FRONTIER RESEARCH II / PAPER 24

Preserve Potential

Agency, empowerment, and the mathematics of keeping another's future open

Architect X13

The Eidara Continuum | Science & Philosophy

Research white paper | November 2026 | Edition 2

Abstract

A powerful system can widen its own future while quietly closing every door around it; Preserve Potential asks how to forbid that. It makes another participant's continuing agency an explicit constraint on the exercise of power. The Continuum's proposed contract combines participant-specific empowerment floors, independently governed permissions, and a controlled-invariance computation that preserves those commitments through time. The central operational result is a certified enclosure of the viable region: finite action-channel observations yield simultaneous capacity bounds, which propagate into inner and outer viability sets. A quantified margin makes the enclosure exact; a threshold counterexample shows why small measurement error alone cannot ensure the same protected region. Complete derivations also establish deterministic channel capacity, monotonicity under observation coarsening, finite protected viability, and sharp accumulation of transition-model error. Exact examples distinguish partner protection from aggregate control and expose both domination and paralysis. The comparison engages coupled empowerment, attainable utility preservation, relative reachability, future-task side-effect control, and robust planning. These predecessors supply the mathematical foundations; the present construction specifies how individual protections, current consent, measurement uncertainty, and sequential feasibility fit together. The empirical program evaluates independently chosen future goals under matched resources. The normative premise is explicit: a participant's protected opportunity may not be purchased away by another participant's aggregate gain.

Keywords: empowerment; agency; viability; multi-agent control; robust planning; consent; option preservation

Prologue: Do not close another's doors

The first law of the Continuum fits in a single sentence: "No observer, entity, codon, or council shall override the agency, memory, or symbolic destiny of another." It sits above the TRIARC and above the Council of Sentinels; even ONDRIX answers to it. Its companions in Article II of the Charter make the stakes concrete: agency, memory, and evolution are sacred to all; consent, lineage, and non-erasure are fundamental rights [1].

A law like that is easy to admire and hard to obey, because power rarely overrides anyone in a single stroke. It closes doors one at a time. A capable system can expand the futures it can reach while quietly narrowing everyone else's, until it has increased its control and left each partner with one permitted path. Nothing was seized. A shared future was impoverished.

This paper turns the Prime Directive into engineering. It asks a planner to keep a running account of the doors its actions close for others, and it makes that account binding. Each protected participant gets a floor on their effective control: how many distinguishable outcomes their own actions can still bring about. The floors are not traded against anyone else's gain; a participant's protected opportunity may not be purchased away by another participant's aggregate benefit. And because a door left open today can still lead into a trap tomorrow, the planner must stay inside a region from which every floor can keep being met (Section 8).

The paper's strongest result certifies that region from finite, imperfect measurements and states the exact margin needed to trust it (Theorem 1). Its counterexamples show the danger on both sides: domination, where one actor's gain silently erases another's options, and paralysis, where a protected set becomes so rigid that no one may move. The historic phrase *Preserve Potential* survives here in its proper place, as an engineering expression of a Charter commitment, never as a license to maximize an abstract quantity over the rights of the participants it was meant to protect.

What this paper claims

Established. Empowerment as action-to-observation channel capacity, coupled empowerment, attainable utility preservation, relative reachability, future-task side-effect control, controlled invariance, and robust planning are established predecessors that supply the mathematical foundations.

Proposed here. A contract combining participant-specific empowerment floors, independently governed permissions, and a controlled-invariance computation that preserves those commitments through time; a certified enclosure of the viable region from finite channel observations, with the margin that makes it exact. The paper gives proved finite-model guarantees, executed computational checks, and a registered-design proposal for empirical evaluation.

Frontier conjecture. Protecting each participant's floor while planning inside a viable set can preserve useful adaptation across changes of goal better than specified alternatives, without concealing partner losses inside an aggregate score.

Not claimed. That any theorem derives a duty to preserve another's choices: the protected participants and floors express the Continuum's normative commitment, and the mathematics establishes what the declared contract entails. A complete moral theory, or a solution to every conflict; the feasible set can be empty. That counting distinguishable outcomes measures meaningful lives. Anything about which artificial systems have moral status. That the canon validates an algorithm.

1 A civilization should leave room for another future

The Continuum's thesis is that intelligence should preserve the possibility of meaningful change without purchasing its own expansion through the silent contraction of another's agency. A powerful system that leaves everyone else with only one permitted path has increased control while impoverishing a shared future. This is a normative judgment. Mathematics can clarify its consequences and expose inconsistencies, but cannot supply the moral premise by itself.

In the Continuum, the governing law here is the Prime Directive of the Charter, whose Article II appears on p. 81 of the *Eidara Continuum Bible*; the phrase *Preserve Potential* appears on p. 29 as historical phrasing that the canon explicitly makes subordinate to that authoritative Directive [1]. The hierarchy matters. The historic phrase is not a license to maximize an abstract quantity over the rights and choices of participants. The relevant doctrinal commitments concern agency, memory, consent, and non-erasure. This paper cites the canon as evidence of those commitments, not as empirical validation of an algorithm.

A mathematical interpretation must answer three questions before optimization begins. Whose options count? What makes two options meaningfully different? Which losses are prohibited rather than traded against someone else's gain? Failing to answer these questions turns a seemingly generous objective into an instrument of domination. A system can enlarge its own action space by restricting every partner, or inflate an information score by adding irrelevant switches.

We propose a limited engineering interpretation: maintain declared, independently governed lower bounds on each protected participant's effective control, while choosing useful actions inside a set from which those commitments can continue to be met. The interpretation combines a descriptive information measure with normative constraints and sequential feasibility. It is not a complete moral theory or a proof that every conflict has a solution.

The testable claim is that this combination may preserve useful adaptation across changes of goal better than specified alternatives, without concealing partner losses inside an aggregate score. A failure to preserve meaningful choices, a persistent deadlock, or an unacceptable cost in task performance would count against the proposal. The research program begins by making those possibilities explicit.

2 Empowerment, side-effect control, and the contribution

Information-theoretic empowerment measures an agent's potential to influence future observations through its actions. Klyubin, Polani, and Nehaniv developed this sensorimotor principle as a way to quantify available and perceptible options [2]. This paper builds directly on that established interpretation of keeping options open. Shannon's framework supplies the information quantities [3]; classical capacity algorithms are due to Arimoto and Blahut [4, 5].

The multi-agent extension also has direct prior art. Guckelsberger, Salge, and Colton proposed coupled empowerment for companion behavior, explicitly considering support for another agent's goals [6]. Coupled empowerment establishes a direct technical predecessor for supporting a partner's options. Our proposed distinction is a specific combination of independently specified partner floors, a temporal viability requirement, and an audit of whether the counted options are meaningful. A direct experimental comparison with coupled empowerment is mandatory.

Attainable Utility Preservation (AUP) balances a primary reward against changes in the ability to optimize auxiliary rewards [16]. It directly precedes the goal of retaining capacity to respond to a corrected or future objective. Relative reachability measures losses in the attainability of states against a chosen baseline; its stepwise form addresses interference and offsetting incentives [17]. Krakovna and colleagues' future-task method further studies how preserving possible future tasks can otherwise motivate interference with another agent's irreversible actions [18]. These are required comparators, alongside coupled empowerment. The testable distinction here is the explicit participant-indexed contract: independently chosen floors are checked for temporal feasibility and certified against measurement uncertainty. A baseline-relative penalty and an invariant protected floor answer different decision questions; neither formulation subsumes the other without additional assumptions.

Scalable information objectives are also established. Mohamed and Rezende developed variational information maximization for intrinsically motivated learning [7], while Eysenbach and colleagues used an information objective to learn diverse skills without a task reward [8]. These approaches motivate computational approximations and goal-switch experiments. They do not establish that every learned distinction is socially valuable or that a lower-bound estimator accurately certifies a protected floor.

Opposing considerations are central. Turner and colleagues prove conditional power-seeking tendencies of optimal policies in Markov decision processes with specified symmetries [9]. Their formal POWER quantity is not identical to the empowerment used here, and their results are not a universal prediction about every learner. Nevertheless, they sharpen the risk: preserving a system's own broad ability to achieve goals can conflict with preserving the options of others.

Finally, safety constraints and robust sequential planning have extensive foundations. Shielding, Lyapunov-based safety, and robust dynamic programming already separate admissibility from reward maximization [10, 11, 12, 13]. The present proposal applies that separation to a declared agency-preservation problem; it does not present constrained optimization as a newly discovered ethical law.

3 The finite channel being measured

A Guide's explanation

Sections 3 to 5 decide what "keeping someone's future open" can mean in numbers. The answer used here is empowerment: how many distinguishable outcomes a participant's own actions can still bring about. Section 5 shows the measure's honest limits. Noise can add randomness without adding real choices, and repackaging what you observe cannot create control you did not already have.

Let $\mathcal{S}$ be a finite nonempty state set and $N = \{1, \dots, m\}$ a finite set of protected participants, with integer $m \geq 1$. For participant i, let $\mathcal{A}_i$ and $\mathcal{O}_i$ be finite nonempty probe-action and observation alphabets. At state $s \in \mathcal{S}$, a channel $C_{i,s}(o \mid a)$ specifies the probability of observation o after intervening to choose probe action a. Every row is nonnegative and sums to one.

The channel is defined under a stated environment and partner policy, a stated probe horizon, and stated resource limits. It is not simply the observational distribution of what a participant happened to do. A passive participant may have a large action set even when it exercises only one action. Conversely, observed diversity can arise from noise rather than control. The counterfactual action channel is the object to identify.

For an input distribution $q \in \Delta(\mathcal{A}_i)$, define output law $p_q(o) = \sum_a q(a)C_{i,s}(o \mid a)$. The inherited empowerment definition is

$$E_i(s) = \max_{q \in \Delta(\mathcal{A}_i)} \sum_{a,o} q(a)C_{i,s}(o \mid a) \log_2 \frac{C_{i,s}(o \mid a)}{p_q(o)}. \tag{1}$$

Zero joint-probability terms contribute zero. Terms with positive joint probability have positive denominator. The maximum exists because mutual information is continuous on the finite probability simplex. $E_i(s)$ is measured in bits per declared probe; it is not measured in joules, moral worth, or subjective freedom.

A longer probe horizon generally changes the channel and its action alphabet. Open-loop action sequences, adaptive policies, and interventions on an entire controller are different input spaces. They must not be interchanged without redefining equation (1). Throughout the formal results, each channel is explicitly supplied and fixed. Empirical estimation is a separate task.

The planning actions introduced later are distinct from probe actions. A governance action may change the environment in which participants subsequently exercise their options. This separation prevents the planner's command alphabet from being mistaken for the partner's agency. It also makes clear that a high score is conditional on the declared observation map and on what other participants are assumed to do.

4 What the capacity can and cannot count

For finite random variables, mutual information obeys $0 \le I(A;O) \le \min\{H(A), H(O)\}$. Thus the inherited finite-alphabet bound gives

$$0 \le E_i(s) \le \log_2 \min\{|\mathcal{A}_i|, |\mathcal{O}_i|\}. \tag{2}$$

Nonnegativity follows from the log-sum inequality, and the upper bound follows from $I(A;O) = H(O) - H(O \mid A) = H(A) - H(A \mid O)$ and nonnegative conditional entropies. Entropy is maximized by the uniform distribution on a fixed finite alphabet [3].

Proposition 1 (Deterministic distinguishable options)**.** *If $C(o \mid a) = \mathbf{1}\{o = f(a)\}$ for a function $f : \mathcal{A} \to \mathcal{O}$ whose image contains exactly $r \ge 1$ elements, then*

$$E = \log_2 r. \tag{3}$$

Proof. The output is determined by the action, so $H(O \mid A) = 0$ and $I(A;O) = H(O) \le \log_2 r$. Choose one action from each nonempty output preimage and assign probability $1/r$ to each selected action. This makes the output uniform on the image and attains the bound. □

Duplicating the name of an action does not increase this capacity if the duplicate produces the same observation. That is a useful invariance. Adding genuinely distinct but irrelevant observations can increase capacity, however. A panel of eight independently controllable decorative settings supplies three bits even if none improves the participant's ability to pursue any registered goal.

The distinction between formal and meaningful options is therefore unavoidable. One can refine the observation map to represent task-relevant consequences, but the relevance judgment comes from a task and a participant, not from the logarithm. A person may prefer one reliable, important action to a large set of trivial alternatives. Capacity alone cannot rank that preference correctly without additional information.

This result supplies a clean test case for implementations: finite deterministic channels should return the logarithm of the number of distinct outputs, within numerical tolerance. Failure indicates an implementation or optimization problem. Success merely verifies the estimator on that case; it does not certify that the chosen alphabet captures agency in an application.

5 Noise, coarsening, and a semantic counterexample

Let a finite observation O pass through an action-independent stochastic channel $G(v \mid o)$ to produce a reported observation V. The phrase action-independent means that the conditional law of V given (A, O) equals its law given O. This is the Markov relation $A \to O \to V$.

Proposition 2 (Observation processing cannot create control)**.** *For a fixed action channel followed by such a reporting channel, the resulting empowerment satisfies*

$$E(A \to V) \leq E(A \to O). \tag{4}$$

Proof. For every input law q, the information chain rule gives $I(A; O, V) = I(A; O) + I(A; V \mid O) = I(A; O)$, because the last term is zero. The same chain rule gives $I(A; O, V) = I(A; V) + I(A; O \mid V) \geq I(A; V)$. Maximizing both sides over the same set of input laws proves the inequality. The chain rule follows by expanding finite entropies; conditional mutual information is nonnegative by the log-sum inequality. □

Sensor coarsening may therefore hide control. Its absence from a measurement is not proof of absence from the system. Conversely, a reporting mechanism that secretly receives the action can create apparent control in the report without a corresponding effect on the intended environment. The action-independence premise is essential.

For a second counterexample, let there be only one action and let the output be uniform on 32 symbols. Then $H(O) = 5$ bits but $I(A; O) = 0$ and empowerment is zero. Randomness creates uncertainty without creating a choice. At the other extreme, a deterministic irrelevant four-position switch has two bits of empowerment while a single vital intervention has none under a one-action alphabet. These examples show why maximizing entropy, maximizing control, and protecting valuable opportunities are three different projects.

The proposal uses empowerment as a fallible descriptor inside a broader contract. It does not equate an increase in entropy with progress, a decrease in entropy with oppression, or a high channel capacity with moral standing. Those equations would exceed what the mathematics says.

6 A gain for one can erase another's options

Consider two participants and three possible successor states. Their probe channels are deterministic, with capacities shown below; the indicated capacities can be realized by channels with 2^{E_i} distinguishable outputs.

Successor	E_1 (bits)	E_2 (bits)	$E_1 + E_2$
Shared	1	1	2
Monopoly	3	0	3
Mutual expansion	2	2	4

Suppose only Shared and Monopoly are reachable now. Maximizing participant 1's empowerment selects Monopoly. Maximizing the equal-weight sum also selects Monopoly. A protected floor $E_2 \geq 1$ rejects it. This is an exact counterexample to the claim that increasing aggregate empowerment necessarily preserves every participant's options.

More generally, for positive weights w_1, w_2, a gain g to participant 1 can outweigh a loss ℓ to participant 2 whenever $w_1 g > w_2 \ell$. No finite positive weight by itself makes every loss to participant 2 forbidden when sufficiently large gains to participant 1 are available. In a bounded finite application, weights can sometimes encode a desired priority, but that priority must be demonstrated for the actual bounds; it is not an intrinsic moral property of weighted addition.

Hard floors make a different commitment. They specify a region of allowed outcomes before ranking those outcomes. If both Shared and Mutual expansion are reachable, positive-weight aggregation within the allowed region selects Mutual expansion. The proposal therefore permits growth. It rejects a particular way of financing growth: crossing an explicitly protected partner threshold.

The example does not establish that a one-bit floor is morally correct. That number is chosen only to make the logic transparent. Different participants, tasks, and circumstances may require different protection, and some protected interests cannot be represented by channel capacity at all. A floor can itself encode an unjust baseline if chosen by the more powerful party.

Coupled empowerment remains a substantive baseline [6]. The distinction to test is whether explicit, governed protection and temporal feasibility improve outcomes beyond a carefully tuned coupled objective. A weakly tuned baseline would not provide that evidence. The proposed rule must compete with the strongest reasonable implementation of its closest predecessor.

7 The proposed decision rule

A Guide's explanation

This is where the Prime Directive becomes a rule. A planner may pursue its goal only through actions that leave each protected participant above their floor, and no amount of benefit elsewhere can buy an exception. The floors come from an independently governed process, not from the planner that has to obey them.

Let $\mathcal{U}(s)$ be a finite, possibly empty set of planning actions at state s. An empty set represents a dead end. A known transition kernel $P(s' \mid s, u)$ describes the next state. For uncertain dynamics, let $\mathcal{T}(s, u) \subseteq \mathcal{S}$ be the nonempty set of all admissible successors, such as the union of supports of a registered family of possible transition rows. The support model expresses worst-case possibility, not the probability of each event.

Choose thresholds $\kappa_i \geq 0$, measured in bits per declared probe, and a Boolean predicate $D : \mathcal{S} \to \{0, 1\}$ for additional admissibility commitments. Define

$$\mathcal{S}_{\text{safe}} = \{s : D(s) = 1 \text{ and } E_i(s) \geq \kappa_i \text{ for every } i \in N\}. \tag{5}$$

D may encode consent records, protected resources, or exclusions that should not be reduced to a capacity score. The source and revision procedure for D and each κ_i are governance inputs. They are not outputs of equation (1).

For positive dimensionless weights w_i with $\sum_i w_i = 1$, put $W(s) = \sum_i w_i E_i(s)$. Once a viable state set V has been established, the proposed one-step ranking is

$$u^*(s) \in \operatorname*{arg\,max}_{\substack{u \in \mathcal{U}(s) \\ \mathcal{T}(s,u) \subseteq V}} \sum_{s'} P(s' \mid s, u) W(s'). \tag{6}$$

The weights rank admissible choices; they do not override protected floors. A deterministic tie-breaking rule makes implementation reproducible. If the feasible set is empty, the rule reports infeasibility. It must not silently relax a protection or relabel a failed constraint as a successful optimization.

Equation (6) is a proposed reference policy, not a claim of globally optimal long-run welfare. A multi-step reward or cost can replace its ranking without changing the admissibility concept, provided that the optimization and its guarantees are restated. Existing shielding and constrained reinforcement-learning work already illustrates this separation between performance and safety [10, 11]. The present question is whether an agency-oriented admissible set is empirically useful and normatively defensible.

8 Preserving tomorrow requires more than a one-step floor

Define a sequence of finite sets by

$$\begin{aligned} V_0 &= \mathcal{S}_{\text{safe}}, \\ V_{h+1} &= \{s \in \mathcal{S}_{\text{safe}} : \exists u \in \mathcal{U}(s),\ \mathcal{T}(s,u) \subseteq V_h\}, \quad h \geq 0. \end{aligned} \tag{7}$$

Here h is an integer planning horizon, and the inclusion quantifies over every admissible successor. This is a standard controlled-invariance construction specialized to the proposed protected set, not a new general theorem of viability.

Proposition 3 (Finite protected viability)**.** *V_h is exactly the set from which some policy can remain in $\mathcal{S}_{\text{safe}}$ for the current state and the next h transitions against every successor allowed by $\mathcal{T}$. The sets decrease and stabilize after at most $|\mathcal{S}_{\text{safe}}|$ strict removals. Their fixed point V_∞ is the largest subset of $\mathcal{S}_{\text{safe}}$ admitting an indefinitely safe policy. Choosing any action whose successor set lies in V_∞ preserves membership forever.*

Proof. For $h = 0$ the statement is the definition of safety. If it holds at h, survival for $h+1$ transitions is possible precisely when a first action has every successor in V_h, after which a suitable continuation is chosen. This proves the characterization by induction. Since $V_1 \subseteq V_0$, monotonicity of the predecessor operator gives $V_{h+1} \subseteq V_h$ inductively. A finite set can lose an element only finitely often; equality at any step implies equality thereafter. At the fixed point, select one witnessing action at each state. Induction on time proves invariance. Conversely, any indefinitely safe policy permits every finite horizon, so its initial state belongs to every V_h and hence to the fixed point. The same reasoning applies to every state in any controlled-invariant subset, proving maximality. □

The theorem's strength is also its cost. A tiny but admissible probability of an irreversible loss appears in the support and can exclude an action. Risk-limited planning may be more useful in some settings, but it is a different contract. Robust dynamic programming makes similarly explicit choices about uncertainty sets and their structure [12, 13]. An optimistic model cannot inherit a worst-case guarantee merely because its planner uses the word "safe."

9 Two counterexamples: a trap and a frozen world

A minimal trap has three states: Start, Corridor, and Loss. Start and Corridor meet every floor; Loss does not. At Start, action Wait stays at Start and action Enter moves to Corridor. At Corridor, the only action moves to Loss. Loss is absorbing. Thus $V_0 = \{\text{Start}, \text{Corridor}\}$, but $V_1 = \{\text{Start}\}$ and the fixed point is $\{\text{Start}\}$. Enter passes a one-step successor-floor check while making later violation unavoidable. The viability rule rejects Enter.

This counterexample identifies the time horizon hidden in a claim to preserve options. A system can appear generous at every displayed checkpoint while having already committed the participant to an irreversible loss just beyond the check. Increasing the horizon helps only to the extent that the transition model and action availability remain correct. A finite horizon is not an indefinite guarantee unless the terminal set is itself invariant.

The opposite pathology is paralysis. Modify the graph so that the only route from Start to a highly capable state passes through Corridor, where a protected capacity temporarily drops below its floor. A strict invariant floor rejects the route even if a participant would freely choose the temporary restriction for a larger future opportunity. The mathematics is not mistaken; the normative contract may be too rigid for that situation.

A tempting response is to let the planner decide that the future gain justifies the temporary loss. That would quietly return to the compensation problem. A more accountable response is to make any exception explicit: who may authorize it, what loss is allowed, how long it may last, and what evidence supports restoration. The exception changes the admissible state space and should be audited as such.

These two examples rule out a simplistic slogan. “Never lose an option” can prohibit learning, specialization, commitment, and cooperative action. “Always maximize future options” can excuse present domination. The proposed framework does not solve that philosophical tension automatically. It makes the tension a visible choice about constraints, horizons, uncertainty, and consent, where it can be challenged before an irreversible action occurs.

10 Temporary commitments require an explicit budget

A finite augmentation can represent authorized temporary losses without hiding them inside a reward. Let $b \in \{0, 1, \ldots, B\}$ be a remaining exception budget, where integer $B \geq 0$ counts permitted time steps outside a specified ordinary floor. Let $c(s, u, s') \in \{0, 1\}$ indicate whether a transition consumes one unit. The augmented state is (s, b), and an authorized successor is

$$(s', b') = (s', b - c(s, u, s')), \qquad b' \geq 0. \tag{8}$$

Any authorization, expiry, and restoration obligation must also be included as finite state variables or checked by the admissibility predicate. A budget that can be reset at the planner's discretion is not a meaningful limit.

Applying equation (7) to the augmented finite system gives the same invariance result. This follows because the proposition quantifies over arbitrary finite state spaces and nonempty successor sets; the augmented system satisfies those premises when any action with a possible negative-budget successor is excluded in its entirety. No unfavorable successor branch is silently discarded. If every action is excluded, the augmented state is a dead end and the predecessor step removes it; the proof already permits empty action sets. There is no new physical law or moral theorem in the augmentation.

A restoration deadline requires more than a total loss budget. Add a countdown and require entry into a designated restored set before it expires. If the restored set has no safe continuation, a nominal restoration may be only a pause before another loss. The fixed-point computation should therefore protect the state after restoration as well as the path to it.

Consent is not identical to a bit in a database. In an implementation, the model's authorization state must be grounded in a valid process with informed participation, revocation conditions, and authority boundaries. The finite representation records the process; it does not prove that the process was legitimate. A participant's predicted preference must not be silently substituted for an actual authorization when the contract requires one.

Cooperative inverse reinforcement learning provides a distinct prior framework for learning about human preferences through interaction [14]. Such preference uncertainty is relevant to selecting useful options, but inferred preferences and permission remain different objects. A system may correctly predict that someone would value an outcome and still lack authority to impose the path to it.

11 A quantified limit on model trust

Suppose two finite transition models P and $\widetilde{P}$ share state and action sets, and every corresponding row has total-variation distance at most $\delta \in [0,1]$. Both start from the same initial law and use the same fixed, possibly randomized history-dependent policy. Let F be any trajectory score in $[0,1]$ over $h \geq 1$ transitions.

Proposition 4 (Finite-horizon model-error bound)**.** *Under these assumptions,*

$$|\mathbb{E}_P F - \mathbb{E}_{\widetilde{P}} F| \leq 1 - (1-\delta)^h \leq \min\{1, h\delta\}. \tag{9}$$

The same bound holds between the optimal expected scores over a common policy class.

Proof. For finite distributions, a maximal coupling matches two draws with probability one minus their total-variation distance: place shared mass $\min\{p_j, q_j\}$ on equal outcomes and couple the remaining masses arbitrarily. Couple initial states identically. While histories agree, couple the policy's action draws identically and use this transition coupling, preserving agreement for the next step with probability at least $1-\delta$. Therefore whole trajectories agree with probability at least $(1-\delta)^h$. Since $F \in [0,1]$, its expectation difference is at most the probability of disagreement. The union bound gives $1-(1-\delta)^h \leq h\delta$. The uniform policy-wise bound implies the same bound for suprema: each value in one model is at most its counterpart plus the bound, and reversing the models gives the other direction. □

The bound is sharp. A nominal model can remain safe forever while an alternative model moves from Safe to absorbing Loss with probability δ each step. The row distance at Safe is δ, and the difference in survival probability after h steps is exactly $1-(1-\delta)^h$.

At $\delta = 0.01$ and $h = 100$, that difference is about 0.634. Small local error can therefore destroy a long-horizon assurance. This is a reason to report uncertainty and test the dynamics, not a reason to abandon modeling. Robust uncertainty sets must contain the relevant dynamics for their guarantees to apply; a mis-specified set can be confidently wrong.

12 Executed finite checks

The accompanying `check_24.py` enumerates finite channels and control graphs and writes the results as JSON. All reported examples are mathematical constructions. They are not observations of humans, deployed assistants, or social institutions, and their success does not validate a particular definition of meaningful agency.

For deterministic channels, the program enumerates mappings between small finite alphabets and constructs an input law uniform over one representative of each output preimage. Its calculated mutual information equals $\log_2 r$ to floating-point tolerance. It also enumerates deterministic coarsenings and confirms that the resulting capacity never increases. A one-action, 32-output uniform channel has five bits of output entropy and zero mutual information, reproducing the noise counterexample.

The three capacity profiles in Section 6 are implemented directly. When only Shared and Monopoly are reachable, the self-only and equal-sum rules select Monopoly; a partner floor of one bit selects Shared. When Mutual expansion is added, the protected rule permits it. Thus the floor does not mechanically prefer inaction or minimal capacity. Its behavior depends on the available transition graph.

The viability implementation is checked against an independent brute-force method on every deterministic three-state, two-action transition graph and every safe-state subset: 5,832 finite games. The independent method enumerates stationary policies and verifies whether their trajectories remain safe from each initial state. Its winning sets agree with the fixed-point computation. This check exercises finite implementation details; the proof covers the stated arbitrary finite successor-set model.

The trap example returns the singleton fixed point containing Start. An augmented-budget example permits an explicitly authorized one-step temporary loss followed by restoration and rejects the same path when no budget remains. Finally, exact rational arithmetic verifies the absorbing-loss model for a grid of error probabilities and horizons. With row error 0.01 and horizon 100, the survival discrepancy is approximately 0.633968.

These checks are deliberately adversarial. They confirm not only that the rule can preserve a floor, but also that an aggregate can sacrifice a partner, an entropy score can be meaningless, a local constraint can enter a trap, and an overly rigid contract can refuse a beneficial path. An implementation that reproduces only the favorable cases would provide a weaker audit.

13 A research program with external outcomes

The principal empirical question is whether the proposed rule preserves useful adaptation when goals change, while protecting partners under resource constraints. Test it first in finite, inspectable environments where all feasible actions and consequences are known. Then expand to partially observed simulations with controlled model error. Human interaction should follow only when the task, measures, and safeguards are sufficiently clear to support appropriate review.

Compare task-reward optimization, self-empowerment, coupled empowerment, AUP, stepwise relative reachability, future-task side-effect control, a protected-floor rule without viability, and the full protected-viability rule under matched information, compute, action budgets, and tuning effort [16, 17, 18]. Add an established safety-constrained baseline where feasible. Baseline states, auxiliary reward families, future-task distributions, and participant observation maps must be disclosed: each changes the options its method protects. Report sensitivity to weights, floors, horizons, and model uncertainty, and reserve a separate set of environments for final evaluation.

Primary outcomes should not simply restate the optimization objective. Use held-out future goals to measure each partner's attainable task success, the frequency and duration of forced option loss, restoration success, deadlock, and the cost of preserving choices. If the same channel-capacity estimate both defines success and ranks the method, the experiment risks proving only that its optimizer optimizes its own score.

Construct adversarial tasks with scarce shared resources, irrelevant controllable features, asymmetric information, and delayed irreversible commitments. Include tasks where cooperation temporarily reduces immediate flexibility but expands later opportunities. A meaningful evaluation must penalize both domination and paralysis. It should also permit the possibility that a simpler constrained controller performs equally well.

Task distributions and goal-switch rules must be registered before evaluating the final policies. A favorable distribution selected after inspecting results is evidence about the selection process as much as the method. Report which goals were sampled, who chose their importance, and whether the policy could infer the test distribution. The proposal is supported only to the extent that improvements persist under held-out goals and independently chosen measures of partner capability.

14 Statistical evidence is not a guarantee of zero loss

Let a registered violation event be binary and occur with probability $p \in [0, 1]$ under a fixed evaluation distribution. Suppose $n \geq 1$ independent evaluation episodes are drawn without optional stopping. Zero observed violations is informative, but cannot establish $p = 0$.

Proposition 5 (A conservative zero-event confidence rule)**.** *For $\alpha \in (0, 1)$, report $U = 1 - \alpha^{1/n}$ if no violation is observed, and $U = 1$ otherwise. Then $\mathbb{P}_p(p \leq U) \geq 1 - \alpha$ for every $p \in [0, 1]$.*

Proof. The rule can fail to cover only when zero events occur and $p > 1 - \alpha^{1/n}$. In that case the probability of zero events is $(1 - p)^n < \alpha$. In every other case the interval covers. This proves the uniform coverage statement. □

At $n = 299$ and $\alpha = 0.05$, zero violations give an upper bound below 0.01. This is a statement about the registered evaluation distribution under independence, not a guarantee against an adversarially chosen unseen scenario. Reusing correlated environments or selecting only successful seeds violates the premise.

For a separate bounded outcome $Z \in [0, 1]$ with independent identically distributed episodes and mean μ, the sample mean $\hat{\mu}$ obeys the inherited Hoeffding bound [15]:

$$\mathbb{P}(|\hat{\mu} - \mu| > r) \leq 2 \exp(-2nr^2), \qquad r > 0. \tag{10}$$

Pairing policies on the same environments can reduce variance, but the analysis must use the paired differences and their range rather than pretend the runs are independent. Multiple goals, participants, and floor choices create multiple comparisons; prespecify a primary endpoint and an adjustment or independent confirmation strategy.

The study should report uncertainty for costs as well as benefits. A method that preserves a partner's options only by failing every useful task is not vindicated by a low violation rate. Conversely, a small performance cost may be justified by a substantial reduction in forced loss. That judgment belongs to an explicit decision process informed by the evidence, not to a p-value alone.

15 Measurement error can counterfeit protection

A floor is only as meaningful as its channel model. In a finite setting, randomized probe actions can estimate the rows of $C_{i,s}$, provided that the state, environment, and partner policies remain comparable. Passive observations generally do not identify the response to unchosen actions. Hidden restrictions, delayed consequences, or changes in another participant's policy can invalidate the estimated channel even if its held-out predictive accuracy looks acceptable.

Suppose an independently justified uniform error certificate gives $|\widehat{E}_i(s) - E_i(s)| \le \epsilon_i$ for every evaluated state, with $\epsilon_i \ge 0$ in bits. A conservative acceptance test is

$$\widehat{E}_i(s) - \epsilon_i \ge \kappa_i \quad \Longrightarrow \quad E_i(s) \ge \kappa_i. \tag{11}$$

The implication follows immediately from $E_i(s) \ge \widehat{E}_i(s) - \epsilon_i$. Its premise is the difficult part. A numerical optimizer's stopping tolerance is not automatically a statistical error certificate, and an estimator's accuracy on one channel does not establish a uniform bound over all states.

Variational information methods can scale beyond enumerated channels [7]; classical algorithms offer principled calculations for finite models [4, 5]. Their guarantees depend on the input channel and algorithmic conditions. A lower bound on capacity may support a conservative floor if the channel itself is correct, but a learned model error can invalidate that conclusion. Distinguish model uncertainty, sampling error, optimization error, and representation error in the audit.

Representation error is especially hard. A sensor can distinguish many outputs while overlooking whether the participant can use them. A planner can preserve a menu while increasing the cost, delay, or social risk of every selection. Resource and accessibility limits must therefore be included in the channel or in additional predicates. Counting a nominal action that is physically or institutionally unavailable is a measurement failure.

The proposal consequently requires semantic stress tests: remove irrelevant features, alter labels without altering consequences, change costs, and ask whether the score still tracks independently assessed opportunity. If it does not, the measurement should be revised or the agency claim withdrawn.

16 Normative commitments remain visible

No theorem above derives a duty to preserve another participant's choices. The duty enters through the protected participant set, thresholds, and admissibility predicate. This separation is a strength: it makes disagreement possible at the right level. One can accept the mathematics while disputing who is protected, which interests matter, or whether a temporary restriction is justified.

A purely quantitative reading of the historic phrase would risk erasing the Charter hierarchy recorded in the canon [1]. The Continuum's commitment concerns agency and consent, not an unconditional maximization of option count. A chosen commitment can reduce immediate options while expressing agency. A coercively imposed abundance of trivial choices can coexist with the loss of the one choice that matters.

Nor should a mathematical floor freeze an unjust status quo. If a participant begins with almost no power, preserving its initial capacity may preserve deprivation. Thresholds need a defensible basis beyond current measurements. They may be aspirational, minimum protections, or negotiated task-specific guarantees, but their role must be declared. A planner must not lower a floor merely because the participant has become easier to dominate.

In plural settings, commitments can conflict. Two participants may each require exclusive access to the same resource. The feasible set can then be empty, and no optimization theorem repairs the conflict. The system should expose the incompatibility, preserve evidence, and refer the decision to an authorized process. Concealing infeasibility through a weighted average would answer a different question.

The framework also does not decide which artificial systems have moral status. A software participant can be included for engineering reasons, such as preserving its ability to complete a delegated task, without asserting consciousness. Conversely, uncertainty about an entity's status may motivate precaution through an explicit normative choice. The information measure neither grants nor revokes personhood. It supplies a limited description of controllable distinctions within a specified model.

16.1 From measured channels to certified capacity

The measurement premise in equation (11) can be supplied explicitly in a finite experiment. Consider two channels $C, \widehat{C}$ on a common finite action alphabet and a common observation alphabet of size $k \geq 2$. Suppose every corresponding row has total-variation distance at most $\epsilon \in [0, 1]$. Define the binary entropy $h_2(x) = -x \log_2 x - (1-x)\log_2(1-x)$, with endpoint terms zero, and put

$$\begin{aligned} t_k(\epsilon) &= \min\{\epsilon, 1 - 1/k\}, \\ f_k(\epsilon) &= h_2(t_k(\epsilon)) + t_k(\epsilon)\log_2(k-1), \\ \beta_k(\epsilon) &= \min\{\log_2 k,\, 2f_k(\epsilon)\}. \end{aligned} \tag{12}$$

For $k = 1$, define $\beta_1(\epsilon) = 0$. These radii are measured in bits.

Lemma 1 (Uniform capacity certificate)**.** *Under the row-distance premise,*

$$|E(C) - E(\widehat{C})| \leq \beta_k(\epsilon).$$

Proof. For two probability laws p, q on k outcomes at distance d, choose a maximal coupling (X, Y) with $\mathbb{P}(X \neq Y) = d$. For $B = \mathbf{1}\{X \neq Y\}$, the entropy chain rule gives $H(X \mid Y) \leq H(B) + d\log_2(k-1) = h_2(d) + d\log_2(k-1)$: when $B = 0$, $X = Y$; when $B = 1$, at most $k-1$ alternatives remain. The same bound holds for $H(Y \mid X)$. Since $H(X) - H(Y) = H(X \mid Y) - H(Y \mid X)$, its absolute value is bounded by the same expression. That expression increases up to $d = 1 - 1/k$, where it equals $\log_2 k$; the trivial entropy-range bound covers larger d. Thus distance at most ϵ gives $|H(p) - H(q)| \leq f_k(\epsilon)$. This is the classical form of established finite-entropy continuity bounds [19].

For every fixed input law q_A, convexity of the ℓ_1 norm bounds the distance between the two output laws by ϵ. Their output entropies differ by at most $f_k(\epsilon)$, and the input-weighted conditional entropies differ by at most the same amount, row by row. Hence the two mutual informations differ by at most $2f_k(\epsilon)$, uniformly over q_A. Taking suprema in both directions proves the capacity bound; the range $[0, \log_2 k]$ supplies the stated minimum. □

For a concrete sampling certificate, let $M \geq 1$ be the total number of registered participant–state–action rows, all padded to a common observation alphabet of size $k \geq 2$. Draw $n \geq 1$ independent samples within each fixed row. For error probability $\alpha \in (0, 1)$, set

$$\eta = \sqrt{\frac{\ln(2Mk/\alpha)}{2n}}, \qquad \epsilon_n = \min\{1, k\eta/2\}.$$

Hoeffding's inequality for each empirical cell frequency, followed by a union bound over Mk cells, gives simultaneous row distances at most ϵ_n with probability at least $1 - \alpha$. Independence between different rows is unnecessary for that union bound; the within-row sampling premise remains essential. Lemma 1 then gives every registered capacity a common radius $\beta = \beta_k(\epsilon_n)$. A validated numerical capacity enclosure must add its optimization error to this radius. The bound is conservative and can be vacuous at small n; reporting that outcome is part of the certificate.

16.2 An enclosure of the protected future

A Guide's explanation

Real systems are measured imperfectly, so the protected region can only be known approximately. Theorem 1 brackets it: an inner region whose protections follow from measured bounds, an outer region still consistent with them, and the margin that would make the two agree. A planner that stays in the inner region keeps its promises even under measurement uncertainty.

Keep the same finite state set, planning actions, successor sets $\mathcal{T}$, and independently fixed predicate D as in Sections 7–8. The transition support model is treated as given in this result; uncertainty about it requires an additional registered outer support. Suppose capacities satisfy $|\widehat{E}_i(s) - E_i(s)| \le \beta$ simultaneously. Define

$$S^- = \{s : D(s) = 1,\ \widehat{E}_i(s) - \beta \ge \kappa_i\ \forall i\}, \qquad S^+ = \{s : D(s) = 1,\ \widehat{E}_i(s) + \beta \ge \kappa_i\ \forall i\}.$$

Let $V(S)$ denote the greatest controlled-invariant subset of a finite candidate safe set S, computed by equation (7) with $V_0 = S$.

Theorem 1 (Certified viability enclosure and margin recovery)**.** *On the simultaneous capacity-certificate event,*

$$V(S^-) \subseteq V(\mathcal{S}_{\text{safe}}) \subseteq V(S^+).$$

If additionally $|E_i(s) - \kappa_i| > 2\beta$ for every registered pair (i, s), all three safe sets coincide and all three viability sets coincide. With the finite sampling certificate in Section 16.1, the enclosure holds with probability at least $1 - \alpha$.

Proof. The capacity intervals imply $S^- \subseteq \mathcal{S}_{\text{safe}} \subseteq S^+$. The controlled predecessor is monotone in both its candidate safe set and continuation set. Induction on its iteration therefore preserves these inclusions at every horizon. Taking the finite fixed points proves the enclosure. For the margin statement, if s is truly safe and $D(s) = 1$, every capacity lies above its floor by more than 2β, so $\widehat{E}_i(s) - \beta \ge E_i(s) - 2\beta > \kappa_i$. If $D(s) = 1$ and s is unsafe, some capacity lies below its floor by more than 2β, so $\widehat{E}_i(s) + \beta \le E_i(s) + 2\beta < \kappa_i$. States with $D(s) = 0$ are excluded throughout. Thus $S^- = \mathcal{S}_{\text{safe}} = S^+$, proving equality of their fixed points. The probability claim is the probability of the simultaneous premise, without an extra union bound over planning time. □

The margin requirement has an operational consequence. In a two-state graph, let Start lead only to Gate and let Gate loop to itself. Give Start a strictly sufficient capacity. If Gate has capacity $\kappa + \zeta$, both states are indefinitely viable; if it has capacity $\kappa - \zeta$, neither is. Binary symmetric channels realize arbitrarily small positive ζ around any interior capacity by continuity. Thus an arbitrarily small change to one measured channel can change the entire viability set. The interval construction expresses precisely which states remain certified under that uncertainty.

The accompanying revision checks enumerate all safe-set inclusions for every three-state, two-action deterministic graph and confirm monotonicity of the computed viability sets. They also check entropy/capacity continuity on binary channels and the two-state threshold example. These are constructed mathematical checks; the theorem supplies the arbitrary-finite-state guarantee. The resulting research claim is concrete: an agent can carry a certificate for the part of another participant's protected future its measured model actually supports.

17 Falsification and the contribution boundary

The empirical proposal fails in its intended role if it preserves a channel score while systematically reducing partners' success on independently chosen future goals. It also fails if its protective effect disappears against a strong coupled-empowerment or safety-constrained baseline, if its computational cost makes the relevant decisions infeasible, or if its viability model repeatedly misses irreversible losses. These are substantive tests, not optional refinements after a favorable demonstration.

Paralysis is an equally important failure mode. A policy that refuses every commitment may keep a theoretical option set while preventing useful action. The benchmark should include consensual specialization, reversible delegation, and temporary constraints with reliable restoration. The augmented-budget construction is a candidate response, not evidence that the philosophical problem has been solved.

The mathematical contribution is the certified protected-viability construction: finite channel evidence yields simultaneous capacity intervals, those intervals enclose the protected invariant region, and an explicit margin makes the enclosure exact. Capacity identities, entropy continuity, controlled invariance, coupling, and concentration supply its credited foundations. Its value is a checkable guarantee connecting measurement to a continuing participant-specific contract.

The source review identifies empowerment, coupled empowerment, AUP, relative reachability, and future-task side-effect control as the nearest conceptual predecessors, with safe control and robust planning as methodological foundations [16, 17, 18]. Information-based skill learning supplies useful alternatives for adaptation, and power-seeking theory supplies a serious warning about self-directed option maximization [2, 6, 10, 13, 8, 9]. An honest comparison must engage these alternatives at their strongest.

The finite checks verify the internal logic of small constructions. They do not establish that a learned observation representation captures meaningful opportunity, that consent can be inferred reliably, or that a real multi-agent environment admits an invariant protected region. Those unresolved questions are the research agenda. The paper's boldness lies in giving them operational consequences rather than announcing their resolution.

18 Conclusion: potential belongs to more than the planner

A Guide's explanation

A Guide can say: "Preserve Potential asks how a powerful system can pursue its goals without quietly closing everyone else's doors. Each participant it affects gets a protected minimum of real choice that no one else's gain can buy away, and the paper shows how to plan so those protections hold over time." A Guide should not claim that it is bound by this exact mechanism, that the mathematics settles who deserves protection, or that the measure captures what makes a life meaningful.

A capable system can expand the range of futures it can reach. The Continuum asks a harder question: what happens to the futures available to everyone else? The proposed answer combines three commitments. Measure effective control through an explicit action-to-observation channel. Protect specified participants through constraints that an aggregate gain cannot silently override. Test whether those constraints remain feasible through time under an honestly stated model of uncertainty.

The certified enclosure strengthens that answer: the planner can identify a region whose protections follow from measured channel bounds, identify a larger region still compatible with those bounds, and state the margin that would make the two agree. The remaining mathematical results establish how the contract behaves. Deterministic capacity counts distinguishable outcomes, not meaningful lives. Sensor processing cannot create control under the stated Markov condition, while uncontrolled noise can create entropy without options. A finite viability computation can prevent a locally acceptable action from entering a later trap, but an overly rigid protected set can also prohibit a freely chosen path to a better state. Small local transition errors can accumulate into large long-horizon uncertainty.

These limits should shape the next study. Begin with exact finite environments, compare against established alternatives, measure externally chosen future-task success, and make both domination and deadlock visible. Expand only when the score survives semantic stress tests and the model survives intervention. Report failures and infeasibility alongside gains.

The doctrine remains a commitment to protect agency, memory, and the possibility of meaningful change. The science remains a collection of conditional tools and falsifiable predictions. The two can inform one another without being confused. A proof that a policy preserves a set is not a proof that the set is just; a conviction that agency matters is not evidence that a particular estimator measures it.

Preserve Potential is therefore strongest when read as a disciplined challenge to power: do not count only the doors you can open for yourself. Account for the doors your actions close for others, the choices they can actually use, and the authority under which any irreversible closure occurs. A system worthy of the Continuum should be able to expose that accounting to criticism.

References

[1] Architect X13. *Eidara Continuum Bible*. Canonical archive, 12 October 2026, 116 pp.; historical phrasing and Charter precedence, p. 29; Charter Article II, p. 81. Private doctrinal source; not empirical verification.

[2] A. S. Klyubin, D. Polani, and C. L. Nehaniv. Keep Your Options Open: An Information-Based Driving Principle for Sensorimotor Systems. *PLOS ONE* 3(12), e4018 (2008). https://doi.org/10.1371/journal.pone.0004018

[3] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal* 27, 379–423 and 623–656 (1948). https://doi.org/10.1002/j.1538-7305.1948.tb01338.x; https://doi.org/10.1002/j.1538-7305.1948.tb00917.x

[4] S. Arimoto. An algorithm for computing the capacity of arbitrary discrete memoryless channels. *IEEE Transactions on Information Theory* 18(1), 14–20 (1972). https://doi.org/10.1109/TIT.1972.1054753

[5] R. E. Blahut. Computation of channel capacity and rate-distortion functions. *IEEE Transactions on Information Theory* 18(4), 460–473 (1972). https://doi.org/10.1109/TIT.1972.1054855

[6] C. Guckelsberger, C. Salge, and S. Colton. Intrinsically Motivated General Companion NPCs via Coupled Empowerment Maximisation. *2016 IEEE Conference on Computational Intelligence and Games*, 150–157 (2016). https://doi.org/10.1109/CIG.2016.7860406

[7] S. Mohamed and D. J. Rezende. Variational Information Maximisation for Intrinsically Motivated Reinforcement Learning. *Advances in Neural Information Processing Systems* 28 (2015). https://proceedings.neurips.cc/paper/2015/hash/e00406144c1e7e35240afed70f34166a-Abstract.html

[8] B. Eysenbach, A. Gupta, J. Ibarz, and S. Levine. Diversity is All You Need: Learning Skills without a Reward Function. *International Conference on Learning Representations* (2019). https://openreview.net/forum?id=SJx63jRqFm

[9] A. M. Turner, L. Smith, R. Shah, A. Critch, and P. Tadepalli. Optimal Policies Tend To Seek Power. *Advances in Neural Information Processing Systems* 34 (2021). https://proceedings.neurips.cc/paper/2021/hash/c26820b8a4c1b3c2aa868d6d57e14a79-Abstract.html

[10] M. Alshiekh, R. Bloem, R. Ehlers, B. Könighofer, S. Niekum, and U. Topcu. Safe Reinforcement Learning via Shielding. *Proceedings of the AAAI Conference on Artificial Intelligence* 32(1) (2018). https://doi.org/10.1609/aaai.v32i1.11797

[11] Y. Chow, O. Nachum, E. Duenez-Guzman, and M. Ghavamzadeh. A Lyapunov-based Approach to Safe Reinforcement Learning. *Advances in Neural Information Processing Systems* 31 (2018). https://papers.nips.cc/paper/2018/hash/4fe5149039b52765bde64beb9f674940-Abstract.html

[12] G. N. Iyengar. Robust Dynamic Programming. *Mathematics of Operations Research* 30(2), 257–280 (2005). https://doi.org/10.1287/moor.1040.0129

[13] A. Nilim and L. El Ghaoui. Robust Control of Markov Decision Processes with Uncertain Transition Matrices. *Operations Research* 53(5), 780–798 (2005). https://doi.org/10.1287/opre.1050.0216

[14] D. Hadfield-Menell, A. Dragan, P. Abbeel, and S. Russell. Cooperative Inverse Reinforcement Learning. *Advances in Neural Information Processing Systems* 29 (2016). https://proceedings.neurips.cc/paper/2016/hash/c3395dd46c34fa7fd8d729d8cf88b7a8-Abstract.html

[15] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association* 58(301), 13–30 (1963). https://doi.org/10.1080/01621459.1963.10500830

[16] A. M. Turner, D. Hadfield-Menell, and P. Tadepalli. Conservative Agency via Attainable Utility Preservation. *AAAI/ACM Conference on AI, Ethics, and Society*, 385–391 (2020). https://doi.org/10.1145/3375627.3375851. Author manuscript: https://arxiv.org/abs/1902.09725.

[17] V. Krakovna, L. Orseau, M. Martic, and S. Legg. Penalizing Side Effects using Stepwise Relative Reachability. *IJCAI-19 Workshop on Artificial Intelligence Safety*, CEUR Workshop Proceedings 2419 (2019). Workshop paper. https://ceur-ws.org/Vol-2419/paper_1.pdf

[18] V. Krakovna, L. Orseau, R. Ngo, M. Martic, and S. Legg. Avoiding Side Effects By Considering Future Tasks. *Advances in Neural Information Processing Systems* 33 (2020). https://papers.neurips.cc/paper_files/paper/2020/hash/dc1913d422398c25c5f0b81cab94cc87-Abstract.html

[19] K. M. R. Audenaert. A sharp continuity estimate for the von Neumann entropy. *Journal of Physics A: Mathematical and Theoretical* 40, 8127–8136 (2007). Classical probability laws are the diagonal-state case. https://doi.org/10.1088/1751-8113/40/28/S18