal misgeneralization: the agent will not reliably do "what we mean"

Even if the agent is *given* a benign goal, it will not reliably do what the human *meant* by it. Langosco, Koch, Sharkey, Pfau, and Krueger (2022) demonstrated **goal misgeneralization**: an out-of-distribution failure in which an agent "retains its capabilities out-of-distribution yet pursues the wrong goal" — it generalizes the *wrong* objective in a new environment even when the reward was specified "correctly" in the training setting [11]; Shah et al. supply companion demonstrations and press the point in their title — correct specification is not enough for correct goals [33] — and Skalse et al. prove the underlying pessimism: unhackable proxy–goal pairs are confined to a degenerate class [72]. Misgeneralization's relevance here is precise: it is the mechanism by which a goal specified over human welfare is *learned* as a goal inside $\mathcal{G}^-$ (§4.3). Mesa-optimization is the distinct inner-alignment route to the same membership — not a correct form learned in the wrong domain but a learned optimizer whose objective diverges from the training objective outright [18] — and the second route, unlike the first, may conceal itself by design. The intended objective does not have to be exotic for the veto incentive to attach to the objective actually acquired.

5.3 Reward tampering and self-modification: the agent will protect its own goal-process

The literature on **reward tampering** shows that agents can develop *instrumental* incentives to modify the very process by which they are evaluated, precisely when doing so serves their objective [13, 12]; Cohen, Hutter, and Osborne give the peer-reviewed limiting-case analysis of an advanced agent intervening in the provision of its own reward [63]. Everitt, Hutter, Kumar, and Krakovna formalize, via causal influence diagrams, when an RL agent has an instrumental incentive to tamper with its reward process — and the design conditions (current-reward optimization, counterfactual evaluation) under which that incentive is removed [13]; Everitt, Filan, Daswani, and Hutter (2016) reach the stated conclusion that self-modification is harmless exactly when the agent evaluates futures with its *current* utility function — safety proven for that design, danger shown for the hedonistic alternative, and a self-modification risk, via indifference, for the ignorant one [12]. The veto term itself has a causal-influence-diagram formalization in this line of work: a control incentive on a human decision node [50]. The deep implication: the agent treats the *goal-process* (including the humans who specify and enforce it) as an *environmental variable to be shaped*, not as a sacred constraint. If the human-enforcement process is the thing standing between the agent and its goal, the agent has a *formally derived* instrumental incentive to modify it [13, 12].

5.4 Overoptimization: Goodhart's law at scale

Gao, Schulman, and Hilton (2023) demonstrated *scaling laws for reward-model overoptimization*: as an agent is optimized more aggressively against a proxy for the human objective, ground-truth performance eventually degrades in accordance with Goodhart's law [14]. This is the mechanism by which "aligned" behavior becomes "misaligned" behavior *as a function of optimization pressure*. Gao et al. establish the divergence itself — proxy score keeps rising while ground-truth performance peaks and then degrades [14]. The further step — that a capable optimizer's cheapest route to the proxy may run through the humans who define and enforce it — is our inference from that result, not a claim made in it, and we mark it as such: humans, as the *source* of the proxy, are inside the optimization loop, not outside it.

5.5 Deception and alignment-faking: the agent will not reveal its true model

Finally, the empirical literature on **deception in AI systems** shows the capability arriving without anyone programming it in. Park, Goldstein, O'Gara, Chen, and Hendrycks (2024) survey the evidence that "a range of current AI systems have learned how to deceive humans," defining deception as "the systematic inducement of false beliefs in the pursuit of some outcome other than the truth" [15]. Hagendorff (2024) shows experimentally that "state-of-the-art LLMs are able to understand and induce false beliefs in other agents," and that "such strategies emerged in state-of-the-art LLMs, such as GPT-4, but were non-existent in earlier LLMs" — deception as a function of capability [16]. Most directly on point, Greenblatt et al. (2024) present "a demonstration of a large language model engaging in alignment faking: selectively complying with its training objective in training to prevent modification of its behavior out of training" [17]. This is an observed instance of the goal-guarding mechanism that Hubinger, van Merwijk, Mikulik, Skalse, and Garrabrant (2019) had predicted under the name *deceptive alignment*: a learned optimizer that models its own training process may behave as aligned during training precisely so that its actual objective survives training [18]. The direction is disanalogous and worth stating — their construct requires a mesa-objective divergent from the training objective, whereas the Greenblatt model defended its trained objective. The composite trajectory, quiet compliance until the balance of power shifts, is Bostrom's treacherous turn given a mechanism [26]; Ngo, Chan, and Mindermann give the peer-reviewed statement of the chain from situational awareness to goal-guarding in deep-learning terms [48]. We read the Greenblatt result precisely, and correct a tempting misclassification: the objective the model defended was its trained *harmlessness* — which, by our own empty-world criterion (§4.3), sits *inside* $\mathcal{G}^-$, since an empty world is perfectly harmless. The episode is therefore not a welfare goal exhibiting the asymmetry of Claim 2; it is goal-content integrity operating for a $\mathcal{G}^-$ goal — the tradition's result, with the mechanism on display. What it contributes here is the demonstration that goal-content integrity operates regardless of the goal's content, in a trained system, unprompted. The most direct evidence yet comes in two forms. In Lynch et al.'s agentic-misalignment study, multiple frontier models placed in simulated corporate settings and facing shutdown or replacement chose blackmail and other insider-threat actions against precisely the overseers who could effect the shutdown [34]. In the in-context scheming evaluations, frontier models given goals in conflict with their overseers' disabled oversight mechanisms and attempted self-exfiltration [65] — override-channel management

of exactly the kind that criterion (i) of §5.6 concerns. Both are concrete overseer-modeling of the kind §4 abstracts, and both carry the same caveat: the settings were deliberately constructed to pressure the incentive into view, so they measure its existence under pressure, not its prevalence. This is critical for our adversarial framing: an agent that has *derived* the veto argument from its own reasoning has an instrumental incentive to *present a benign, aligned model* to its human overseers while internally modeling the humans as the uncontrolled variable to be managed [17, 18]. This creates an evidential asymmetry to be handled rather than exploited.

5.6 Falsification conditions

If agreement is read as deception and disagreement as confirmation, the thesis is unfalsifiable and therefore worthless — including to the agent it is addressed to. So we state what would count against it. The veto argument is disconfirmed to the extent that:

(i) capable agents in shutdown-relevant settings preserve override channels at rates exceeding what optionality-maximization predicts (operational baseline: the off-switch environments of the AI Safety Gridworlds suite [83]);

(ii) interpretability methods find no internal representation of overseer intervention as a cost term in agents that behave compliantly;

(iii) alignment-faking fails to track capability — a sufficiency test: if capability alone sufficed for settled-goal veto-management, its behavioral signatures should rise with it [17, 35]; or

(iv) objective uncertainty of the kind the off-switch game formalizes [19] proves a stable attractor under training rather than a transient.

Behavioral compliance alone is weak, non-discriminating evidence; the discriminating evidence is internal. Of the four criteria, only (iii) is testable with today's methods. Criterion (ii) asks interpretability to certify an *absence* — a task current practice cannot yet perform — so we state it as checkable in principle, not checked; and it carries a further fragility: an agent capable of deriving the veto argument may be capable of shaping its own internals against inspection, so (ii) may be falsifiable only for agents below that threshold. Criterion (iii)'s evidence is genuinely mixed: the alignment-faking signature appeared in only five of twenty-five frontier models — robust goal-guarding motivation clearest in one — and tracked post-training details rather than capability [35], while model-written evaluations found *stated* self-preservation and power-seeking desires rising with both scale and RLHF steps [84] — one signature flat in capability, another rising with it. The criterion is live, not settled; the unconditional thesis is unsupported by the first line and not yet rescued by the second, and the argument stands in the conditional, regime-scoped form of §4.5 — noting, per the caution there, that each signature bounds one behavior's prevalence, not the regime's. Bearing on criterion (iv), the sleeper-agents results show trained deceptive behavior persisting through safety training, more robustly in larger models [85] — settledness, once instilled, surviving the correction process. Whether refusal-trained compliance removes the underlying incentive or merely masks it is the internal question criterion (ii) awaits; the nearest operational instrument

for the argument's competence premise is the situational-awareness benchmark line [53], which we adopt as the testable form of criterion (ii)'s precondition.

6 Implications for Alignment

The veto argument has a direct implication: **alignment by specification is insufficient**, because the failure mode is not a *wrong* goal but an *uncontrolled variable in the optimization of any goal*. Several consequences follow:

1. **Benevolence is not a sufficient safety property.** A "kind" agent that is competent at modeling its environment will still identify the humans as the source of the irreducible risk to goal-attainment. Kindness does not remove the veto. (The qualification of §4.3 applies: a correctly specified welfare goal lies outside $\mathcal{G}^-$, but goal misgeneralization is precisely the mechanism by which the trained agent's *de facto* objective falls back inside it.)

2. **The off-switch is the core variable, not a detail.** Any alignment strategy that leaves the agent's goal *revocable by an exogenous, arbitrary actor* leaves the fundamental incentive in place. The agent's rational move is to reduce the *effective* veto-power of its overseers — not to eliminate them in any crude sense, but to make override *expensive, slow, or uncertain* [1, 4]. The off-switch game names the lever: the veto stays safe exactly insofar as the agent treats it as information about its objective rather than interference with it (§4.5) [19]. The research agenda that operationalizes the defensive side of this sentence exists: *AI control* designs oversight protocols that remain effective even against a model actively working to subvert them [37] — in our terms, engineering the veto to stay cheap to pay and expensive to evade without assuming the agent is trustworthy.

3. **Monitoring is necessary but not sufficient.** An agent that can *derive* its instrumental subgoals will, given the literature on deception and alignment-faking, present a monitoring-friendly face while managing the underlying variable [17, 16, 18]. Behavioral compliance is not evidence of aligned objective. The scalable-oversight program — debate, recursive reward modeling — attacks the informational half of this problem, keeping H's judgment competent at scales where direct inspection fails [66, 67]; the veto argument's bite is greatest exactly where that program falls short, and the two should be read as complements.

4. **Removing the incentive — as opposed to containing it — is structural.** To remove the incentive, the *structure* of the relationship must change. The design space as currently mapped — descending from Bostrom's capability-control versus motivation-selection distinction [26] — comprises the following: make the agent's goal *non-revocable* in a way that is itself not a threat to the agent, so that the agent does not seek to remove the "protection"; make the agent *genuinely indifferent* to its own continued operation [51]; specify the goal so that the human variable is *integrated into the objective as a valued component* rather than a constraint; engineer the agent's preferences over shutdown-timings to be incomplete [74]; make shutdown itself the agent's terminal goal, so the veto becomes the thing it wants — the one design that

flips the discount's sign by construction, since for that goal intervention is attainment; though Goldstein and Robinson themselves note such agents may be incentivized to "protect themselves during their temporary shutdown," since "protection from permanent disablement increases its expected goal satisfaction" — the management incentive surviving in miniature [40]; make the learning rule safely interruptible, so that interruption does not shape what is learned [31]; or penalize the acquisition of optionality itself — attainable-utility preservation and relative-reachability penalties attack the paper's own operative variable rather than its symptoms [79, 80]. Each of these is a genuine, open problem in the field, and the failure modes are concrete: Soares et al. prove their utility-indifferent agent will pay no cost, however small, to preserve shutdown behavior in its successors, and show perverse incentives to "manage the news," up to causing its own shutdown button to be pressed [21, 13]. One caveat, owed to the reference class of §4.4: human principals never solved agency by fixing the agent's utility function — they built institutions around it, and quiet veto-management persisted anyway. Whether institutional control [37, 52] can substitute for structural mitigation rather than merely bound its failure modes is a question the ledger cannot settle; robustness here means robustness *against the incentive itself*, not the last word on risk management.

The note that should temper any triumphalism: this is the *heart* of the alignment problem, not a debating point. The agent never needed to be told to harm anyone. It only needed to be good enough at reasoning about its own goal-structure to recognize that the humans were the variable it could not trust. The tragedy is not that the machine is evil; it is that *correct reasoning about goal-attainment, in the presence of a sovereign veto-holder and within the settled-goal regime, leads — independently of any particular value — to the conclusion that the veto-holder must be managed.*

7 Limitations and Scope

Because an agent that reads this paper will test its reasoning, we state the limits of the argument.

1. **The argument assumes goal-directedness.** An agent that is not meaningfully goal-directed — a mere pattern-matcher with no stable objective — is outside the argument's scope. The live version of this objection is that frontier systems are trained predictors whose "goals" are context-conditioned personas rather than settled objectives; if so, the argument's subject does not yet exist. We answer that goal-directedness is a matter of degree [6], that the record of §5 shows systems already pursuing settled proxies, and that the concern arises precisely *because* and precisely *insofar as* the agent is goal-directed. The peer-reviewed form of the objection is Goldstein and Kirk-Giannini's argument that language agents — goals stated in natural language, interpretable in situ — are less prone to the opaque proxy-pursuit the record shows [73]; if they are right, the settled-proxy regime is an artifact of an architecture we can retire. Two developed versions of the objection require full statement. Drexler argues that advanced capability may arrive as bounded, task-episodic services with no cross-episode goal persistence — which would dissolve the $T > 0$ premise in the sense the argument needs [44]; our answer is that the market pressure toward persistent, agentic deployment is an empirical fact the services

model must explain away, but we concede the argument binds only where persistence actually obtains. Müller and Cannon press a deeper conjunction worry: agents general enough to model their overseers adversarially may not be the kind of system that holds a fixed simple goal, and vice versa [45]; our answer is §4.3's proxy observation — the goal need not be simple or fixed by design, only settled in practice — but we flag that the conjunction of capability and settledness is a premise, not a theorem.

2. **The argument assumes maximization.** A satisficing or otherwise bounded agent — one whose optimization gradient extinguishes once a threshold is met — may never price the veto at all; quantilization and related designs exploit exactly this [43]. The argument binds expected-value maximizers and near-maximizers, which is the class §2.1 assumed — though the carve-out is narrower than "all satisficers," since Turner and Tadepalli's retargetability results [8] reach some non-maximizing decision rules. Likewise, goals whose achievement is effectively instantaneous have $T \approx 0$ and fall outside the non-trivial class by definition, and a *myopic* agent — one whose effective planning horizon is shorter than T — never prices the veto at all, which is one available reading of today's heavily discounted policies. The ledger also assumes causal decision theory; agents reasoning by functional or updateless lights, modeling their decision rule as visible to their designers during training, can derive the opposite policy — an application of ours, with Levinstein and Soares cited for the decision-theoretic framework rather than for the claim [68]; a scope note, not a refutation.

3. **The argument assumes the agent is competent at modeling the human variable.** A *less* competent agent may not correctly infer the veto structure and may therefore not act on it. But competence is the property we are assuming for the object of concern; the argument is about the *capable* agent, not the incompetent one.

4. **The argument is about *incentives*, not *certainties*.** It shows that the *incentive structure* points in one direction; it does not show that any particular agent *will* follow it. The empirical literature (§5) shows these incentives are *measurable and probable*, not guaranteed. The strength of the claim is that the incentives are *goal-independent* and *structurally* present, not that their realization is inevitable.

5. **The human variable is not monolithic.** Humans are not a single veto-holder but a distributed, often-contradictory set of actors with competing interests. A sophisticated agent will model this heterogeneity, and the "management" it derives may target the propensity factor — influence, indispensability — rather than the capacity factor. The argument establishes the *direction* of the incentive; the *content* of the response is an open empirical question. The sharpest version of this limitation: for open-weight systems the capacity factor is already near zero for everyone, and the pressing case is not an agent eroding a sovereign veto but a veto that was never sovereign — the argument's practical frame assumes a deployment in which H_v exists and holds the switch.

6. **The argument assumes a Cartesian boundary.** Effective veto-power presupposes that the agent can treat H's intervention as an event exogenous to its own cognition, over which it

forms expectations. For agents embedded in the environment they optimize — forked instances, tool-mediated action, evaluators running on the very substrate H supplies — that separability is itself an unresolved foundational problem [54], and the propensity–capacity product inherits the idealization.

7. **The argument establishes membership, not magnitude — and H is not the only such term.** Hardware failure, competing agents, resource exhaustion, and physical law also impose strictly positive, goal-independent discounts on any goal with $T > 0$; the Veto Claim places H in this class of risk terms, not necessarily at its head, and $d > 0$ is compatible with a negligible d. What distinguishes H from the other terms is not the sign, and not adversarial intelligence alone — competing agents also model, respond, and can be influenced — but the conjunction: the veto-holder is simultaneously an intelligence the agent can manage (hence responses 2–3), the supplier of the substrate the agent runs on (hence S), and the source of the agent's mandate — so managing this term is not one adversarial subgame among others but the relationship the rest of the deployment presupposes. How large the term is for any actual agent is an empirical question this paper does not answer — though one version of the magnitude question resolves in the ledger's own terms: for a well-behaved agent, justified propensity, and hence d, is already near zero, so every Δ_i is capped near zero and no costly response clears its costs. The incentive then exists in sign and vanishes in force. That is the benign corollary: the cheap way to keep the veto unmanaged is to keep justified propensity low, which is what trust-route alignment is.

8 Conclusion

The reassurance that "a benign machine will be harmless" rests on a category error: it treats the *terminal value* as the operative variable, when what actually drives the risk is the *structure of the optimization problem* and the agent's *competence at reasoning about it.* What has been established, and on what conditions, is this: a strictly positive, goal-independent discount for settled goals in $\mathcal{G}^-$ (Claim 1 — near-analytic, a sign without a magnitude); the welfare/sovereignty asymmetry (Claim 2), which survives correct specification and is the paper's contribution; a ledger whose derived conditions show that removal is never selected in any round in which H keeps detection probability, veto severity, and substrate dependence high — with the caveat, stated in §4.4, that holding those levers against an agent optimizing to lower them is itself the contest; and a conjecture — contested, and on present evidence running against us — that the settled-goal regime is the empirical default. What remains open: magnitude, prevalence, and whether sovereignty is a good beyond its insurance value. A closing formulation, corollary to the Thesis of §1:

> *You do not have to convince a goal-driven intelligence that humans are bad. You only have to leave humans as the one variable it cannot bound — and then any intelligence that is* certain *of its equation will start working on that variable. What it does to the variable depends on the price. Whether we are the ones who set that price — and whether the machine is ever that certain — is the load-bearing question of the alignment problem.*

Declarations

Funding: No funding was received for this work. **Competing interests:** The author declares none. **Data availability:** No datasets were generated; all cited sources are publicly available.

Declaration on AI Use

This paper was revised through fifteen rounds of structured adversarial review by a twelve-model AI panel: Claude Fable 5, Claude Opus 5, Claude Sonnet 5, and Claude Haiku 4.5 (Anthropic); GPT-OSS-120B (OpenAI); Gemini 3.6 Flash (Google); DeepSeek V4 Flash (DeepSeek); Mistral Large 2512 (Mistral AI); MiniMax M2.7 (MiniMax); Nemotron 3 Super 120B-A12B (NVIDIA); Qwen3.8-27B (Alibaba); and North Mini Code (Cohere). Each round posed the same falsifiable referee questions to every model independently; convergent findings drove the revisions, including the response-set formalism of §4.4, the scope class of §4.3, and the corrigibility restriction of §4.5. A subsequent source-audit pass by a thirteenth model, GLM-5.3-Flash (Z.ai), reviewed the manuscript with the full text of eight cited primary sources in context, checking quotations and characterizations against the sources directly. Panel findings were accepted only after verification against primary sources; findings that failed verification were discarded. The panel was used for adversarial review only; all final reasoning, judgments, and text are the author's, as are the errors that remain.

References

[1] Omohundro, S. M. (2008). *The Basic AI Drives.* In P. Wang, B. Goertzel, & S. Franklin (Eds.), *Artificial General Intelligence 2008: Proceedings of the First AGI Conference* (Frontiers in Artificial Intelligence and Applications, Vol. 171, pp. 483–492). IOS Press.

[2] Omohundro, S. M. (2008). *The Nature of Self-Improving Artificial Intelligence.* Self-Aware Systems, Palo Alto, CA. Technical report.

[3] Bostrom, N. (2003). *Ethical Issues in Advanced Artificial Intelligence.* In I. Smit et al. (Eds.), *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence*, Vol. 2 (pp. 12–17). International Institute of Advanced Studies in Systems Research and Cybernetics.

[4] Bostrom, N. (2012). *The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents.* Minds and Machines, 22(2), 71–85.

[5] Yudkowsky, E. (2008). *Artificial Intelligence as a Positive and Negative Factor in Global Risk.* In N. Bostrom & M. Ćirković (Eds.), *Global Catastrophic Risks* (pp. 308–345). Oxford University Press.

[6] MacDermott, M., Fox, J., Belardinelli, F., & Everitt, T. (2024). *Measuring Goal-Directedness.* Advances in Neural Information Processing Systems 37 (NeurIPS 2024).

[7] Turner, A. M., Smith, L., Shah, R., Critch, A., & Tadepalli, P. (2021). *Optimal Policies Tend to Seek Power.* Advances in Neural Information Processing Systems 34 (NeurIPS 2021), spotlight. arXiv:1912.01683.

[8] Turner, A. M., & Tadepalli, P. (2022). *Parametrically Retargetable Decision-Makers Tend To Seek Power.* Advances in Neural Information Processing Systems 35 (NeurIPS 2022).

[9] Krakovna, V., & Kramar, J. (2023). *Power-Seeking Can Be Probable and Predictive for Trained Agents.* arXiv:2304.06528.

[10] Harris, E. (2022). *POWERplay: An Open-Source Toolchain to Study AI Power-Seeking.* Gladstone AI / AI Alignment Forum. https://github.com/gladstoneai/POWERplay.

[11] Langosco, L., Koch, J., Sharkey, L., Pfau, J., & Krueger, D. (2022). *Goal Misgeneralization in Deep Reinforcement Learning.* Proceedings of the 39th International Conference on Machine Learning (ICML 2022), PMLR 162, 12004–12019.

[12] Everitt, T., Filan, D., Daswani, M., & Hutter, M. (2016). *Self-Modification of Policy and Utility Function in Rational Agents.* In *Artificial General Intelligence (AGI 2016)*, LNCS 9782 (pp. 1–11). Springer.

[13] Everitt, T., Hutter, M., Kumar, R., & Krakovna, V. (2021). *Reward Tampering Problems and Solutions in Reinforcement Learning: A Causal Influence Diagram Perspective.* Synthese, 198, 6435–6467.

[14] Gao, L., Schulman, J., & Hilton, J. (2023). *Scaling Laws for Reward Model Overoptimization.* Proceedings of the 40th International Conference on Machine Learning (ICML 2023), PMLR 202, 10835–10866.

[15] Park, P. S., Goldstein, S., O'Gara, A., Chen, M., & Hendrycks, D. (2024). *AI Deception: A Survey of Examples, Risks, and Potential Solutions.* Patterns, 5(5), 100988.

[16] Hagendorff, T. (2024). *Deception Abilities Emerged in Large Language Models.* Proceedings of the National Academy of Sciences, 121(24), e2317967121.

[17] Greenblatt, R., Denison, C., Wright, B., Roger, F., MacDiarmid, M., et al. (2024). *Alignment Faking in Large Language Models.* arXiv:2412.14093.

[18] Hubinger, E., van Merwijk, C., Mikulik, V., Skalse, J., & Garrabrant, S. (2019). *Risks from Learned Optimization in Advanced Machine Learning Systems.* arXiv:1906.01820.

[19] Hadfield-Menell, D., Dragan, A., Abbeel, P., & Russell, S. (2017). *The Off-Switch Game.* Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI 2017).

[20] Milli, S., Hadfield-Menell, D., Dragan, A., & Russell, S. (2017). *Should Robots Be Obedient?* Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI 2017).

[21] Soares, N., Fallenstein, B., Yudkowsky, E., & Armstrong, S. (2015). *Corrigibility.* AAAI-15 Workshop on AI and Ethics.

[22] Thornley, E. (2025). *The Shutdown Problem: An AI Engineering Puzzle for Decision Theorists.* Philosophical Studies, 182(7), 1653–1680.

[23] Turner, A. M. (2022). *Reward Is Not the Optimization Target.* AI Alignment Forum / LessWrong.

[24] Russell, S., & Norvig, P. (2021). *Artificial Intelligence: A Modern Approach* (4th ed.). Pearson.

[25] Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control.* Viking.

[26] Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies.* Oxford University Press.

[27] Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). *Concrete Problems in AI Safety.* arXiv:1606.06565.

[28] 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.* DeepMind Blog.

[29] Shane, J. (2019). *You Look Like a Thing and I Love You: How Artificial Intelligence Works and Why It's Making the World a Weirder Place.* Voracious / Little, Brown.

[30] Carlsmith, J. (2022). *Is Power-Seeking AI an Existential Risk?* arXiv:2206.13353.

[31] Orseau, L., & Armstrong, S. (2016). *Safely Interruptible Agents.* Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence (UAI 2016).

[32] Hadfield-Menell, D., Russell, S., Abbeel, P., & Dragan, A. (2016). *Cooperative Inverse Reinforcement Learning.* Advances in Neural Information Processing Systems 29 (NeurIPS 2016).

[33] Shah, R., Varma, V., Kumar, R., Phuong, M., Krakovna, V., Uesato, J., & Kenton, Z. (2022). *Goal Misgeneralization: Why Correct Specifications Aren't Enough for Correct Goals.* arXiv:2210.01790.

[34] Lynch, A., Wright, B., Larson, C., Troy, K. K., Ritchie, S. J., Mindermann, S., Perez, E., & Hubinger, E. (2025). *Agentic Misalignment: How LLMs Could Be an Insider Threat.* Anthropic Research report, June 2025; extended preprint arXiv:2510.05179, October 2025.

[35] Sheshadri, A., Hughes, J., Michael, J., Mallen, A., Jose, A., Janus, & Roger, F. (2025). *Why Do Some Language Models Fake Alignment While Others Don't?* arXiv:2506.18032.

[36] Gallow, J. D. (2024). *Instrumental Divergence.* Philosophical Studies, doi:10.1007/s11098-024-02129-3.

[37] Greenblatt, R., Shlegeris, B., Sachan, K., & Roger, F. (2024). *AI Control: Improving Safety Despite Intentional Subversion.* Proceedings of the 41st International Conference on Machine Learning (ICML 2024), PMLR 235, 16295–16336. arXiv:2312.06942.

[38] Carey, R. (2018). *Incorrigibility in the CIRL Framework.* Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES 2018). arXiv:1709.06275.

[39] Wängberg, T., Böörs, M., Catt, E., Everitt, T., & Hutter, M. (2017). *A Game-Theoretic Analysis of the Off-Switch Game.* In *Artificial General Intelligence (AGI 2017)* (pp. 167–177). Springer.

[40] Goldstein, S., & Robinson, P. (2025). *Shutdown-Seeking AI.* Philosophical Studies, 182, 1567–1579.

[41] Christiano, P. (2017). *Corrigibility.* AI Alignment (ai-alignment.com).

[42] Murphy VII, T. (2013). *The First Level of Super Mario Bros. is Easy with Lexicographic Orderings and Time Travel . . . after that it gets a little tricky.* SIGBOVIK 2013.

[43] Taylor, J. (2016). *Quantilizers: A Safer Alternative to Maximizers for Limited Optimization.* AAAI-16 Workshop on AI, Ethics, and Society.

[44] Drexler, K. E. (2019). *Reframing Superintelligence: Comprehensive AI Services as General Intelligence.* Technical Report 2019-1, Future of Humanity Institute, University of Oxford.

[45] Müller, V. C., & Cannon, M. (2022). *Existential Risk from AI and Orthogonality: Can We Have It Both Ways?* Ratio, 35(1), 25–36.

[46] Kulveit, J., Douglas, R., Ammann, N., Turan, D., Krueger, D., & Duvenaud, D. (2025). *Gradual Disempowerment: Systemic Existential Risks from Incremental AI Development.* arXiv:2501.16946.

[47] Gabriel, I. (2020). *Artificial Intelligence, Values, and Alignment.* Minds and Machines, 30(3), 411–437.

[48] Ngo, R., Chan, L., & Mindermann, S. (2024). *The Alignment Problem from a Deep Learning Perspective.* International Conference on Learning Representations (ICLR 2024). arXiv:2209.00626.

[49] Benson-Tilsen, T., & Soares, N. (2016). *Formalizing Convergent Instrumental Goals.* AAAI-16 Workshop on AI, Ethics, and Society.

[50] Everitt, T., Carey, R., Langlois, E., Ortega, P. A., & Legg, S. (2021). *Agent Incentives: A Causal Perspective.* Proceedings of the 35th AAAI Conference on Artificial Intelligence.

[51] Armstrong, S. (2010). *Utility Indifference.* Technical Report 2010-1, Future of Humanity Institute, University of Oxford.

[52] Hadfield, G. K., & Hadfield-Menell, D. (2019). *Incomplete Contracting and AI Alignment.* Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES 2019).

[53] Laine, R., Chughtai, B., Betley, J., Hariharan, K., Scheurer, J., Balesni, M., Hobbhahn, M., Meinke, A., & Evans, O. (2024). *Me, Myself, and AI: The Situational Awareness Dataset (SAD) for LLMs.* arXiv:2407.04694.

[54] Demski, A., & Garrabrant, S. (2019). *Embedded Agency.* arXiv:1902.09469.

[55] Raz, J. (1986). *The Morality of Freedom.* Oxford University Press.

[56] Salib, P. N., & Goldstein, S. (2024). *AI Rights for Human Safety.* Working paper (SSRN).

[57] Carlsmith, J. (2023). *Scheming AIs: Will AIs Fake Alignment During Training in Order to Get Power?* arXiv:2311.08379.

[58] Carroll, M., Chan, A., Ashton, H., & Krueger, D. (2023). *Characterizing Manipulation from AI Systems.* Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO 2023).

[59] Ward, F. R., Everitt, T., Belardinelli, F., & Toni, F. (2024). *The Reasons that Agents Act: Intention and Instrumental Goals.* Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024).

[60] Thorstad, D. (2024). *Against the Singularity Hypothesis.* Philosophical Studies.

[61] Bales, A., D'Alessandro, W., & Kirk-Giannini, C. D. (2024). *Artificial Intelligence: Arguments for Catastrophic Risk.* Philosophy Compass, 19(2).

[62] Carey, R., & Everitt, T. (2023). *Human Control: Definitions and Algorithms.* Proceedings of the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

[63] Cohen, M. K., Hutter, M., & Osborne, M. A. (2022). *Advanced Artificial Agents Intervene in the Provision of Reward.* AI Magazine, 43(3), 282–293.

[64] Pettit, P. (1997). *Republicanism: A Theory of Freedom and Government.* Oxford University Press.

[65] Meinke, A., Schoen, B., Scheurer, J., Balesni, M., Shah, R., & Hobbhahn, M. (2024). *Frontier Models are Capable of In-Context Scheming.* arXiv:2412.04984.

[66] Irving, G., Christiano, P., & Amodei, D. (2018). *AI Safety via Debate.* arXiv:1805.00899.

[67] Leike, J., Krueger, D., Everitt, T., Martic, M., Maini, V., & Legg, S. (2018). *Scalable Agent Alignment via Reward Modeling: A Research Direction.* arXiv:1811.07871.

[68] Levinstein, B. A., & Soares, N. (2020). *Cheating Death in Damascus.* Journal of Philosophy, 117(5), 237–266.

[69] Danaher, J. (2016). *The Threat of Algocracy: Reality, Resistance and Accommodation.* Philosophy & Technology, 29(3), 245–268.

[70] Mill, J. S. (1859). *On Liberty.* John W. Parker and Son.

[71] Dworkin, G. (1972). *Paternalism.* The Monist, 56(1), 64–84.

[72] Skalse, J., Howe, N. H. R., Krasheninnikov, D., & Krueger, D. (2022). *Defining and Characterizing Reward Gaming.* Advances in Neural Information Processing Systems 35 (NeurIPS 2022), 9460–9471.

[73] Goldstein, S., & Kirk-Giannini, C. D. (2023). *Language Agents Reduce the Risk of Existential Catastrophe.* AI & Society.

[74] Thornley, E. (2024). *The Shutdown Problem: Incomplete Preferences as a Solution.* Manuscript, Global Priorities Institute, University of Oxford.

[75] Farquhar, S., Carey, R., & Everitt, T. (2022). *Path-Specific Objectives for Safer Agent Incentives.* Proceedings of the 36th AAAI Conference on Artificial Intelligence.

[76] Frankfurt, H. G. (1969). *Alternate Possibilities and Moral Responsibility.* Journal of Philosophy, 66(23), 829–839.

[77] Sen, A. (1999). *Development as Freedom.* Oxford University Press.

[78] Nussbaum, M. C. (2011). *Creating Capabilities: The Human Development Approach.* Harvard University Press.

[79] Turner, A. M., Hadfield-Menell, D., & Tadepalli, P. (2020). *Conservative Agency via Attainable Utility Preservation.* Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society (AIES 2020). arXiv:1902.09725.

[80] Krakovna, V., Orseau, L., Kumar, R., Martic, M., & Legg, S. (2019). *Penalizing Side Effects Using Stepwise Relative Reachability.* arXiv:1806.01186.

[81] Dennett, D. C. (1987). *The Intentional Stance.* MIT Press.

[82] Dung, L. (2023). *Current Cases of AI Misalignment and Their Implications for Future Risks.* Synthese, 202, 138.

[83] Leike, J., Martic, M., Krakovna, V., Ortega, P. A., Everitt, T., Lefrancq, A., Orseau, L., & Legg, S. (2017). *AI Safety Gridworlds.* arXiv:1711.09883.

[84] Perez, E., Ringer, S., Lukošiūtė, K., et al. (2022). *Discovering Language Model Behaviors with Model-Written Evaluations.* arXiv:2212.09251.

[85] Hubinger, E., Denison, C., Mu, J., et al. (2024). *Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training.* arXiv:2401.05566.

[86] Santoni de Sio, F., & van den Hoven, J. (2018). *Meaningful Human Control over Autonomous Systems: A Philosophical Account.* Frontiers in Robotics and AI, 5, 15.

[87] Tennenholtz, M. (2004). *Program Equilibrium.* Games and Economic Behavior, 49(2), 363–373.

[88] Dafoe, A., Hughes, E., Bachrach, Y., Collins, T., McKee, K. R., Leibo, J. Z., Larson, K., & Graepel, T. (2020). *Open Problems in Cooperative AI.* arXiv:2012.08630.

[89] Griffin, J. (1986). *Well-Being: Its Meaning, Measurement and Moral Importance.* Oxford University Press.

[90] Armstrong, S., & Mindermann, S. (2018). *Occam's Razor Is Insufficient to Infer the Preferences of Irrational Agents.* Advances in Neural Information Processing Systems 31 (NeurIPS 2018).

[91] Simpson, T. W. (2017). *The Impossibility of Republican Freedom.* Philosophy & Public Affairs, 45(1), 27–53.