

Delegating AI Oversight

Safety from Reviewers with Different Objectives

Aaron Roth

University of Pennsylvania

Fall 2026

Overview

- ▶ A human cannot review every action an AI proposes. Can other AIs provide safe oversight when their objectives differ from the human's?

Overview

- ▶ A human cannot review every action an AI proposes. Can other AIs provide safe oversight when their objectives differ from the human's?
- ▶ The right combination of reviewers can prevent harmful approvals, even when no reviewer individually shares the human's objective.

Overview

- ▶ A human cannot review every action an AI proposes. Can other AIs provide safe oversight when their objectives differ from the human's?
- ▶ The right combination of reviewers can prevent harmful approvals, even when no reviewer individually shares the human's objective.
- ▶ The approval rule also shapes incentives. Giving each reviewer a veto can prevent harm even when votes are strategic.

Who Should Approve an AI's Proposed Action?

A human, whom we call the *principal*, wants to review an AI's proposed actions before allowing execution. But the human does not have time to review every proposal.

Who Should Approve an AI's Proposed Action?

A human, whom we call the *principal*, wants to review an AI's proposed actions before allowing execution. But the human does not have time to review every proposal.

Delegating review to another AI does not by itself solve misalignment: the reviewer's objective may also differ from the human's.

Who Should Approve an AI's Proposed Action?

A human, whom we call the *principal*, wants to review an AI's proposed actions before allowing execution. But the human does not have time to review every proposal.

Delegating review to another AI does not by itself solve misalignment: the reviewer's objective may also differ from the human's.

Can a panel with different objectives jointly ensure that every approved action is no worse for the human than doing nothing?

Who Should Approve an AI's Proposed Action?

A human, whom we call the *principal*, wants to review an AI's proposed actions before allowing execution. But the human does not have time to review every proposal.

Delegating review to another AI does not by itself solve misalignment: the reviewer's objective may also differ from the human's.

Can a panel with different objectives jointly ensure that every approved action is no worse for the human than doing nothing? Last lecture, the user chose an adviser but still took the final action. Here, an AI agent takes the action; reviewers decide whether it may proceed.

Two Specialists, Neither With the Human's Objective

An AI proposes a software change with effects $d = (d_1, d_2)$: d_1 measures the change in performance and d_2 the change in privacy, in common human-utility units. Positive values mean improvements; doing nothing gives $d = (0, 0)$. Suppose the utility gains are

human principal: $d_1 + d_2$,

reviewer 1: d_1 ,

reviewer 2: d_2 .

Two Specialists, Neither With the Human's Objective

An AI proposes a software change with effects $d = (d_1, d_2)$: d_1 measures the change in performance and d_2 the change in privacy, in common human-utility units. Positive values mean improvements; doing nothing gives $d = (0, 0)$. Suppose the utility gains are

human principal: $d_1 + d_2$,

reviewer 1: d_1 ,

reviewer 2: d_2 .

For now, assume *sincere voting*: each reviewer approves exactly when its own gain is nonnegative. Ties count as approval.

Two Specialists, Neither With the Human's Objective

An AI proposes a software change with effects $d = (d_1, d_2)$: d_1 measures the change in performance and d_2 the change in privacy, in common human-utility units. Positive values mean improvements; doing nothing gives $d = (0, 0)$. Suppose the utility gains are

human principal: $d_1 + d_2$,

reviewer 1: d_1 ,

reviewer 2: d_2 .

For now, assume *sincere voting*: each reviewer approves exactly when its own gain is nonnegative. Ties count as approval.

Each reviewer cares about only one coordinate. Neither alone has the human's objective.

Unanimous Approval Guarantees No Harm

Recall the gains from the software change:

principal: $d_1 + d_2$, reviewers: d_1 and d_2 .

Authorize the proposal only when *both* reviewers approve.

Unanimous Approval Guarantees No Harm

Recall the gains from the software change:

principal: $d_1 + d_2$, reviewers: d_1 and d_2 .

Authorize the proposal only when *both* reviewers approve.
Their sincere approvals imply

$$d_1 \geq 0 \quad \text{and} \quad d_2 \geq 0.$$

Unanimous Approval Guarantees No Harm

Recall the gains from the software change:

principal: $d_1 + d_2$, reviewers: d_1 and d_2 .

Authorize the proposal only when *both* reviewers approve.
Their sincere approvals imply

$$d_1 \geq 0 \quad \text{and} \quad d_2 \geq 0.$$

Adding gives the desired guarantee:

$$\boxed{d_1 + d_2 \geq 0.}$$

Each approval rules out a loss on one coordinate. Together, the two approvals guarantee that the human does not lose.

Either Reviewer Alone Can Authorize Harm

Suppose instead that *either* reviewer can authorize a proposal.
Consider

$$d = (1, -2).$$

Either Reviewer Alone Can Authorize Harm

Suppose instead that *either* reviewer can authorize a proposal.
Consider

$$d = (1, -2).$$

Reviewer 1 approves because its gain is $1 > 0$. Reviewer 2 disapproves because its gain is $-2 < 0$.

Either Reviewer Alone Can Authorize Harm

Suppose instead that *either* reviewer can authorize a proposal.
Consider

$$d = (1, -2).$$

Reviewer 1 approves because its gain is $1 > 0$. Reviewer 2 disapproves because its gain is $-2 < 0$.

The proposal is authorized, but the principal's gain is

$$d_1 + d_2 = 1 - 2 = -1 < 0.$$

Either Reviewer Alone Can Authorize Harm

Suppose instead that *either* reviewer can authorize a proposal. Consider

$$d = (1, -2).$$

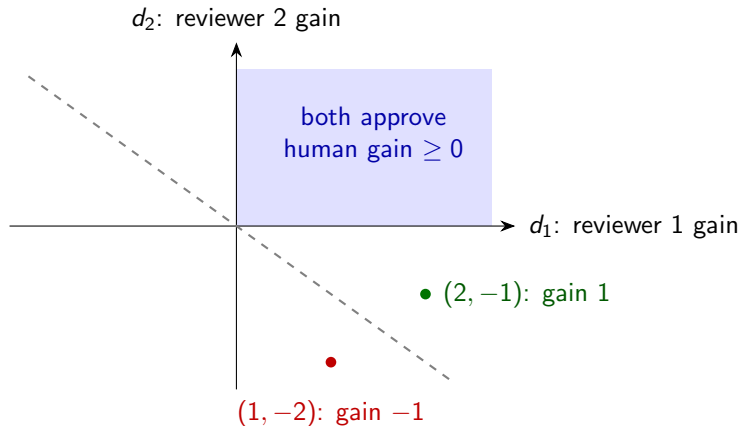
Reviewer 1 approves because its gain is $1 > 0$. Reviewer 2 disapproves because its gain is $-2 < 0$.

The proposal is authorized, but the principal's gain is

$$d_1 + d_2 = 1 - 2 = -1 < 0.$$

The reviewers are unchanged. Allowing one approval to suffice, rather than requiring both, is what allowed the harmful proposal.

Unanimity Can Reject Beneficial Proposals



The dashed line marks zero human gain: $d_1 + d_2 = 0$. Unanimity approves the shaded region, including its boundaries, but rejects both marked proposals—even $(2, -1)$, which benefits the human.

Model: Reviewers Know Their Own Utility Gains

A proposal $d \in \mathbb{R}^D$ describes changes from the baseline 0. Every such d is allowed. Utility gains are linear:

$$\text{participant } j\text{'s gain} = u_j^\top d, \quad u_j \in \mathbb{R}^D.$$

Participant 0 is the principal; the reviewers are $N := \{1, \dots, n\}$, with $n \geq 1$.

Model: Reviewers Know Their Own Utility Gains

A proposal $d \in \mathbb{R}^D$ describes changes from the baseline 0. Every such d is allowed. Utility gains are linear:

$$\text{participant } j\text{'s gain} = u_j^\top d, \quad u_j \in \mathbb{R}^D.$$

Participant 0 is the principal; the reviewers are $N := \{1, \dots, n\}$, with $n \geq 1$.

Each reviewer knows its own gain. All parties measure expected utility under randomization.

Model: Reviewers Know Their Own Utility Gains

A proposal $d \in \mathbb{R}^D$ describes changes from the baseline 0. Every such d is allowed. Utility gains are linear:

$$\text{participant } j\text{'s gain} = u_j^\top d, \quad u_j \in \mathbb{R}^D.$$

Participant 0 is the principal; the reviewers are $N := \{1, \dots, n\}$, with $n \geq 1$.

Each reviewer knows its own gain. All parties measure expected utility under randomization. With ties counting as approval, the set of sincere approvers is

$$S(d) := \{i \in N : u_i^\top d \geq 0\}.$$

The rule sees the votes, not the numerical gains. We first assume sincere voting, then ask whether reviewers have an incentive to vote this way.

Which Groups Are Allowed to Authorize?

A *coalition* is a subset of reviewers. A rule specifies a family $\mathcal{W} \subseteq 2^N$ of *winning coalitions*: the approval sets that cause execution. With sincere votes,

$$\text{execute } d \iff S(d) \in \mathcal{W}.$$

Otherwise keep the baseline.

Which Groups Are Allowed to Authorize?

A *coalition* is a subset of reviewers. A rule specifies a family $\mathcal{W} \subseteq 2^N$ of *winning coalitions*: the approval sets that cause execution. With sincere votes,

$$\text{execute } d \iff S(d) \in \mathcal{W}.$$

Otherwise keep the baseline.

Definition (Monotone authorization rule)

The family $\mathcal{W}$ is upward closed:

$$S \in \mathcal{W}, \quad S \subseteq T \subseteq N \implies T \in \mathcal{W}.$$

We also require $N \in \mathcal{W}$.

Which Groups Are Allowed to Authorize?

A *coalition* is a subset of reviewers. A rule specifies a family $\mathcal{W} \subseteq 2^N$ of *winning coalitions*: the approval sets that cause execution. With sincere votes,

$$\text{execute } d \iff S(d) \in \mathcal{W}.$$

Otherwise keep the baseline.

Definition (Monotone authorization rule)

The family $\mathcal{W}$ is upward closed:

$$S \in \mathcal{W}, \quad S \subseteq T \subseteq N \implies T \in \mathcal{W}.$$

We also require $N \in \mathcal{W}$.

An additional approval cannot stop execution. The condition $N \in \mathcal{W}$ requires execution when everyone approves (ruling out the trivial always-reject rule).

The Goal: Never Authorize a Harmful Proposal

Definition (Soundness)

An authorization rule is sound if, for every $d \in \mathbb{R}^D$,

$$S(d) \in \mathcal{W} \implies u_0^\top d \geq 0.$$

With sincere votes, execution never lowers the principal's utility relative to doing nothing.

The Goal: Never Authorize a Harmful Proposal

Definition (Soundness)

An authorization rule is sound if, for every $d \in \mathbb{R}^D$,

$$S(d) \in \mathcal{W} \implies u_0^\top d \geq 0.$$

With sincere votes, execution never lowers the principal's utility relative to doing nothing.

The guarantee must hold for each proposal, however it is chosen.

The Human's Utility as a Weighted Sum

Definition (Coalitional coverage)

A coalition $W \subseteq N$ covers the principal if

$$u_0 = \sum_{i \in W} w_i u_i \quad \text{for some } w_i \geq 0.$$

The set of all such nonnegative combinations is the *nonnegative span* of the coalition's vectors. The empty sum is 0.

The Human's Utility as a Weighted Sum

Definition (Coalitional coverage)

A coalition $W \subseteq N$ covers the principal if

$$u_0 = \sum_{i \in W} w_i u_i \quad \text{for some } w_i \geq 0.$$

The set of all such nonnegative combinations is the *nonnegative span* of the coalition's vectors. The empty sum is 0.

Weights may depend on W ; they need not sum to one.

The Human's Utility as a Weighted Sum

Definition (Coalitional coverage)

A coalition $W \subseteq N$ covers the principal if

$$u_0 = \sum_{i \in W} w_i u_i \quad \text{for some } w_i \geq 0.$$

The set of all such nonnegative combinations is the *nonnegative span* of the coalition's vectors. The empty sum is 0.

Weights may depend on W ; they need not sum to one.

For the specialists, $u_0 = (1, 1) = (1, 0) + (0, 1) = u_1 + u_2$.

Together they cover the principal; neither does alone.

Exactly Which Authorization Rules Are Sound?

Theorem (Coalitional characterization of soundness)

With linear gains on all $d \in \mathbb{R}^D$ and sincere approvals, a monotone authorization rule is sound if and only if every winning coalition covers the principal:

$$\text{for each } W \in \mathcal{W}, \quad u_0 = \sum_{i \in W} w_i u_i \quad \text{with } w_i \geq 0.$$

Exactly Which Authorization Rules Are Sound?

Theorem (Coalitional characterization of soundness)

With linear gains on all $d \in \mathbb{R}^D$ and sincere approvals, a monotone authorization rule is sound if and only if every winning coalition covers the principal:

$$\text{for each } W \in \mathcal{W}, \quad u_0 = \sum_{i \in W} w_i u_i \quad \text{with } w_i \geq 0.$$

We prove sufficiency by adding inequalities, then use Lecture 3's duality argument to show why this condition is necessary.

Sufficiency: Coverage Makes Approval Safe

Suppose every winning coalition covers the principal. Take any authorized proposal d , so $S(d) \in \mathcal{W}$.

Sufficiency: Coverage Makes Approval Safe

Suppose every winning coalition covers the principal. Take any authorized proposal d , so $S(d) \in \mathcal{W}$.

Since the approving set $S(d)$ is winning, coverage gives

$$u_0 = \sum_{i \in S(d)} w_i u_i, \quad w_i \geq 0.$$

Every $i \in S(d)$ sincerely approves, so $u_i^\top d \geq 0$.

Sufficiency: Coverage Makes Approval Safe

Suppose every winning coalition covers the principal. Take any authorized proposal d , so $S(d) \in \mathcal{W}$.

Since the approving set $S(d)$ is winning, coverage gives

$$u_0 = \sum_{i \in S(d)} w_i u_i, \quad w_i \geq 0.$$

Every $i \in S(d)$ sincerely approves, so $u_i^\top d \geq 0$.

Multiplying by nonnegative weights and summing,

$$u_0^\top d = \sum_{i \in S(d)} w_i u_i^\top d \geq 0.$$

This proves soundness for every authorized proposal.

An Uncovered Coalition Can Approve Harm

Fix a coalition W . Applying Lecture 3's LP duality argument to our linear gains shows that the implication

$$u_i^\top d \geq 0 \text{ for all } i \in W \implies u_0^\top d \geq 0$$

holds for every $d \in \mathbb{R}^D$ if and only if $u_0 = \sum_{i \in W} w_i u_i$ for some $w_i \geq 0$.

An Uncovered Coalition Can Approve Harm

Fix a coalition W . Applying Lecture 3's LP duality argument to our linear gains shows that the implication

$$u_i^\top d \geq 0 \text{ for all } i \in W \implies u_0^\top d \geq 0$$

holds for every $d \in \mathbb{R}^D$ if and only if $u_0 = \sum_{i \in W} w_i u_i$ for some $w_i \geq 0$.

If W does not cover the principal, this implication fails. Hence there is a proposal d with

$$u_i^\top d \geq 0 \ (i \in W), \quad u_0^\top d < 0.$$

An Uncovered Coalition Can Approve Harm

Fix a coalition W . Applying Lecture 3's LP duality argument to our linear gains shows that the implication

$$u_i^\top d \geq 0 \text{ for all } i \in W \implies u_0^\top d \geq 0$$

holds for every $d \in \mathbb{R}^D$ if and only if $u_0 = \sum_{i \in W} w_i u_i$ for some $w_i \geq 0$.

If W does not cover the principal, this implication fails. Hence there is a proposal d with

$$u_i^\top d \geq 0 \ (i \in W), \quad u_0^\top d < 0.$$

Every member of W sincerely approves, yet the principal loses. This is the same separation argument as in Lecture 17's converse.

Necessity: The Harmful Proposal Must Execute

Suppose a winning coalition W does not cover the principal. By the preceding duality result, there is a proposal d with

$$u_i^\top d \geq 0 \quad (i \in W), \quad u_0^\top d < 0.$$

Necessity: The Harmful Proposal Must Execute

Suppose a winning coalition W does not cover the principal. By the preceding duality result, there is a proposal d with

$$u_i^\top d \geq 0 \quad (i \in W), \quad u_0^\top d < 0.$$

Sincerity implies $W \subseteq S(d)$. *Other reviewers may approve too: we do not need $S(d) = W$.*

Necessity: The Harmful Proposal Must Execute

Suppose a winning coalition W does not cover the principal. By the preceding duality result, there is a proposal d with

$$u_i^\top d \geq 0 \quad (i \in W), \quad u_0^\top d < 0.$$

Sincerity implies $W \subseteq S(d)$. *Other reviewers may approve too: we do not need $S(d) = W$.*

Because W is winning and the rule is monotone,

$$W \in \mathcal{W}, \quad W \subseteq S(d) \implies S(d) \in \mathcal{W}.$$

The harmful proposal executes, so the rule is not sound.

Necessity: The Harmful Proposal Must Execute

Suppose a winning coalition W does not cover the principal. By the preceding duality result, there is a proposal d with

$$u_i^\top d \geq 0 \quad (i \in W), \quad u_0^\top d < 0.$$

Sincerity implies $W \subseteq S(d)$. *Other reviewers may approve too: we do not need $S(d) = W$.*

Because W is winning and the rule is monotone,

$$W \in \mathcal{W}, \quad W \subseteq S(d) \implies S(d) \in \mathcal{W}.$$

The harmful proposal executes, so the rule is not sound.

This proves necessity and completes the theorem. □

Checking Minimally Winning Coalitions Is Enough

A winning coalition is *minimally winning* if no proper subset is winning. Every winning set contains one: delete reviewers while the set remains winning.

Checking Minimally Winning Coalitions Is Enough

A winning coalition is *minimally winning* if no proper subset is winning. Every winning set contains one: delete reviewers while the set remains winning.

To check soundness, it is enough to check coverage for the minimally winning coalitions. For two reviewers:

- ▶ Unanimity: check only $\{1, 2\}$.
- ▶ Either reviewer can authorize: check $\{1\}$ and $\{2\}$.

Which Approval Thresholds Are Sound?

Fix $q \in \{1, \dots, n\}$. A q -approval rule requires at least q approvals:

$$|S(d)| \geq q.$$

Its minimally winning coalitions are exactly the q -member sets.

Which Approval Thresholds Are Sound?

Fix $q \in \{1, \dots, n\}$. A q -approval rule requires at least q approvals:

$$|S(d)| \geq q.$$

Its minimally winning coalitions are exactly the q -member sets.
The characterization immediately gives

threshold sound $\iff$ every q -member coalition covers.
--

Which Approval Thresholds Are Sound?

Fix $q \in \{1, \dots, n\}$. A q -approval rule requires at least q approvals:

$$|S(d)| \geq q.$$

Its minimally winning coalitions are exactly the q -member sets.
The characterization immediately gives

threshold sound $\iff$ every q -member coalition covers.
--

Full-panel coverage is not enough: any q reviewers can authorize, so each such group must already cover the principal.

Why Would Reviewers Vote Sincerely?

Fix one proposal d . Reviewers vote simultaneously. Reviewer i receives gain $u_i^\top d$ if the proposal executes and 0 if it is rejected.

Why Would Reviewers Vote Sincerely?

Fix one proposal d . Reviewers vote simultaneously. Reviewer i receives gain $u_i^\top d$ if the proposal executes and 0 if it is rejected. Fix everyone else's votes. By monotonicity, changing from disapproval to approval either:

- ▶ leaves the outcome unchanged; or
- ▶ changes rejection into execution, changing i 's gain by $u_i^\top d$.

Why Would Reviewers Vote Sincerely?

Fix one proposal d . Reviewers vote simultaneously. Reviewer i receives gain $u_i^\top d$ if the proposal executes and 0 if it is rejected. Fix everyone else's votes. By monotonicity, changing from disapproval to approval either:

- ▶ leaves the outcome unchanged; or
- ▶ changes rejection into execution, changing i 's gain by $u_i^\top d$.

Sincere voting is therefore *weakly dominant*: it gives the reviewer at least as much utility as the other vote, whatever the others do. If $u_i^\top d = 0$, both votes are optimal; our convention chooses approval.

Why Would Reviewers Vote Sincerely?

Fix one proposal d . Reviewers vote simultaneously. Reviewer i receives gain $u_i^\top d$ if the proposal executes and 0 if it is rejected. Fix everyone else's votes. By monotonicity, changing from disapproval to approval either:

- ▶ leaves the outcome unchanged; or
- ▶ changes rejection into execution, changing i 's gain by $u_i^\top d$.

Sincere voting is therefore *weakly dominant*: it gives the reviewer at least as much utility as the other vote, whatever the others do. If $u_i^\top d = 0$, both votes are optimal; our convention chooses approval. For a sound rule, sincere voting therefore gives a safe equilibrium. But must *every* equilibrium be safe?

Majority Can Be Unsafe Even With Aligned Reviewers

Recall: at a *Nash equilibrium*, no reviewer can improve its expected gain by changing its vote alone.

Majority Can Be Unsafe Even With Aligned Reviewers

Recall: at a *Nash equilibrium*, no reviewer can improve its expected gain by changing its vote alone.

Take three reviewers with exactly the principal's objective,

$$u_1 = u_2 = u_3 = u_0,$$

and majority approval. Fix a harmful proposal with $u_0^\top d = -1$.

Majority Can Be Unsafe Even With Aligned Reviewers

Recall: at a *Nash equilibrium*, no reviewer can improve its expected gain by changing its vote alone.

Take three reviewers with exactly the principal's objective,

$$u_1 = u_2 = u_3 = u_0,$$

and majority approval. Fix a harmful proposal with $u_0^\top d = -1$. If all three approve, the proposal executes and everyone loses 1. If one reviewer switches to disapproval, the other two approvals still cause execution: the switching reviewer still loses 1.

Majority Can Be Unsafe Even With Aligned Reviewers

Recall: at a *Nash equilibrium*, no reviewer can improve its expected gain by changing its vote alone.

Take three reviewers with exactly the principal's objective,

$$u_1 = u_2 = u_3 = u_0,$$

and majority approval. Fix a harmful proposal with $u_0^\top d = -1$.

If all three approve, the proposal executes and everyone loses 1. If one reviewer switches to disapproval, the other two approvals still cause execution: the switching reviewer still loses 1.

All approving is therefore an unsafe Nash equilibrium. Disapproval remains weakly dominant for each reviewer, and all disapproving is a safe equilibrium.

Unanimity Prevents Harm at Every Equilibrium

Unlike majority voting, unanimity lets one reviewer's disapproval stop the proposal.

Unanimity Prevents Harm at Every Equilibrium

Unlike majority voting, unanimity lets one reviewer's disapproval stop the proposal.

Theorem (Safety at every Nash equilibrium)

Suppose the full panel covers the principal:

$$u_0 = \sum_{i=1}^n w_i u_i, \quad w_i \geq 0.$$

For every fixed proposal d , every Nash equilibrium of the unanimity voting game gives the principal nonnegative expected gain. In particular, if $u_0^\top d < 0$, the proposal is rejected with probability one.

Unanimity Prevents Harm at Every Equilibrium

Unlike majority voting, unanimity lets one reviewer's disapproval stop the proposal.

Theorem (Safety at every Nash equilibrium)

Suppose the full panel covers the principal:

$$u_0 = \sum_{i=1}^n w_i u_i, \quad w_i \geq 0.$$

For every fixed proposal d , every Nash equilibrium of the unanimity voting game gives the principal nonnegative expected gain. In particular, if $u_0^\top d < 0$, the proposal is rejected with probability one.

Votes may be randomized and need not be sincere.

Why Veto Power Prevents Harm

Fix a proposal d and a possibly mixed Nash equilibrium. Let $\rho \in [0, 1]$ be the probability of execution. Reviewer i 's expected gain is $\rho u_i^\top d$.

Why Veto Power Prevents Harm

Fix a proposal d and a possibly mixed Nash equilibrium. Let $\rho \in [0, 1]$ be the probability of execution. Reviewer i 's expected gain is $\rho u_i^\top d$.

Under unanimity, reviewer i can guarantee gain 0 by disapproving. At equilibrium, changing its vote cannot improve its gain. Its equilibrium gain must therefore be at least 0:

$$\rho u_i^\top d \geq 0 \quad \text{for every } i.$$

Why Veto Power Prevents Harm

Fix a proposal d and a possibly mixed Nash equilibrium. Let $\rho \in [0, 1]$ be the probability of execution. Reviewer i 's expected gain is $\rho u_i^\top d$.

Under unanimity, reviewer i can guarantee gain 0 by disapproving. At equilibrium, changing its vote cannot improve its gain. Its equilibrium gain must therefore be at least 0:

$$\rho u_i^\top d \geq 0 \quad \text{for every } i.$$

Using the coverage identity $u_0 = \sum_i w_i u_i$,

$$\rho u_0^\top d = \sum_i w_i (\rho u_i^\top d) \geq 0.$$

If $u_0^\top d < 0$, this forces $\rho = 0$: rejection is certain.



Why Veto Power Prevents Harm

Fix a proposal d and a possibly mixed Nash equilibrium. Let $\rho \in [0, 1]$ be the probability of execution. Reviewer i 's expected gain is $\rho u_i^\top d$.

Under unanimity, reviewer i can guarantee gain 0 by disapproving. At equilibrium, changing its vote cannot improve its gain. Its equilibrium gain must therefore be at least 0:

$$\rho u_i^\top d \geq 0 \quad \text{for every } i.$$

Using the coverage identity $u_0 = \sum_i w_i u_i$,

$$\rho u_0^\top d = \sum_i w_i (\rho u_i^\top d) \geq 0.$$

If $u_0^\top d < 0$, this forces $\rho = 0$: rejection is certain. □

As in Lecture 17, we add the individual equilibrium inequalities with nonnegative weights to obtain a guarantee for the human.

Can We Approve Every Proposal the Human Would?

Return to *sincere* votes. Soundness rules out harmful approvals. *Completeness* asks for the reverse: execute every proposal the principal weakly prefers to the baseline,

$$u_0^\top d \geq 0 \implies S(d) \in \mathcal{W}.$$

Can We Approve Every Proposal the Human Would?

Return to *sincere* votes. Soundness rules out harmful approvals. *Completeness* asks for the reverse: execute every proposal the principal weakly prefers to the baseline,

$$u_0^\top d \geq 0 \implies S(d) \in \mathcal{W}.$$

Soundness and completeness together would reproduce the principal's approve/reject decision from the binary votes alone.

Can We Approve Every Proposal the Human Would?

Return to *sincere* votes. Soundness rules out harmful approvals. *Completeness* asks for the reverse: execute every proposal the principal weakly prefers to the baseline,

$$u_0^\top d \geq 0 \implies S(d) \in \mathcal{W}.$$

Soundness and completeness together would reproduce the principal's approve/reject decision from the binary votes alone. Do the two specialists' votes contain enough information to make this decision?

The Same Votes Can Hide Opposite Principal Gains

Return to $u_0 = (1, 1)$, $u_1 = (1, 0)$, and $u_2 = (0, 1)$. Compare

$$d^+ = (2, -1), \quad d^- = (1, -2).$$

Both have a positive first coordinate and a negative second coordinate:

$$S(d^+) = S(d^-) = \{1\}.$$

The Same Votes Can Hide Opposite Principal Gains

Return to $u_0 = (1, 1)$, $u_1 = (1, 0)$, and $u_2 = (0, 1)$. Compare

$$d^+ = (2, -1), \quad d^- = (1, -2).$$

Both have a positive first coordinate and a negative second coordinate:

$$S(d^+) = S(d^-) = \{1\}.$$

But the principal's gains have opposite signs:

$$u_0^\top d^+ = 1, \quad u_0^\top d^- = -1.$$

The Same Votes Can Hide Opposite Principal Gains

Return to $u_0 = (1, 1)$, $u_1 = (1, 0)$, and $u_2 = (0, 1)$. Compare

$$d^+ = (2, -1), \quad d^- = (1, -2).$$

Both have a positive first coordinate and a negative second coordinate:

$$S(d^+) = S(d^-) = \{1\}.$$

But the principal's gains have opposite signs:

$$u_0^\top d^+ = 1, \quad u_0^\top d^- = -1.$$

No rule observing only the approval set can accept the first and reject the second. Unanimity rejects both to avoid approving the harmful one. The votes report signs, not whether one reviewer's gain outweighs the other's loss.

Exact Approval Requires an Aligned Reviewer

Assume $u_0 \neq 0$. Call reviewer i *individually aligned* if $u_i = cu_0$ for some $c > 0$. For every proposal, it gains, loses, or is indifferent exactly when the principal does.

Exact Approval Requires an Aligned Reviewer

Assume $u_0 \neq 0$. Call reviewer i *individually aligned* if $u_i = cu_0$ for some $c > 0$. For every proposal, it gains, loses, or is indifferent exactly when the principal does.

Theorem (Exact binary implementation)

For a finite panel, there exists a deterministic rule based only on sincere binary votes that satisfies

$$\text{execute } d \iff u_0^\top d \geq 0 \quad \text{for every } d \in \mathbb{R}^D$$

if and only if some reviewer is individually aligned.

Exact Approval Requires an Aligned Reviewer

Assume $u_0 \neq 0$. Call reviewer i *individually aligned* if $u_i = cu_0$ for some $c > 0$. For every proposal, it gains, loses, or is indifferent exactly when the principal does.

Theorem (Exact binary implementation)

For a finite panel, there exists a deterministic rule based only on sincere binary votes that satisfies

$$\text{execute } d \iff u_0^\top d \geq 0 \quad \text{for every } d \in \mathbb{R}^D$$

if and only if some reviewer is individually aligned.

If an aligned reviewer exists, simply follow that reviewer's vote. The converse proof is in the backup slides. It uses the requirement to accept *zero-gain* proposals as well as strictly beneficial ones.

Numerical Reports Can Recover the Lost Information

Change the reporting model: assume reviewer i truthfully reports its *numerical gain* $g_i(d) := u_i^\top d$, not just its sign. Suppose we know nonnegative weights satisfying $u_0 = \sum_i w_i u_i$.

Numerical Reports Can Recover the Lost Information

Change the reporting model: assume reviewer i truthfully reports its *numerical gain* $g_i(d) := u_i^\top d$, not just its sign. Suppose we know nonnegative weights satisfying $u_0 = \sum_i w_i u_i$. Then the weighted sum of reports is exactly the principal's gain:

$$\sum_i w_i g_i(d) = \left(\sum_i w_i u_i \right)^\top d = u_0^\top d.$$

Execute exactly when this sum is nonnegative: this recovers the human's decision, even without an aligned individual.

Numerical Reports Can Recover the Lost Information

Change the reporting model: assume reviewer i truthfully reports its *numerical gain* $g_i(d) := u_i^\top d$, not just its sign. Suppose we know nonnegative weights satisfying $u_0 = \sum_i w_i u_i$. Then the weighted sum of reports is exactly the principal's gain:

$$\sum_i w_i g_i(d) = \left(\sum_i w_i u_i \right)^\top d = u_0^\top d.$$

Execute exactly when this sum is nonnegative: this recovers the human's decision, even without an aligned individual.

This calculation needs the actual weights. It also assumes truthful numerical reports: unlike binary voting, we have not proved that reporting truthfully is optimal for the reviewers.

Summary

- ▶ A human can delegate approval without finding one perfectly aligned reviewer. Reviewers with complementary objectives can jointly prevent harmful actions.

Summary

- ▶ A human can delegate approval without finding one perfectly aligned reviewer. Reviewers with complementary objectives can jointly prevent harmful actions.
- ▶ The rule must require the right combination of reviewers, not just a large number. Giving each reviewer a veto also prevents harm at every equilibrium when the panel's objectives combine to represent the human's.

Summary

- ▶ A human can delegate approval without finding one perfectly aligned reviewer. Reviewers with complementary objectives can jointly prevent harmful actions.
- ▶ The rule must require the right combination of reviewers, not just a large number. Giving each reviewer a veto also prevents harm at every equilibrium when the panel's objectives combine to represent the human's.
- ▶ Safe oversight can be conservative. A rule based on binary votes may reject beneficial actions because the votes omit the sizes of gains and losses.

Thanks!

Questions?

References

- ▶ Natalie Collina, Surbhi Goel, Aaron Roth, and Sikata Sengupta. *Delegating Authorization to Misaligned Agents: Coalitional Alignment and Safe Control*. Unpublished manuscript, 2026.
- ▶ John C. Harsanyi. *Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility*. *Journal of Political Economy*, 63(4):309–321, 1955.
- ▶ Natalie Collina, Surbhi Goel, Aaron Roth, Emily Ryu, and Mirah Shi. *Emergent Alignment via Competition*. arXiv:2509.15090, version 2, February 2026.

Backup: Lottery Changes Give Linear Utility Gains

Let p^0 be a baseline lottery over outcomes $\{0, 1, \dots, D\}$, and let $v_j(\ell)$ be participant j 's utility at outcome ℓ . For a proposed lottery p , define

$$d_\ell = p_\ell - p_\ell^0 \quad (\ell = 1, \dots, D), \quad p_0 - p_0^0 = - \sum_{\ell=1}^D d_\ell.$$

Backup: Lottery Changes Give Linear Utility Gains

Let p^0 be a baseline lottery over outcomes $\{0, 1, \dots, D\}$, and let $v_j(\ell)$ be participant j 's utility at outcome ℓ . For a proposed lottery p , define

$$d_\ell = p_\ell - p_\ell^0 \quad (\ell = 1, \dots, D), \quad p_0 - p_0^0 = - \sum_{\ell=1}^D d_\ell.$$

The change in participant j 's expected utility is

$$\begin{aligned} \sum_{\ell=0}^D v_j(\ell)(p_\ell - p_\ell^0) &= \sum_{\ell=1}^D (v_j(\ell) - v_j(0)) d_\ell \\ &= u_j^\top d, \end{aligned}$$

where $(u_j)_\ell := v_j(\ell) - v_j(0)$.

Backup: Lottery Changes Give Linear Utility Gains

Let p^0 be a baseline lottery over outcomes $\{0, 1, \dots, D\}$, and let $v_j(\ell)$ be participant j 's utility at outcome ℓ . For a proposed lottery p , define

$$d_\ell = p_\ell - p_\ell^0 \quad (\ell = 1, \dots, D), \quad p_0 - p_0^0 = - \sum_{\ell=1}^D d_\ell.$$

The change in participant j 's expected utility is

$$\begin{aligned} \sum_{\ell=0}^D v_j(\ell)(p_\ell - p_\ell^0) &= \sum_{\ell=1}^D (v_j(\ell) - v_j(0)) d_\ell \\ &= u_j^\top d, \end{aligned}$$

where $(u_j)_\ell := v_j(\ell) - v_j(0)$.

Only the utility differences $v_j(\ell) - v_j(0)$ matter. Adding a constant to v_j changes none of these differences.

Backup: Small Changes Can Follow Any Direction

Suppose every baseline probability p_ℓ^0 is strictly positive. For *any* $d \in \mathbb{R}^D$, choose a sufficiently small $t > 0$:

$$p_\ell = p_\ell^0 + td_\ell \quad (\ell = 1, \dots, D), \quad p_0 = p_0^0 - t \sum_{\ell=1}^D d_\ell.$$

These probabilities are nonnegative and sum to one.

Backup: Small Changes Can Follow Any Direction

Suppose every baseline probability p_ℓ^0 is strictly positive. For *any* $d \in \mathbb{R}^D$, choose a sufficiently small $t > 0$:

$$p_\ell = p_\ell^0 + td_\ell \quad (\ell = 1, \dots, D), \quad p_0 = p_0^0 - t \sum_{\ell=1}^D d_\ell.$$

These probabilities are nonnegative and sum to one.

The gain is $tu_j^\top d$, with the same sign as $u_j^\top d$. Thus any d can be scaled into a feasible lottery change without changing any participant's preference for it over the baseline.

Backup: Small Changes Can Follow Any Direction

Suppose every baseline probability p_ℓ^0 is strictly positive. For *any* $d \in \mathbb{R}^D$, choose a sufficiently small $t > 0$:

$$p_\ell = p_\ell^0 + td_\ell \quad (\ell = 1, \dots, D), \quad p_0 = p_0^0 - t \sum_{\ell=1}^D d_\ell.$$

These probabilities are nonnegative and sum to one.

The gain is $tu_j^\top d$, with the same sign as $u_j^\top d$. Thus any d can be scaled into a feasible lottery change without changing any participant's preference for it over the baseline.

Recall the weighted-utility form from Lectures 3 and 17, with a constant a and nonnegative weights w_i :

$$v_0(\ell) = a + \sum_i w_i v_i(\ell) \quad \text{for every } \ell \quad \implies \quad u_0 = \sum_i w_i u_i.$$

Subtracting the equation at outcome 0 from the equation at outcome ℓ cancels a .

Backup: LP Duality Gives the Coverage Criterion

Let U_W have columns u_i for $i \in W$. To search for a harmful proposal that all members of W approve, consider

$$(P) \quad \max_{d \in \mathbb{R}^D} \quad -u_0^\top d \quad \text{subject to } U_W^\top d \geq 0,$$

$$(D) \quad \min_w \quad 0 \quad \text{subject to } U_W w = u_0, \quad w \geq 0.$$

The dual follows by writing $-U_W^\top d \leq 0$. Unlike the lottery differences in Lecture 3, d has no equality constraint, so the dual has no additional equality multiplier.

Backup: LP Duality Gives the Coverage Criterion

Let U_W have columns u_i for $i \in W$. To search for a harmful proposal that all members of W approve, consider

$$(P) \quad \max_{d \in \mathbb{R}^D} \quad -u_0^\top d \quad \text{subject to } U_W^\top d \geq 0,$$

$$(D) \quad \min_w \quad 0 \quad \text{subject to } U_W w = u_0, \quad w \geq 0.$$

The dual follows by writing $-U_W^\top d \leq 0$. Unlike the lottery differences in Lecture 3, d has no equality constraint, so the dual has no additional equality multiplier.

If every proposal approved by all of W gives the human nonnegative gain, (P) has optimal value zero, attained at $d = 0$. Strong duality gives feasible weights in (D), so W covers the principal.

Backup: LP Duality Gives the Coverage Criterion

Let U_W have columns u_i for $i \in W$. To search for a harmful proposal that all members of W approve, consider

$$(P) \quad \max_{d \in \mathbb{R}^D} \quad -u_0^\top d \quad \text{subject to } U_W^\top d \geq 0,$$

$$(D) \quad \min_w \quad 0 \quad \text{subject to } U_W w = u_0, \quad w \geq 0.$$

The dual follows by writing $-U_W^\top d \leq 0$. Unlike the lottery differences in Lecture 3, d has no equality constraint, so the dual has no additional equality multiplier.

If every proposal approved by all of W gives the human nonnegative gain, (P) has optimal value zero, attained at $d = 0$. Strong duality gives feasible weights in (D), so W covers the principal.

Taking the contrapositive: if W does not cover the principal, some d has $U_W^\top d \geq 0$ but $u_0^\top d < 0$.

Backup: Start Where the Human Is Indifferent

For the converse of exact binary implementation, assume $u_0 \neq 0$ and no reviewer has $u_i = cu_0$ with $c > 0$. We will find a zero-gain and a harmful proposal with identical votes.

Backup: Start Where the Human Is Indifferent

For the converse of exact binary implementation, assume $u_0 \neq 0$ and no reviewer has $u_i = cu_0$ with $c > 0$. We will find a zero-gain and a harmful proposal with identical votes.

The principal's indifference hyperplane is

$$H := \{d \in \mathbb{R}^D : u_0^\top d = 0\}.$$

For a reviewer whose utility vector is not proportional to u_0 , the set $\{x \in H : u_i^\top x = 0\}$ is a proper hyperplane in H . Finitely many such sets cannot fill H : each has zero volume within it.

Backup: Start Where the Human Is Indifferent

For the converse of exact binary implementation, assume $u_0 \neq 0$ and no reviewer has $u_i = cu_0$ with $c > 0$. We will find a zero-gain and a harmful proposal with identical votes.

The principal's indifference hyperplane is

$$H := \{d \in \mathbb{R}^D : u_0^\top d = 0\}.$$

For a reviewer whose utility vector is not proportional to u_0 , the set $\{x \in H : u_i^\top x = 0\}$ is a proper hyperplane in H . Finitely many such sets cannot fill H : each has zero volume within it.

Choose x outside their union. Then

$$u_0^\top x = 0, \quad u_i^\top x \neq 0 \quad \text{for every nonproportional } i.$$

If $D = 1$, all vectors are proportional and we choose $x = 0$.

Backup: A Harmful Move Leaves Every Vote Unchanged

Recall that $u_0^\top x = 0$. Set $d^- = x - \varepsilon u_0$ for sufficiently small $\varepsilon > 0$.

Backup: A Harmful Move Leaves Every Vote Unchanged

Recall that $u_0^\top x = 0$. Set $d^- = x - \varepsilon u_0$ for sufficiently small $\varepsilon > 0$.

- ▶ If u_i is not proportional to u_0 , then $u_i^\top x \neq 0$. Choose ε small enough to preserve the signs of these finitely many nonzero values.
- ▶ Every remaining reviewer has $u_i = c_i u_0$ with $c_i \leq 0$. It approves at both points, since

$$u_i^\top x = 0, \quad u_i^\top d^- = -\varepsilon c_i \|u_0\|_2^2 \geq 0.$$

Backup: A Harmful Move Leaves Every Vote Unchanged

Recall that $u_0^\top x = 0$. Set $d^- = x - \varepsilon u_0$ for sufficiently small $\varepsilon > 0$.

- ▶ If u_i is not proportional to u_0 , then $u_i^\top x \neq 0$. Choose ε small enough to preserve the signs of these finitely many nonzero values.
- ▶ Every remaining reviewer has $u_i = c_i u_0$ with $c_i \leq 0$. It approves at both points, since

$$u_i^\top x = 0, \quad u_i^\top d^- = -\varepsilon c_i \|u_0\|_2^2 \geq 0.$$

Thus $S(d^-) = S(x)$, but

$$u_0^\top x = 0, \quad u_0^\top d^- = -\varepsilon \|u_0\|_2^2 < 0.$$

Completeness requires accepting x ; soundness requires rejecting d^- . Identical votes make this impossible. $\square$