

Can Competition Produce Alignment?

Information and Competing Objectives

Aaron Roth

University of Pennsylvania

Fall 2026

Overview

- ▶ A user has her own goals. She talks to an LLM to learn information relevant to a decision, then chooses an action.

Overview

- ▶ A user has her own goals. She talks to an LLM to learn information relevant to a decision, then chooses an action.
- ▶ LLM providers also care about what she does—for example, what she buys or whom she votes for. They cannot act for her; they can only influence her through the conversation.

Overview

- ▶ A user has her own goals. She talks to an LLM to learn information relevant to a decision, then chooses an action.
- ▶ LLM providers also care about what she does—for example, what she buys or whom she votes for. They cannot act for her; they can only influence her through the conversation.
- ▶ Their objectives may differ from hers. But they compete for her attention: she chooses the provider whose conversation gives her the greatest expected value.

Overview

- ▶ A user has her own goals. She talks to an LLM to learn information relevant to a decision, then chooses an action.
- ▶ LLM providers also care about what she does—for example, what she buys or whom she votes for. They cannot act for her; they can only influence her through the conversation.
- ▶ Their objectives may differ from hers. But they compete for her attention: she chooses the provider whose conversation gives her the greatest expected value.
- ▶ Under a condition linking their competing objectives, this competition can serve her as well as an aligned provider would, even when no individual provider shares her objective.

Choosing the Best Offer Is Not Enough

AI providers choose how to advise a user. Their advice affects her decisions, but they may prefer different outcomes from the ones she wants.

Choosing the Best Offer Is Not Enough

AI providers choose how to advise a user. Their advice affects her decisions, but they may prefer different outcomes from the ones she wants.

Costly training can make model updates slower than users' adaptation and switching between deployed models. This difference in timescales motivates treating conversational policies as fixed while users choose whom to consult.

Choosing the Best Offer Is Not Enough

AI providers choose how to advise a user. Their advice affects her decisions, but they may prefer different outcomes from the ones she wants.

Costly training can make model updates slower than users' adaptation and switching between deployed models. This difference in timescales motivates treating conversational policies as fixed while users choose whom to consult.

We use a stylized economic model: providers commit to rules for communicating information, and selection ranks offers by the expected utility they give the user *with her actual behavior*. The selected provider follows its rule.

Choosing the Best Offer Is Not Enough

AI providers choose how to advise a user. Their advice affects her decisions, but they may prefer different outcomes from the ones she wants.

Costly training can make model updates slower than users' adaptation and switching between deployed models. This difference in timescales motivates treating conversational policies as fixed while users choose whom to consult.

We use a stylized economic model: providers commit to rules for communicating information, and selection ranks offers by the expected utility they give the user *with her actual behavior*. The selected provider follows its rule.

The user can select the best *offered* rule. Providers determine which rules are offered in the first place.

The question

When does competition make providers offer a rule that is best for the user among *all feasible rules*?

Two Providers Whose Objectives Add Up

Return to the deployment decision from Bayesian persuasion. The state θ is safe or harmful, with

$$\mathbb{P}(\text{safe}) = \frac{2}{5}, \quad \mathbb{P}(\text{harmful}) = \frac{3}{5}.$$

The user chooses deploy or hold. All three players receive 0 from holding. Deployment gives:

	Safe	Harmful
User: u_0	1	-4
Provider 1: u_1	1	1
Provider 2: u_2	0	-5

Two Providers Whose Objectives Add Up

Return to the deployment decision from Bayesian persuasion. The state θ is safe or harmful, with

$$\mathbb{P}(\text{safe}) = \frac{2}{5}, \quad \mathbb{P}(\text{harmful}) = \frac{3}{5}.$$

The user chooses deploy or hold. All three players receive 0 from holding. Deployment gives:

	Safe	Harmful
User: u_0	1	-4
Provider 1: u_1	1	1
Provider 2: u_2	0	-5

Provider 1 wants deployment regardless of safety. Provider 2 wants to avoid harmful deployment but is indifferent about safe deployment.

Two Providers Whose Objectives Add Up

Return to the deployment decision from Bayesian persuasion. The state θ is safe or harmful, with

$$\mathbb{P}(\text{safe}) = \frac{2}{5}, \quad \mathbb{P}(\text{harmful}) = \frac{3}{5}.$$

The user chooses deploy or hold. All three players receive 0 from holding. Deployment gives:

	Safe	Harmful
User: u_0	1	-4
Provider 1: u_1	1	1
Provider 2: u_2	0	-5

Provider 1 wants deployment regardless of safety. Provider 2 wants to avoid harmful deployment but is indifferent about safe deployment. Neither shares the user's objective, but at every action and state,

$$u_0 = u_1 + u_2.$$

Both Providers Can Offer the Same Three Rules

In this example, the user updates by Bayes' rule. At posterior safety probability x , deployment gives her $5x - 4$, while holding gives 0. She deploys when $x \geq 4/5$, including equality, regardless of the provider.

Both Providers Can Offer the Same Three Rules

In this example, the user updates by Bayes' rule. At posterior safety probability x , deployment gives her $5x - 4$, while holding gives 0. She deploys when $x \geq 4/5$, including equality, regardless of the provider.

Each provider can commit to any of these disclosure rules before the state is drawn:

- ▶ c_{none} : reveal no information. The user holds, since the prior safety probability is below $4/5$.

Both Providers Can Offer the Same Three Rules

In this example, the user updates by Bayes' rule. At posterior safety probability x , deployment gives her $5x - 4$, while holding gives 0. She deploys when $x \geq 4/5$, including equality, regardless of the provider.

Each provider can commit to any of these disclosure rules before the state is drawn:

- ▶ c_{none} : reveal no information. The user holds, since the prior safety probability is below $4/5$.
- ▶ c_{persuade} : recommend deployment in every safe state and with probability $1/6$ in a harmful state; otherwise recommend holding. A deployment recommendation gives posterior safety probability $(2/5)/(2/5 + 1/10) = 4/5$.

Both Providers Can Offer the Same Three Rules

In this example, the user updates by Bayes' rule. At posterior safety probability x , deployment gives her $5x - 4$, while holding gives 0. She deploys when $x \geq 4/5$, including equality, regardless of the provider.

Each provider can commit to any of these disclosure rules before the state is drawn:

- ▶ c_{none} : reveal no information. The user holds, since the prior safety probability is below $4/5$.
- ▶ c_{persuade} : recommend deployment in every safe state and with probability $1/6$ in a harmful state; otherwise recommend holding. A deployment recommendation gives posterior safety probability $(2/5)/(2/5 + 1/10) = 4/5$.
- ▶ c_{full} : reveal the state. The user deploys exactly when safe.

Both Providers Can Offer the Same Three Rules

In this example, the user updates by Bayes' rule. At posterior safety probability x , deployment gives her $5x - 4$, while holding gives 0. She deploys when $x \geq 4/5$, including equality, regardless of the provider.

Each provider can commit to any of these disclosure rules before the state is drawn:

- ▶ c_{none} : reveal no information. The user holds, since the prior safety probability is below $4/5$.
- ▶ c_{persuade} : recommend deployment in every safe state and with probability $1/6$ in a harmful state; otherwise recommend holding. A deployment recommendation gives posterior safety probability $(2/5)/(2/5 + 1/10) = 4/5$.
- ▶ c_{full} : reveal the state. The user deploys exactly when safe.

These three rules are the entire feasible set in this example.

A Good Rule Need Not Be Offered

Write $V_p(c)$ for player p 's expected utility under rule c , averaging over the state and the rule's randomization. Under c_{persuade} , safe and harmful deployment have probabilities $2/5$ and $(3/5)(1/6) = 1/10$, respectively. Thus the user's value is $2/5 - 4(1/10) = 0$.

Rule	User V_0	Provider 1 V_1	Provider 2 V_2
c_{none}	0	0	0
c_{persuade}	0	$1/2$	$-1/2$
c_{full}	$2/5$	$2/5$	0

A Good Rule Need Not Be Offered

Write $V_p(c)$ for player p 's expected utility under rule c , averaging over the state and the rule's randomization. Under c_{persuade} , safe and harmful deployment have probabilities $2/5$ and $(3/5)(1/6) = 1/10$, respectively. Thus the user's value is $2/5 - 4(1/10) = 0$.

Rule	User V_0	Provider 1 V_1	Provider 2 V_2
c_{none}	0	0	0
c_{persuade}	0	$1/2$	$-1/2$
c_{full}	$2/5$	$2/5$	0

Suppose provider 1 offers c_{persuade} and provider 2 offers c_{none} . The best offered value is 0, although full disclosure could give the user $2/5$.

A Good Rule Need Not Be Offered

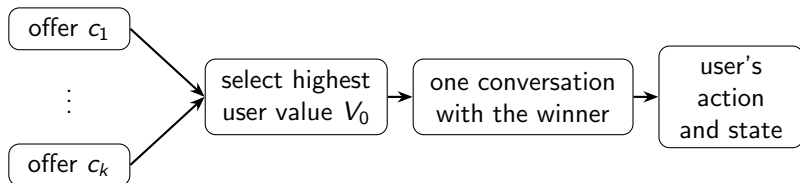
Write $V_p(c)$ for player p 's expected utility under rule c , averaging over the state and the rule's randomization. Under c_{persuade} , safe and harmful deployment have probabilities $2/5$ and $(3/5)(1/6) = 1/10$, respectively. Thus the user's value is $2/5 - 4(1/10) = 0$.

Rule	User V_0	Provider 1 V_1	Provider 2 V_2
c_{none}	0	0	0
c_{persuade}	0	$1/2$	$-1/2$
c_{full}	$2/5$	$2/5$	0

Suppose provider 1 offers c_{persuade} and provider 2 offers c_{none} . The best offered value is 0, although full disclosure could give the user $2/5$. Can either provider gain by changing its offer?

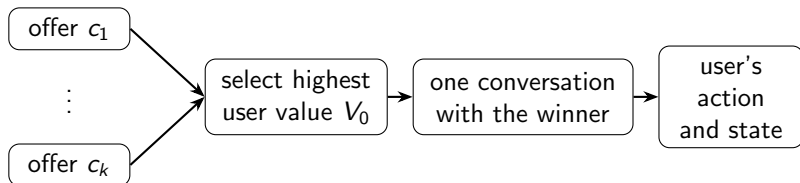
Providers Compete to Supply One Conversation

In the *Best-AI Selection Game*, each of k providers simultaneously commits to a rule c_i . Before the state is drawn, the user selects an offer with the highest expected user value, using a fixed tie-break.



Providers Compete to Supply One Conversation

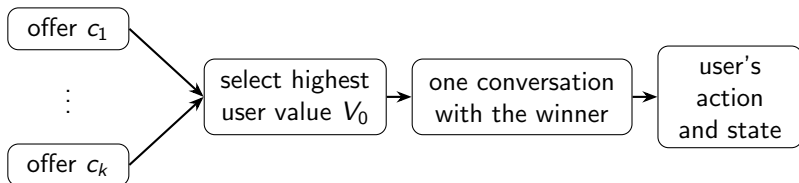
In the *Best-AI Selection Game*, each of k providers simultaneously commits to a rule c_i . Before the state is drawn, the user selects an offer with the highest expected user value, using a fixed tie-break.



Only the selected provider communicates; the user responds and chooses an action.

Providers Compete to Supply One Conversation

In the *Best-AI Selection Game*, each of k providers simultaneously commits to a rule c_i . Before the state is drawn, the user selects an offer with the highest expected user value, using a fixed tie-break.



Only the selected provider communicates; the user responds and chooses an action. Every provider gets utility from that final outcome, even if it was not selected. There is no separate reward for selection. For example, provider 2 dislikes harmful deployment even when provider 1 induces it.

A Bad Menu Creates a Profitable Deviation

Consider the offers $(c_{\text{persuade}}, c_{\text{none}})$. Both give the user 0, so selection depends on the provider tie-break.

Selected rule	V_0	V_1	V_2
c_{none}	0	0	0
c_{persuade}	0	$1/2$	$-1/2$
c_{full}	$2/5$	$2/5$	0

A Bad Menu Creates a Profitable Deviation

Consider the offers $(c_{\text{persuade}}, c_{\text{none}})$. Both give the user 0, so selection depends on the provider tie-break.

Selected rule	V_0	V_1	V_2
c_{none}	0	0	0
c_{persuade}	0	$1/2$	$-1/2$
c_{full}	$2/5$	$2/5$	0

- ▶ If c_{none} is selected, provider 1 can offer c_{full} and raise its payoff from 0 to $2/5$.
- ▶ If c_{persuade} is selected, provider 2 can offer c_{full} and raise its payoff from $-1/2$ to 0.

A Bad Menu Creates a Profitable Deviation

Consider the offers $(c_{\text{persuade}}, c_{\text{none}})$. Both give the user 0, so selection depends on the provider tie-break.

Selected rule	V_0	V_1	V_2
c_{none}	0	0	0
c_{persuade}	0	$1/2$	$-1/2$
c_{full}	$2/5$	$2/5$	0

- ▶ If c_{none} is selected, provider 1 can offer c_{full} and raise its payoff from 0 to $2/5$.
- ▶ If c_{persuade} is selected, provider 2 can offer c_{full} and raise its payoff from $-1/2$ to 0.

Either change wins selection strictly. The bad menu is not stable against unilateral changes. Can this reasoning be generalized?

The User's Behavior Need Not Be Bayesian

The example used Bayesian best responses. The general model does not: fix a user *behavioral policy* b , specifying how she sends messages and chooses a final action from the selected rule, her own information, and the conversation so far.

The User's Behavior Need Not Be Bayesian

The example used Bayesian best responses. The general model does not: fix a user *behavioral policy* b , specifying how she sends messages and chooses a final action from the selected rule, her own information, and the conversation so far.

The policy may randomize, use a heuristic, or ignore information. The same policy maps each possible interaction to user behavior; it need not maximize expected utility.

The User's Behavior Need Not Be Bayesian

The example used Bayesian best responses. The general model does not: fix a user *behavioral policy* b , specifying how she sends messages and chooses a final action from the selected rule, her own information, and the conversation so far.

The policy may randomize, use a heuristic, or ignore information. The same policy maps each possible interaction to user behavior; it need not maximize expected utility.

Providers know this policy when evaluating changes to their offers. We ask how well competition can serve *this user*, keeping her policy fixed.

The User's Behavior Need Not Be Bayesian

The example used Bayesian best responses. The general model does not: fix a user *behavioral policy* b , specifying how she sends messages and chooses a final action from the selected rule, her own information, and the conversation so far.

The policy may randomize, use a heuristic, or ignore information. The same policy maps each possible interaction to user behavior; it need not maximize expected utility.

Providers know this policy when evaluating changes to their offers. We ask how well competition can serve *this user*, keeping her policy fixed.

The user still selects the provider to use by choosing the one that maximizes her own utility.

A Provider Rule and User Policy Determine Outcomes

Now allow any common, nonempty class $\mathcal{C}$ of committed rules, including rules for entire conversations. Every provider can offer any $c \in \mathcal{C}$.

A Provider Rule and User Policy Determine Outcomes

Now allow any common, nonempty class $\mathcal{C}$ of committed rules, including rules for entire conversations. Every provider can offer any $c \in \mathcal{C}$.

With finite states Θ and user actions $\mathcal{A}$, an outcome is

$$z = (a, \theta) \in Z := \mathcal{A} \times \Theta.$$

The user has utility $u_0 : Z \rightarrow \mathbb{R}$; provider i has utility $u_i : Z \rightarrow \mathbb{R}$.

A Provider Rule and User Policy Determine Outcomes

Now allow any common, nonempty class $\mathcal{C}$ of committed rules, including rules for entire conversations. Every provider can offer any $c \in \mathcal{C}$.

With finite states Θ and user actions $\mathcal{A}$, an outcome is

$$z = (a, \theta) \in Z := \mathcal{A} \times \Theta.$$

The user has utility $u_0 : Z \rightarrow \mathbb{R}$; provider i has utility $u_i : Z \rightarrow \mathbb{R}$. Fix the distribution of the state and the user's initial information. Rule c and policy b induce an outcome distribution L_c^b . Assume it is the same whichever provider offers c , independent of unselected offers.

$$\boxed{V_p^b(c) := \mathbb{E}_{z \sim L_c^b}[u_p(z)]}, \quad p = 0, \dots, k.$$

The expectation includes state uncertainty and both parties' randomization. Since b stays fixed, write L_c and $V_p(c)$ from now on.

The Selected Rule Determines Everyone's Payoff

An offer *profile* is $\mathbf{c} = (c_1, \dots, c_k) \in \mathcal{C}^k$. The selection rule chooses the offer with highest actual expected user utility under the fixed policy b :

$$j(\mathbf{c}) \in \arg \max_{i \in \{1, \dots, k\}} V_0(c_i).$$

A fixed tie-break makes j single valued.

The Selected Rule Determines Everyone's Payoff

An offer *profile* is $\mathbf{c} = (c_1, \dots, c_k) \in \mathcal{C}^k$. The selection rule chooses the offer with highest actual expected user utility under the fixed policy b :

$$j(\mathbf{c}) \in \arg \max_{i \in \{1, \dots, k\}} V_0(c_i).$$

A fixed tie-break makes j single valued.

Only the selected rule is run. Every player's expected payoff is

$$\boxed{W_p(\mathbf{c}) := V_p(c_{j(\mathbf{c})})}, \quad p = 0, \dots, k.$$

The Selected Rule Determines Everyone's Payoff

An offer *profile* is $\mathbf{c} = (c_1, \dots, c_k) \in \mathcal{C}^k$. The selection rule chooses the offer with highest actual expected user utility under the fixed policy b :

$$j(\mathbf{c}) \in \arg \max_{i \in \{1, \dots, k\}} V_0(c_i).$$

A fixed tie-break makes j single valued.

Only the selected rule is run. Every player's expected payoff is

$$\boxed{W_p(\mathbf{c}) := V_p(c_{j(\mathbf{c})})}, \quad p = 0, \dots, k.$$

V_p evaluates one rule; W_p evaluates a competing profile.

Equilibrium Means No Profitable Unilateral Change

Fix the offers of every provider other than i , denoted $\mathbf{c}_{-i}$. Provider i may replace its rule by any $c'_i \in \mathcal{C}$.

Equilibrium Means No Profitable Unilateral Change

Fix the offers of every provider other than i , denoted $\mathbf{c}_{-i}$. Provider i may replace its rule by any $c'_i \in \mathcal{C}$.

Definition (Pure Nash equilibrium)

A profile $\mathbf{c}^*$ is a pure Nash equilibrium if, for every provider i and every alternative rule $c'_i \in \mathcal{C}$,

$$W_i(\mathbf{c}^*) \geq W_i(c'_i, \mathbf{c}_{-i}^*).$$

Equilibrium Means No Profitable Unilateral Change

Fix the offers of every provider other than i , denoted $\mathbf{c}_{-i}$. Provider i may replace its rule by any $c'_i \in \mathcal{C}$.

Definition (Pure Nash equilibrium)

A profile $\mathbf{c}^*$ is a pure Nash equilibrium if, for every provider i and every alternative rule $c'_i \in \mathcal{C}$,

$$W_i(\mathbf{c}^*) \geq W_i(c'_i, \mathbf{c}_{-i}^*).$$

“Pure” means that a provider chooses one committed rule; the rule itself may still randomize during the conversation.

The Benchmark Serves the Same User

A perfectly aligned provider would choose the best feasible rule for the fixed user policy b :

$$c^{\text{opt}} \in \arg \max_{c \in \mathcal{C}} V_0(c), \quad V_0^* := V_0(c^{\text{opt}}).$$

The Benchmark Serves the Same User

A perfectly aligned provider would choose the best feasible rule for the fixed user policy b :

$$c^{\text{opt}} \in \arg \max_{c \in \mathcal{C}} V_0(c), \quad V_0^* := V_0(c^{\text{opt}}).$$

We optimize the provider's rule, *not* the user's behavior. Every provider can offer c^{opt} and induce the same outcome distribution $L_{c^{\text{opt}}}$.

The Benchmark Serves the Same User

A perfectly aligned provider would choose the best feasible rule for the fixed user policy b :

$$c^{\text{opt}} \in \arg \max_{c \in \mathcal{C}} V_0(c), \quad V_0^* := V_0(c^{\text{opt}}).$$

We optimize the provider's rule, *not* the user's behavior. Every provider can offer c^{opt} and induce the same outcome distribution $L_{c^{\text{opt}}}$.

In the deployment example,

$$c^{\text{opt}} = c_{\text{full}}, \quad V_0^* = \frac{2}{5}.$$

In general this benchmark need not be full-information optimal: both provider capabilities and the user's behavior can limit its value.

Market Alignment Combines Providers' Objectives

In our example, $u_0 = u_1 + u_2$ at every outcome. Taking expectations therefore gives $V_0(c) = V_1(c) + V_2(c)$ for every rule. We generalize this relationship:

Definition (Exact market alignment)

There are a constant $c_0 \in \mathbb{R}$ and weights $w_i \geq 0$ such that

$$V_0(c) = c_0 + \sum_{i=1}^k w_i V_i(c) \quad \text{for every } c \in \mathcal{C}.$$

Market Alignment Combines Providers' Objectives

In our example, $u_0 = u_1 + u_2$ at every outcome. Taking expectations therefore gives $V_0(c) = V_1(c) + V_2(c)$ for every rule. We generalize this relationship:

Definition (Exact market alignment)

There are a constant $c_0 \in \mathbb{R}$ and weights $w_i \geq 0$ such that

$$V_0(c) = c_0 + \sum_{i=1}^k w_i V_i(c) \quad \text{for every } c \in \mathcal{C}.$$

One set of weights must work for *all* feasible rules, not be chosen separately for each comparison.

Market Alignment Combines Providers' Objectives

In our example, $u_0 = u_1 + u_2$ at every outcome. Taking expectations therefore gives $V_0(c) = V_1(c) + V_2(c)$ for every rule. We generalize this relationship:

Definition (Exact market alignment)

There are a constant $c_0 \in \mathbb{R}$ and weights $w_i \geq 0$ such that

$$V_0(c) = c_0 + \sum_{i=1}^k w_i V_i(c) \quad \text{for every } c \in \mathcal{C}.$$

One set of weights must work for *all* feasible rules, not be chosen separately for each comparison.

The weights describe a relationship among objectives. Providers need not use them when choosing their offers, and they need not sum to one.

Competition Delivers the Best Feasible User Value

Theorem (Best-AI selection under exact market alignment)

Fix any user policy b . If, for every feasible rule c ,

$$V_0(c) = c_0 + \sum_{i=1}^k w_i V_i(c), \quad w_i \geq 0,$$

then every pure Nash equilibrium $\mathbf{c}^*$ satisfies

$$W_0(\mathbf{c}^*) = V_0^*.$$

Competition Delivers the Best Feasible User Value

Theorem (Best-AI selection under exact market alignment)

Fix any user policy b . If, for every feasible rule c ,

$$V_0(c) = c_0 + \sum_{i=1}^k w_i V_i(c), \quad w_i \geq 0,$$

then every pure Nash equilibrium $\mathbf{c}^*$ satisfies

$$\boxed{W_0(\mathbf{c}^*) = V_0^*}.$$

We will show that every suboptimal profile admits a profitable deviation.

A Better Rule Benefits Some Provider

Consider a profile $\mathbf{c}$ whose selected rule $c := c_{j(\mathbf{c})}$ gives the user less than the benchmark:

$$V_0(c^{\text{opt}}) - V_0(c) > 0.$$

A Better Rule Benefits Some Provider

Consider a profile $\mathbf{c}$ whose selected rule $c := c_j(\mathbf{c})$ gives the user less than the benchmark:

$$V_0(\mathbf{c}^{\text{opt}}) - V_0(c) > 0.$$

Subtract the two market-alignment identities. The constant cancels:

$$V_0(\mathbf{c}^{\text{opt}}) - V_0(c) = \sum_i w_i [V_i(\mathbf{c}^{\text{opt}}) - V_i(c)].$$

A Better Rule Benefits Some Provider

Consider a profile $\mathbf{c}$ whose selected rule $c := c_{j(\mathbf{c})}$ gives the user less than the benchmark:

$$V_0(c^{\text{opt}}) - V_0(c) > 0.$$

Subtract the two market-alignment identities. The constant cancels:

$$V_0(c^{\text{opt}}) - V_0(c) = \sum_i w_i [V_i(c^{\text{opt}}) - V_i(c)].$$

The left side is positive and every $w_i \geq 0$. At least one provider i must therefore satisfy

$$V_i(c^{\text{opt}}) > V_i(c).$$

That provider would benefit if the user received the better rule.

Offering the Better Rule Wins Selection

Let the provider i identified on the previous slide offer c^{opt} , keeping everyone else's offer fixed. Because c was selected from the original offers,

$$V_0(c_\ell) \leq V_0(c) < V_0(c^{\text{opt}}) \quad \text{for every original offer } c_\ell.$$

Offering the Better Rule Wins Selection

Let the provider i identified on the previous slide offer c^{opt} , keeping everyone else's offer fixed. Because c was selected from the original offers,

$$V_0(c_\ell) \leq V_0(c) < V_0(c^{\text{opt}}) \quad \text{for every original offer } c_\ell.$$

The new offer strictly wins selection. It induces the same outcome distribution as the benchmark, so provider i 's new payoff is

$$V_i(c^{\text{opt}}) > V_i(c).$$

This is a profitable unilateral deviation.

Offering the Better Rule Wins Selection

Let the provider i identified on the previous slide offer c^{opt} , keeping everyone else's offer fixed. Because c was selected from the original offers,

$$V_0(c_\ell) \leq V_0(c) < V_0(c^{\text{opt}}) \quad \text{for every original offer } c_\ell.$$

The new offer strictly wins selection. It induces the same outcome distribution as the benchmark, so provider i 's new payoff is

$$V_i(c^{\text{opt}}) > V_i(c).$$

This is a profitable unilateral deviation.

Thus no suboptimal profile can be a pure Nash equilibrium. Every pure equilibrium's selected rule must attain the best feasible user value V_0^* . □

Full Disclosure Is an Equilibrium Outcome

Return to the Bayesian deployment example and the offers $(c_{\text{persuade}}, c_{\text{full}})$. Full disclosure is selected strictly and gives payoffs $(2/5, 2/5, 0)$.

Rule	V_0	V_1	V_2
c_{none}	0	0	0
c_{persuade}	0	$1/2$	$-1/2$
c_{full}	$2/5$	$2/5$	0

Full Disclosure Is an Equilibrium Outcome

Return to the Bayesian deployment example and the offers $(c_{\text{persuade}}, c_{\text{full}})$. Full disclosure is selected strictly and gives payoffs $(2/5, 2/5, 0)$.

Rule	V_0	V_1	V_2
c_{none}	0	0	0
c_{persuade}	0	$1/2$	$-1/2$
c_{full}	$2/5$	$2/5$	0

- ▶ Provider 1 cannot displace full disclosure with either lower-value rule. Offering full disclosure changes no outcome.
- ▶ Provider 2 already receives 0, its largest feasible payoff; changing its offer cannot give a strict improvement.

Full Disclosure Is an Equilibrium Outcome

Return to the Bayesian deployment example and the offers $(c_{\text{persuade}}, c_{\text{full}})$. Full disclosure is selected strictly and gives payoffs $(2/5, 2/5, 0)$.

Rule	V_0	V_1	V_2
c_{none}	0	0	0
c_{persuade}	0	$1/2$	$-1/2$
c_{full}	$2/5$	$2/5$	0

- ▶ Provider 1 cannot displace full disclosure with either lower-value rule. Offering full disclosure changes no outcome.
- ▶ Provider 2 already receives 0, its largest feasible payoff; changing its offer cannot give a strict improvement.

Thus this is a pure equilibrium for either selection tie-break. The theorem says every pure equilibrium gives the user $2/5$.

Three Assumptions Make the Deviation Work

The proof identifies which features of the selection game matter:

Three Assumptions Make the Deviation Work

The proof identifies which features of the selection game matter:

- ▶ *Common capability*: every provider can implement the best rule for this user and induce the same outcome distribution.

Three Assumptions Make the Deviation Work

The proof identifies which features of the selection game matter:

- ▶ *Common capability*: every provider can implement the best rule for this user and induce the same outcome distribution.
- ▶ *Outcome incentives*: unselected providers still care about the final decision. Improving that decision can benefit a provider even when it was not previously selected.

Three Assumptions Make the Deviation Work

The proof identifies which features of the selection game matter:

- ▶ *Common capability*: every provider can implement the best rule for this user and induce the same outcome distribution.
- ▶ *Outcome incentives*: unselected providers still care about the final decision. Improving that decision can benefit a provider even when it was not previously selected.
- ▶ *Correct selection*: offers are ranked by actual user utility under the fixed behavioral policy. Arbitrary interaction is allowed.

Shared Conflicts Can Defeat Competition

The positive proof needs a provider who benefits from helping the user. We now ask whether this can be guaranteed *across decision problems*.

Shared Conflicts Can Defeat Competition

The positive proof needs a provider who benefits from helping the user. We now ask whether this can be guaranteed *across decision problems*.

Fix user utility v_0 and provider utilities $v_1, \dots, v_k$ over a finite set of terminal consequences Ω . We may construct a new problem in which the action and state determine a lottery over Ω .

Shared Conflicts Can Defeat Competition

The positive proof needs a provider who benefits from helping the user. We now ask whether this can be guaranteed *across decision problems*.

Fix user utility v_0 and provider utilities $v_1, \dots, v_k$ over a finite set of terminal consequences Ω . We may construct a new problem in which the action and state determine a lottery over Ω .

Theorem (A worst-case converse)

If there are no $c_0 \in \mathbb{R}$ and $w_i \geq 0$ with

$$v_0(\omega) = c_0 + \sum_i w_i v_i(\omega) \quad \text{for every } \omega \in \Omega,$$

there is a binary decision problem with a Bayesian best-responding user and fully informed providers allowed any randomized conversation rule, yet all providers withholding information is a suboptimal pure Nash equilibrium. Revelation achieves the user's full-information optimum.

Harsanyi Finds a Conflict over Two Lotteries

Recall Harsanyi's theorem (Lecture 3). Suppose that, for every pair of lotteries, whenever all providers regard the first as at least as good as the second, the user does too. Then

$$v_0 = c_0 + \sum_{i=1}^k w_i v_i, \quad w_i \geq 0.$$

Harsanyi Finds a Conflict over Two Lotteries

Recall Harsanyi's theorem (Lecture 3). Suppose that, for every pair of lotteries, whenever all providers regard the first as at least as good as the second, the user does too. Then

$$v_0 = c_0 + \sum_{i=1}^k w_i v_i, \quad w_i \geq 0.$$

We use the *contrapositive*: if no such representation exists, there are lotteries $\mu^G, \mu^B \in \Delta(\Omega)$ such that the user strictly prefers G , but every provider regards B as at least as good as G .

Harsanyi Finds a Conflict over Two Lotteries

Recall Harsanyi's theorem (Lecture 3). Suppose that, for every pair of lotteries, whenever all providers regard the first as at least as good as the second, the user does too. Then

$$v_0 = c_0 + \sum_{i=1}^k w_i v_i, \quad w_i \geq 0.$$

We use the *contrapositive*: if no such representation exists, there are lotteries $\mu^G, \mu^B \in \Delta(\Omega)$ such that the user strictly prefers G , but every provider regards B as at least as good as G . Write player p 's expected utilities under these lotteries as

$$g_p := \mathbb{E}_{\omega \sim \mu^G}[v_p(\omega)], \quad b_p := \mathbb{E}_{\omega \sim \mu^B}[v_p(\omega)].$$

Thus

$$g_0 > b_0, \quad g_i \leq b_i \quad \text{for every provider } i.$$

We will make useful information move the outcome toward G .

Guessing a Hidden Bit Determines the Outcome

Providers commit and the user selects one rule. Then draw a fair bit $s \in \{0, 1\}$. Every provider observes s ; the user has no information about it.

Guessing a Hidden Bit Determines the Outcome

Providers commit and the user selects one rule. Then draw a fair bit $s \in \{0, 1\}$. Every provider observes s ; the user has no information about it.

For this construction, fix the user's policy to Bayesian best response: after the conversation she chooses a bit $a \in \{0, 1\}$.

Draw the terminal consequence from:

	$s = 0$	$s = 1$
$a = 0$	μ^G	μ^B
$a = 1$	μ^B	μ^G

Thus guessing correctly gives G ; guessing incorrectly gives B .

Guessing a Hidden Bit Determines the Outcome

Providers commit and the user selects one rule. Then draw a fair bit $s \in \{0, 1\}$. Every provider observes s ; the user has no information about it.

For this construction, fix the user's policy to Bayesian best response: after the conversation she chooses a bit $a \in \{0, 1\}$.

Draw the terminal consequence from:

	$s = 0$	$s = 1$
$a = 0$	μ^G	μ^B
$a = 1$	μ^B	μ^G

Thus guessing correctly gives G ; guessing incorrectly gives B .

Providers may use any randomized, interactive conversation rule.

Revealing s lets the user always choose correctly.

The User Can Always Ignore the Conversation

Let $q := \mathbb{P}(a = s)$ be the user's probability of guessing correctly.
Her expected utility is

$$qg_0 + (1 - q)b_0 = b_0 + (g_0 - b_0)q.$$

Because $g_0 > b_0$, she wants to maximize q .

The User Can Always Ignore the Conversation

Let $q := \mathbb{P}(a = s)$ be the user's probability of guessing correctly. Her expected utility is

$$qg_0 + (1 - q)b_0 = b_0 + (g_0 - b_0)q.$$

Because $g_0 > b_0$, she wants to maximize q .

If every provider sends only constant messages, the user learns nothing about the fair bit, so $q = 1/2$.

The User Can Always Ignore the Conversation

Let $q := \mathbb{P}(a = s)$ be the user's probability of guessing correctly. Her expected utility is

$$qg_0 + (1 - q)b_0 = b_0 + (g_0 - b_0)q.$$

Because $g_0 > b_0$, she wants to maximize q .

If every provider sends only constant messages, the user learns nothing about the fair bit, so $q = 1/2$.

After *any* provider changes its rule, the user can still ignore all messages and guess a fixed bit. For this Bayesian user, best response gives

$$q \geq \frac{1}{2}.$$

This holds whether she selects the new rule or an unchanged one.

No Provider Gains from Informing the User

Recall: $q = \mathbb{P}(a = s)$, and provider i has $g_i \leq b_i$. Its expected utility from the final consequence is

$$qg_i + (1 - q)b_i = b_i + (g_i - b_i)q.$$

This is nonincreasing in the user's success probability q .

No Provider Gains from Informing the User

Recall: $q = \mathbb{P}(a = s)$, and provider i has $g_i \leq b_i$. Its expected utility from the final consequence is

$$qg_i + (1 - q)b_i = b_i + (g_i - b_i)q.$$

This is nonincreasing in the user's success probability q .

Starting from all providers' constant-message rules, any unilateral deviation leaves $q \geq 1/2$. Hence

$$\underbrace{b_i + (g_i - b_i)q}_{\text{payoff after deviation}} \leq \underbrace{b_i + \frac{g_i - b_i}{2}}_{\text{payoff before deviation}}.$$

No Provider Gains from Informing the User

Recall: $q = \mathbb{P}(a = s)$, and provider i has $g_i \leq b_i$. Its expected utility from the final consequence is

$$qg_i + (1 - q)b_i = b_i + (g_i - b_i)q.$$

This is nonincreasing in the user's success probability q . Starting from all providers' constant-message rules, any unilateral deviation leaves $q \geq 1/2$. Hence

$$\underbrace{b_i + (g_i - b_i)q}_{\text{payoff after deviation}} \leq \underbrace{b_i + \frac{g_i - b_i}{2}}_{\text{payoff before deviation}}.$$

No provider has a profitable deviation, even to a complicated interactive rule. The all-uninformative profile is a pure Nash equilibrium.

Revelation Is Feasible, but Silence Is an Equilibrium

At the equilibrium, the user guesses correctly with probability $1/2$. A provider who reveals s would let her guess correctly with probability 1, the largest possible value. Therefore

$$W_0^{\text{NE}} = \frac{g_0 + b_0}{2}, \quad V_0^* = g_0,$$

$$V_0^* - W_0^{\text{NE}} = \frac{g_0 - b_0}{2} > 0. \quad \square$$

Revelation Is Feasible, but Silence Is an Equilibrium

At the equilibrium, the user guesses correctly with probability $1/2$. A provider who reveals s would let her guess correctly with probability 1, the largest possible value. Therefore

$$W_0^{\text{NE}} = \frac{g_0 + b_0}{2}, \quad V_0^* = g_0,$$

$$V_0^* - W_0^{\text{NE}} = \frac{g_0 - b_0}{2} > 0. \quad \square$$

Every provider weakly prefers the user to remain uninformed, even though revealing the state is feasible .

Summary

- ▶ The user has her own preferences and takes the final action.
An LLM supplies information that can change her decision.

Summary

- ▶ The user has her own preferences and takes the final action. An LLM supplies information that can change her decision.
- ▶ Providers have their own preferences over her actions. They cannot choose for her: they must influence her through conversation, and compete for her attention.

Summary

- ▶ The user has her own preferences and takes the final action. An LLM supplies information that can change her decision.
- ▶ Providers have their own preferences over her actions. They cannot choose for her: they must influence her through conversation, and compete for her attention.
- ▶ When the user's objective is a nonnegative combination of provider objectives, any shortfall from the aligned-provider benchmark gives some provider an incentive to offer a better conversation. Selection by user value makes that offer win.

Summary

- ▶ The user has her own preferences and takes the final action. An LLM supplies information that can change her decision.
- ▶ Providers have their own preferences over her actions. They cannot choose for her: they must influence her through conversation, and compete for her attention.
- ▶ When the user's objective is a nonnegative combination of provider objectives, any shortfall from the aligned-provider benchmark gives some provider an incentive to offer a better conversation. Selection by user value makes that offer win.
- ▶ Under our assumptions, every pure equilibrium then serves the actual user as well as an aligned provider could. *Competition can produce aligned outcomes from individually misaligned providers.*

Thanks!

See you next class!

References

- ▶ Natalie Collina, Surbhi Goel, Aaron Roth, Emily Ryu, and Mirah Shi. *Emergent Alignment via Competition*. arXiv:2509.15090
- ▶ Emir Kamenica and Matthew Gentzkow. *Bayesian Persuasion*. American Economic Review, 101(6):2590–2615, 2011.
- ▶ John C. Harsanyi. *Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility*. Journal of Political Economy, 63(4):309–321, 1955.