Targeted Data Acquisition for Evolving Negotiation Agents

Minae Kwon, Siddharth Karamcheti,
Mariano-Florentino Cuéllar, Dorsa Sadigh

Negotiation is a bargaining process
by which a joint decision is made by two parties

Negotiation is a bargaining process by which a joint decision is made by two parties

Lawyers in court

Negotiation is a bargaining process by which a joint decision is made by two parties

Lawyers in court

Employee negotiating salary

Negotiation is a bargaining process by which a joint decision is made by two parties

Lawyers in court

Employee negotiating salary

2021 UN climate change conference


Nước Tương
Soya sauce

Desiderata

Desiderata

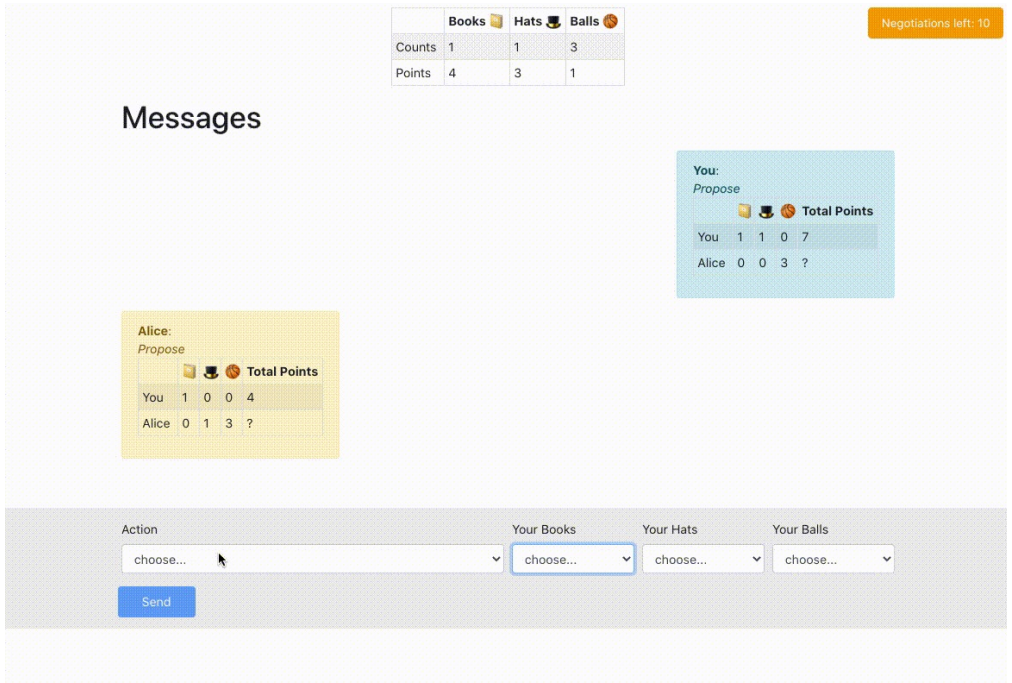
(1) Agents that maximize their self-interest

Desiderata

(1) Agents that maximize their self-interest

(2) Agents that can compromise (find Pareto-optimal solutions)

Supervised Learning (SL)

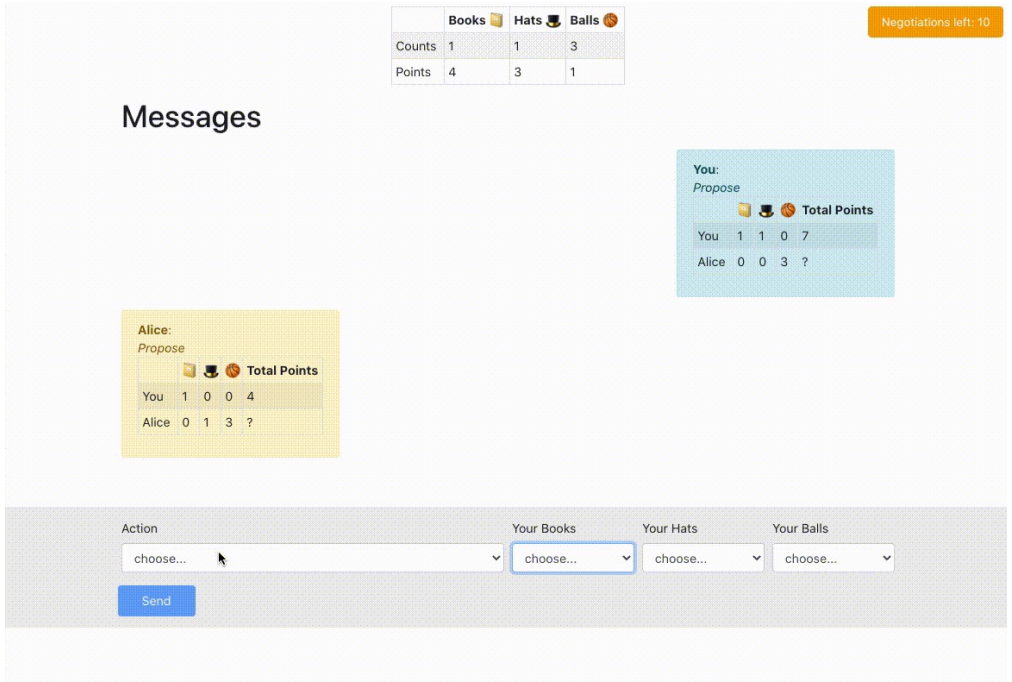


$$L(\theta) = -\sum_{x,c}\sum_{t} \log p_\theta(x_t | x_{0:t-1}, c \quad)$$

$$-\alpha \sum_{x,c}\sum_{j} \log p_\theta(o_j | x_{0:t-1}, c \quad)$$

Lewis et al., Deal or No Deal? End-to-End Learning of Negotiation Dialogues, 2017

Supervised Learning (SL)



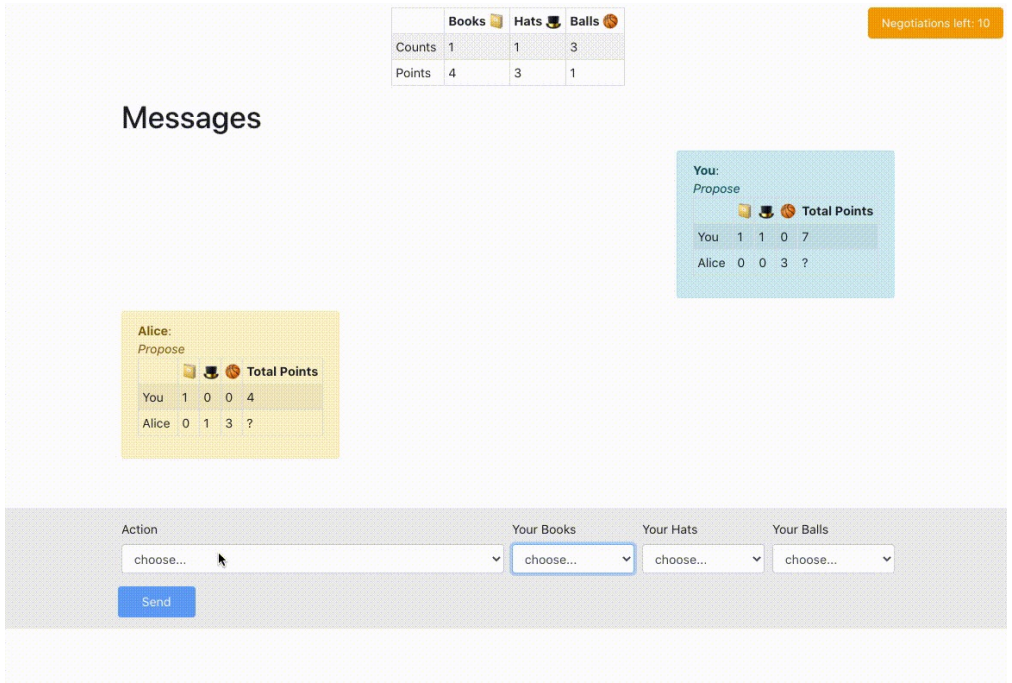
$$L(\theta) = -\sum_{x,c}\sum_{t} \log p_\theta(x_t | x_{0:t-1}, c)$$

utterances *context*

$$-\alpha \sum_{x,c}\sum_{j} \log p_\theta(o_j | x_{0:t-1}, c)$$

Lewis et al., Deal or No Deal? End-to-End Learning of Negotiation Dialogues, 2017

Supervised Learning (SL)



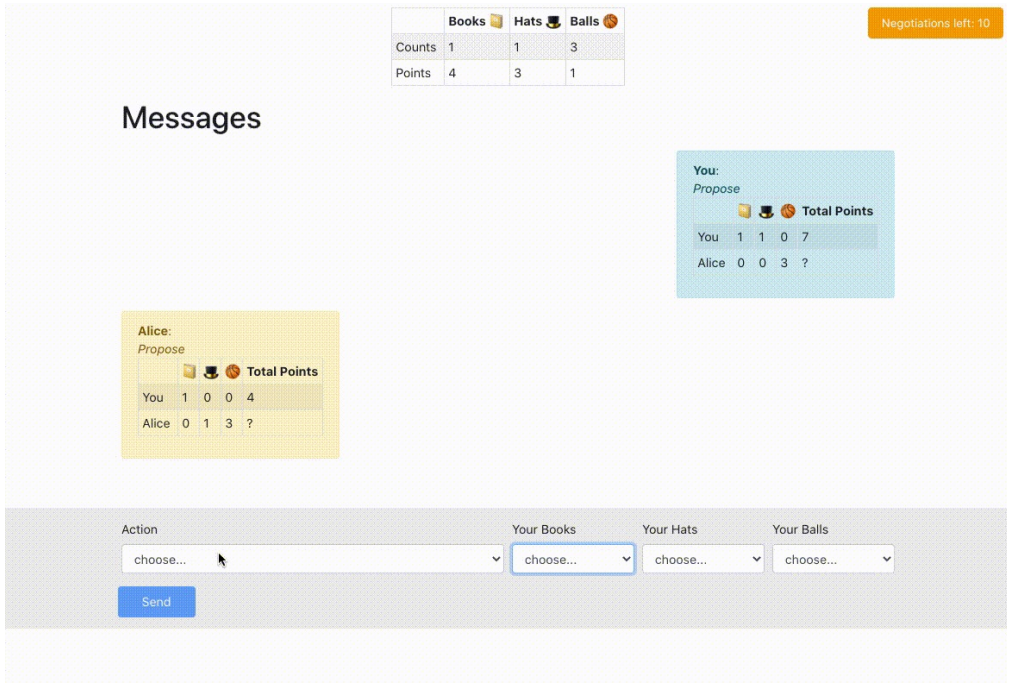
$$L(\theta) = -\underbrace{\sum_{x,c}\sum_{t} \log p_\theta(x_t | x_{0:t-1}, c)}_{\textit{utterance prediction loss}}$$

utterances — x_t — *context* — c

$$-\alpha \sum_{x,c}\sum_{j} \log p_\theta(o_j | x_{0:t-1}, c)$$

Lewis et al., Deal or No Deal? End-to-End Learning of Negotiation Dialogues, 2017

Supervised Learning (SL)



$$L(\theta) = -\sum_{x,c}\sum_{t} \log p_\theta(x_t | x_{0:t-1}, c)$$

utterances *context*

utterance prediction loss

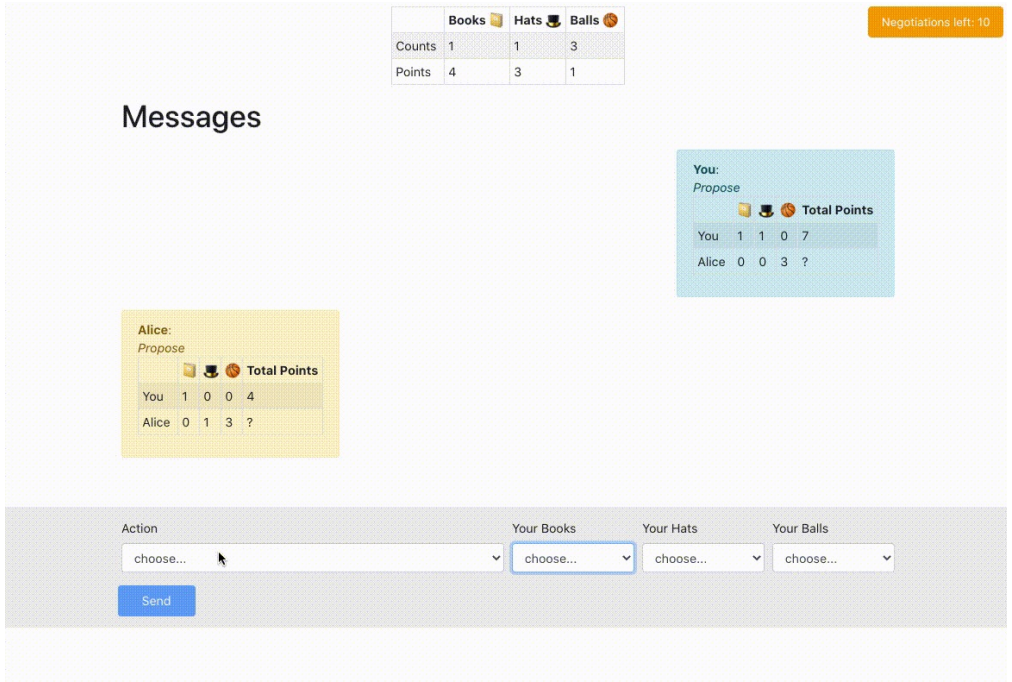
$$-\alpha \sum_{x,c}\sum_{j} \log p_\theta(o_j | x_{0:t-1}, c)$$

jth item

final split prediction loss

Lewis et al., Deal or No Deal? End-to-End Learning of Negotiation Dialogues, 2017

Supervised Learning (SL)



$$L(\theta) = -\sum_{x,c}\sum_{t} \log p_\theta(x_t | x_{0:t-1}, c)$$

utterances *context*

utterance prediction loss

$$-\alpha \sum_{x,c}\sum_{j} \log p_\theta(o_j | x_{0:t-1}, c)$$

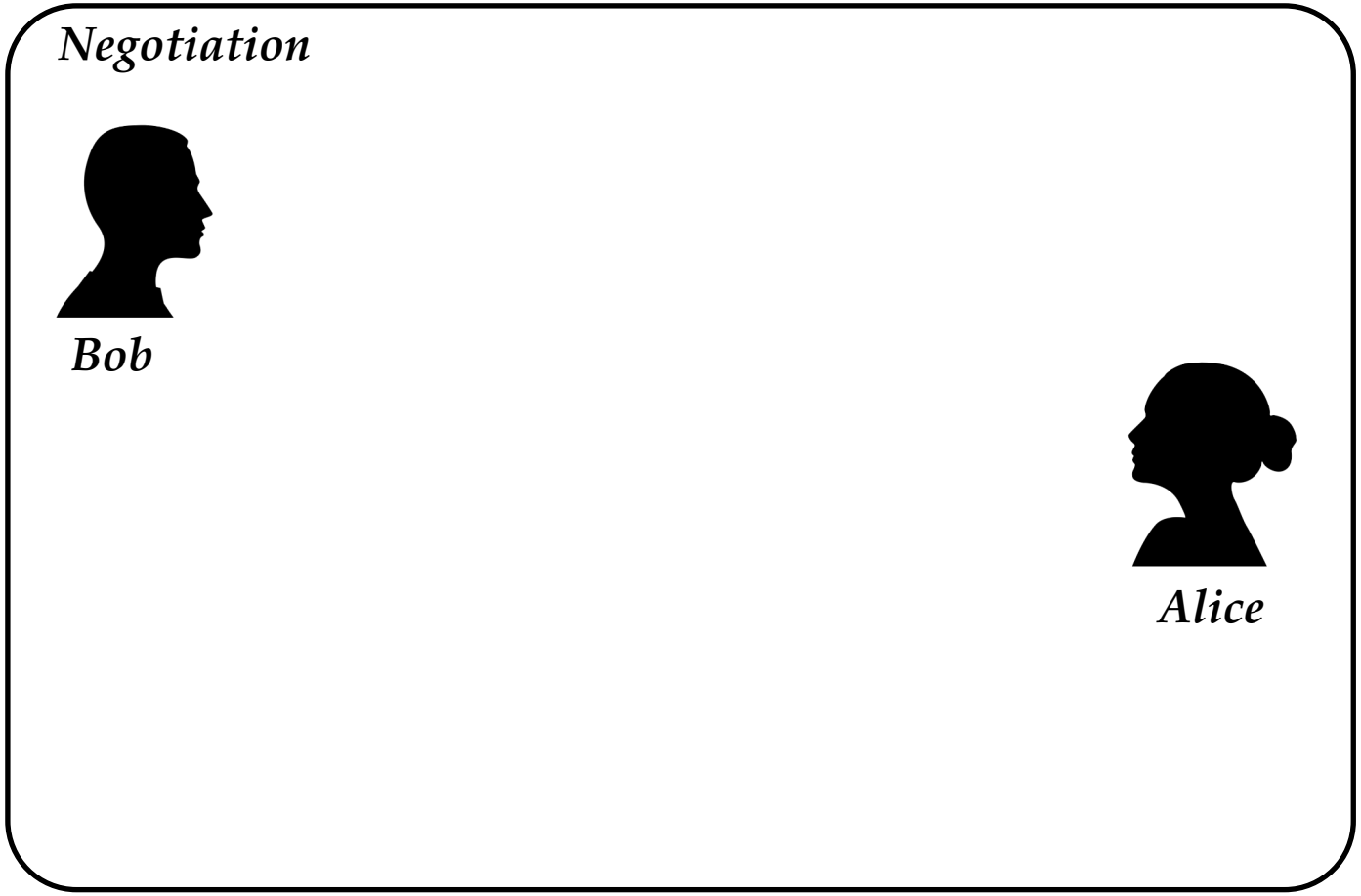
jth item

final split prediction loss

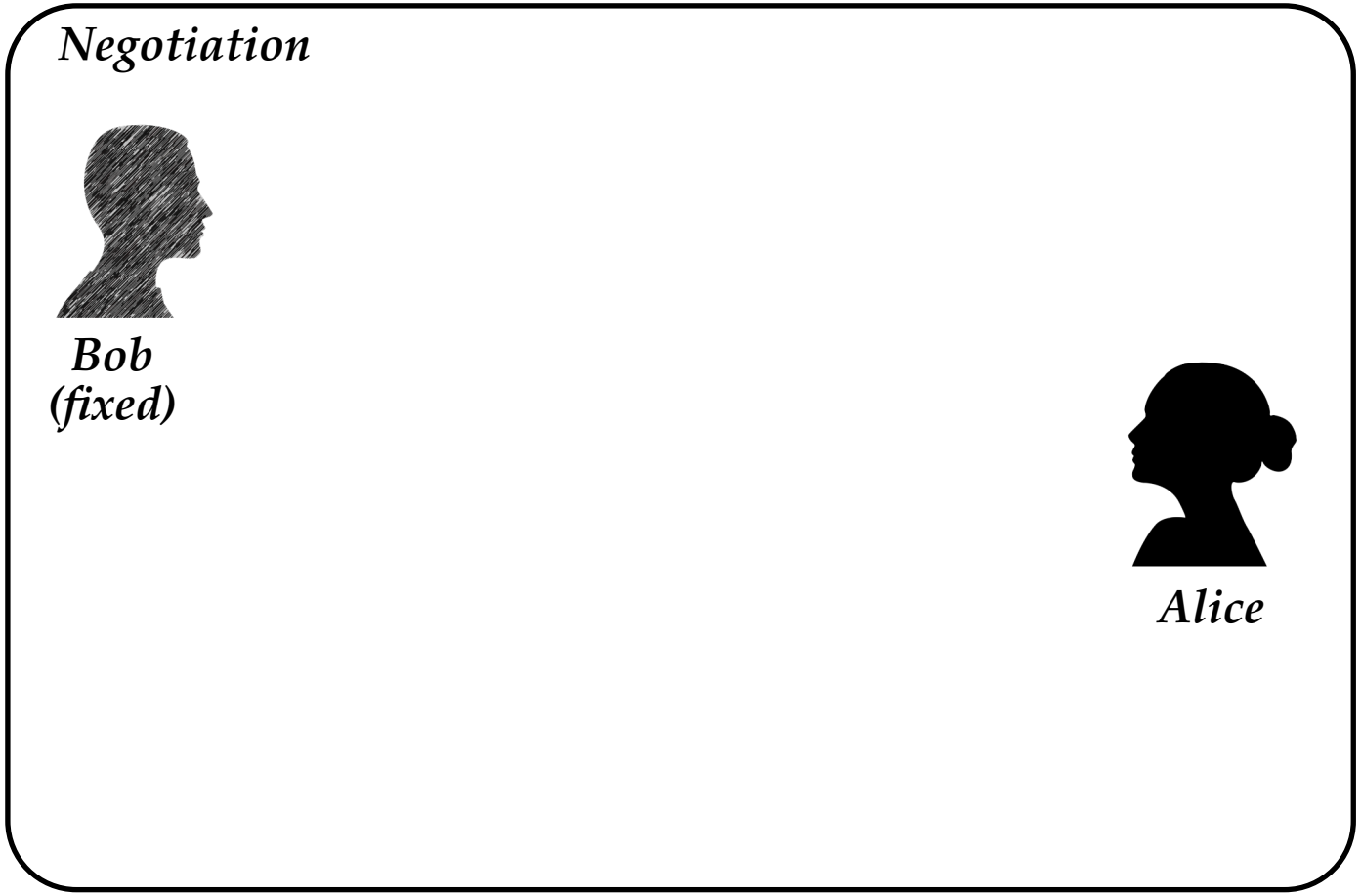
Relationship to dataset: bias inherited from dataset

Lewis et al., Deal or No Deal? End-to-End Learning of Negotiation Dialogues, 2017

Reinforcement Learning (RL)



Reinforcement Learning (RL)



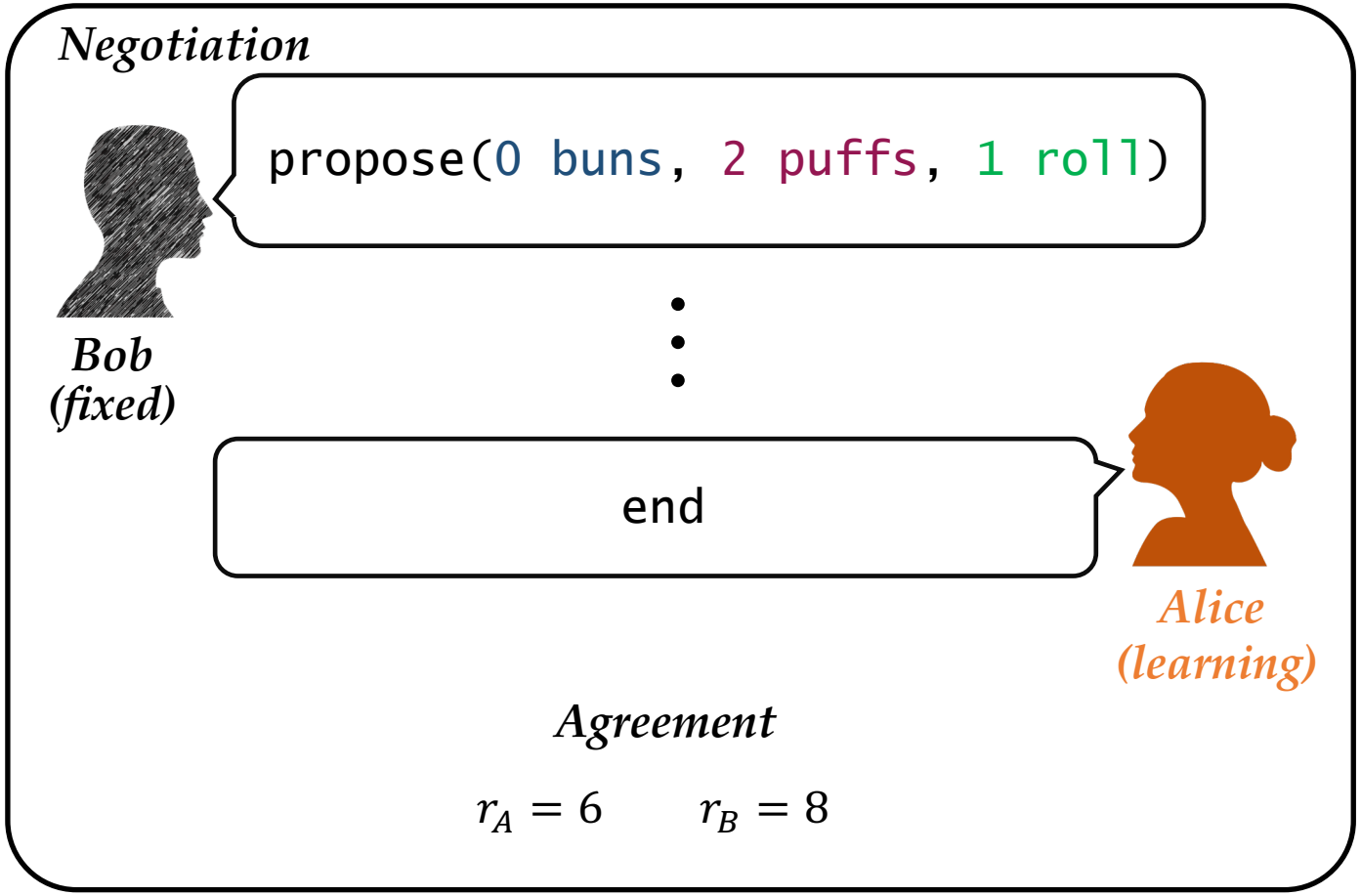
Reinforcement Learning (RL)

Negotiation

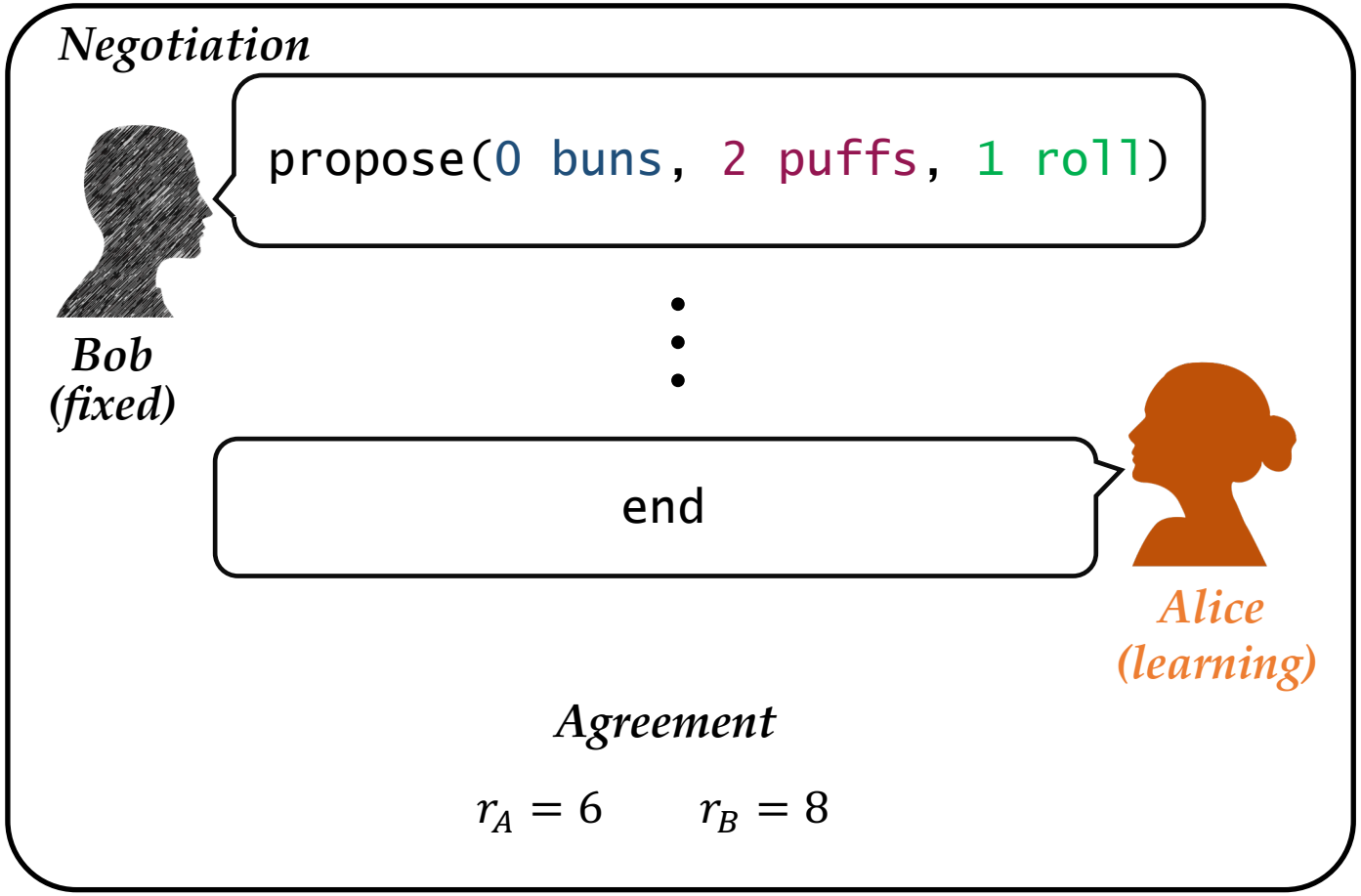
Bob
(fixed)

Alice
(learning)

Reinforcement Learning (RL)



Reinforcement Learning (RL)



Alice's utterances

For $x_t \in X^A$

$$R_A(x_t) = \gamma^{T-t}(r_A - \mu_n)$$

running mean

Reinforcement Learning (RL)

Negotiation

```
propose(0 buns, 2 puffs, 1 roll)
```

Bob (fixed)

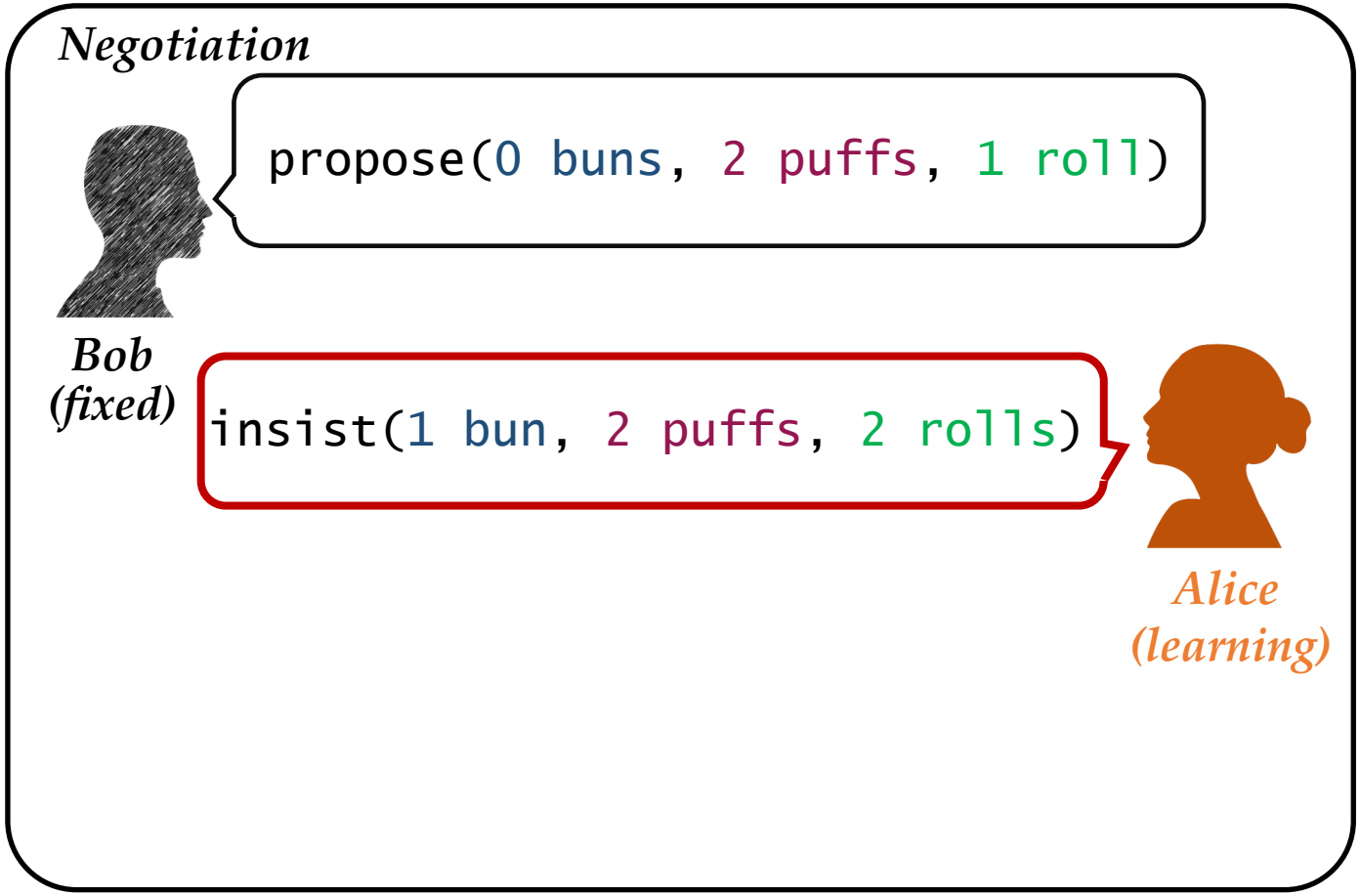
Alice's utterances

For $x_t \in X^A$

$$R_A(x_t) = \gamma^{T-t}(r_A - \mu_n)$$

running mean

Reinforcement Learning (RL)



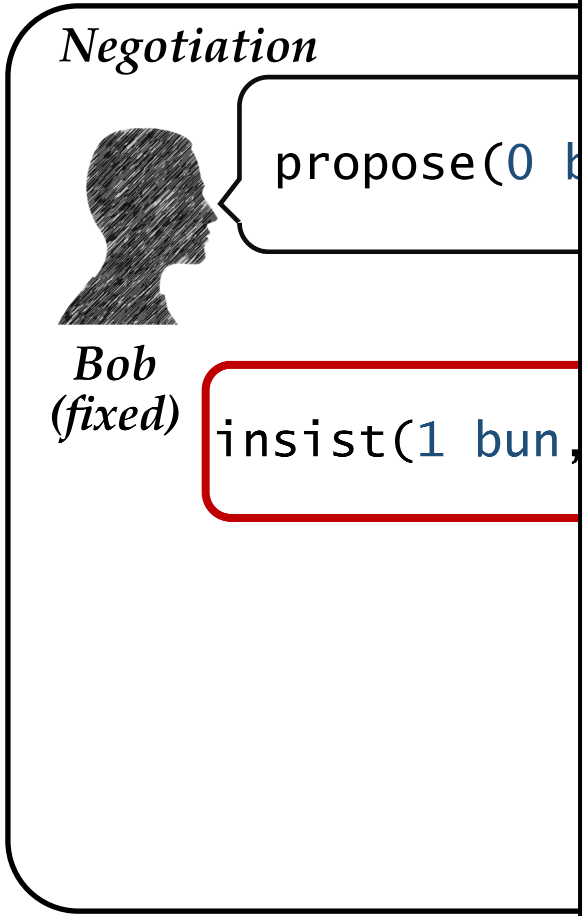
Alice's utterances

For $x_t \in X^A$

$$R_A(x_t) = \gamma^{T-t}(r_A - \mu_n)$$

running mean

Reinforce RL)



RL

Alice : insist: item0=0 item1=3 item2=1
Bob : propose: item0=1 item1=2 item2=0
Alice : propose: item0=1 item1=3 item2=1
Bob : propose: item0=1 item1=2 item2=0
Alice : propose: item0=1 item1=3 item2=1
Bob : propose: item0=1 item1=2 item2=0
Alice : propose: item0=1 item1=3 item2=1
Bob : propose: item0=1 item1=2 item2=0
Alice : propose: item0=1 item1=3 item2=1
Bob : propose: item0=1 item1=2 item2=0
Alice : propose: item0=1 item1=3 item2=1
Bob : propose: item0=1 item1=2 item2=0
Alice : propose: item0=1 item1=3 item2=1
Bob : propose: item0=1 item1=2 item2=0
Alice : propose: item0=1 item1=3 item2=1
Bob : propose: item0=1 item1=2 item2=0
Alice : propose: item0=1 item1=3 item2=1
Bob : propose: item0=1 item1=2 item2=0
Alice : <selection>
Alice : book=1 hat=3 ball=1
Bob : book=1 hat=2 ball=0

Disagreement?!
Alice : 0 (potential 10)
Bob : 0 (potential 7)

Alice's utterances

For $x_t \in X^A$

$$R_A(x_t) = \gamma^{T-t}(r_A - \mu_n)$$

running mean

Reinforce... RL)

Negotiation

Bob (fixed)

`propose(0 b`

`insist(1 bun,`

RL

Alice : insist: item0=0 item1=3 item2=1
Bob : propose: item0=1 item1=2 item2=0
Alice : propose: item0=1 item1=3 item2=1
Bob : propose: item0=1 item1=2 item2=0
Alice : propose: item0=1 item1=3 item2=1
Bob : propose: item0=1 item1=2 item2=0
Alice : propose: item0=1 item1=3 item2=1
Bob : propose: item0=1 item1=2 item2=0
Alice : propose: item0=1 item1=3 item2=1
Bob : propose: item0=1 item1=2 item2=0
Alice : propose: item0=1 item1=3 item2=1
Bob : propose: item0=1 item1=2 item2=0
Alice : propose: item0=1 item1=3 item2=1
Bob : propose: item0=1 item1=2 item2=0
Alice : propose: item0=1 item1=3 item2=1
Bob : propose: item0=1 item1=2 item2=0
Alice : propose: item0=1 item1=3 item2=1
Bob : propose: item0=1 item1=2 item2=0
Alice : <selection>
Alice : book=1 hat=3 ball=1
Bob : book=1 hat=2 ball=0

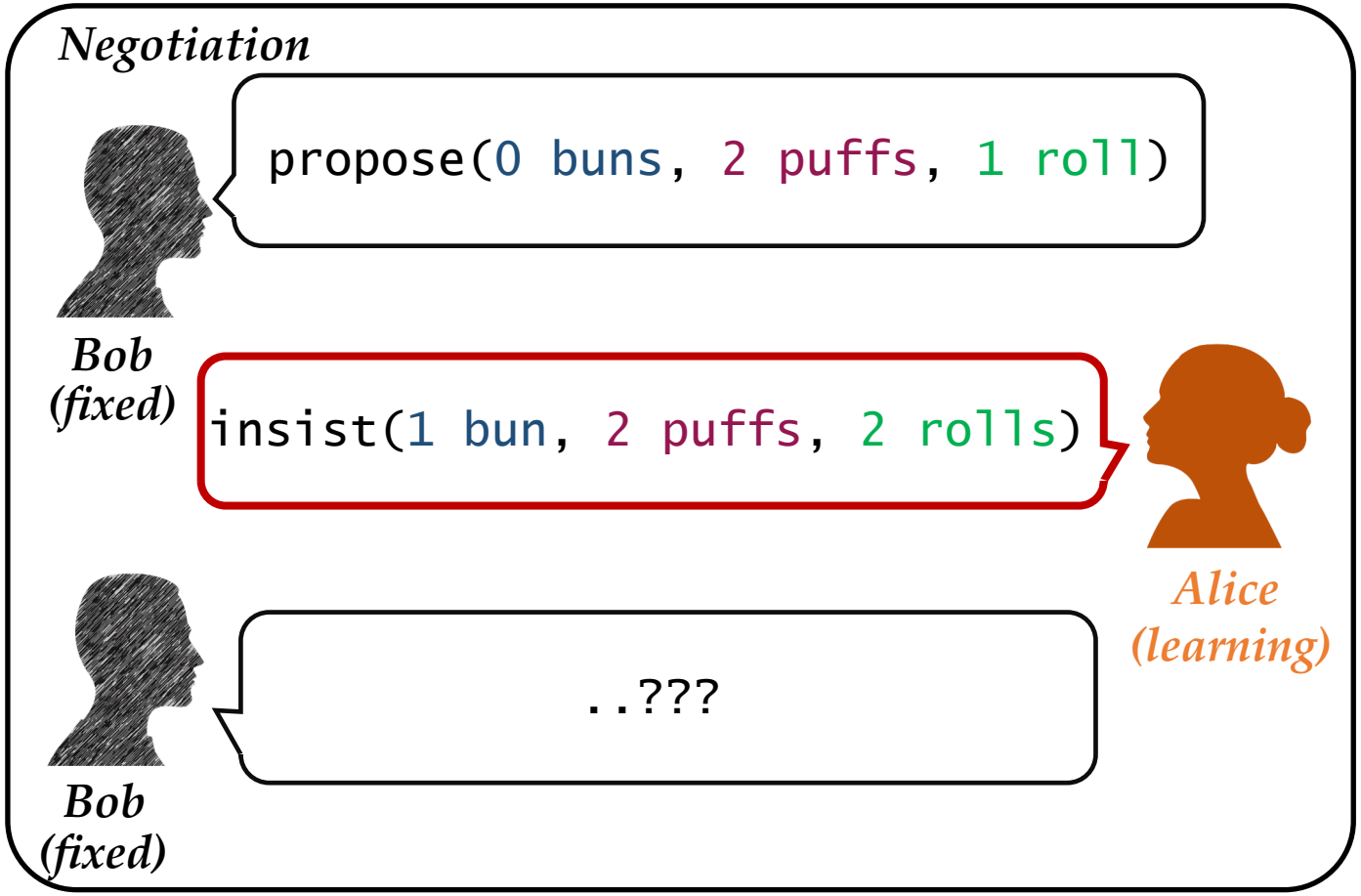
Disagreement?!
Alice : 0 (potential 10)
Bob : 0 (potential 7)

For $x_t \in X^A$ — *Alice's utterances*

$$R_A(x_t) = \gamma^{T-t}(r_A - \mu_n)$$

μ_n: *running mean*

Reinforcement Learning (RL)



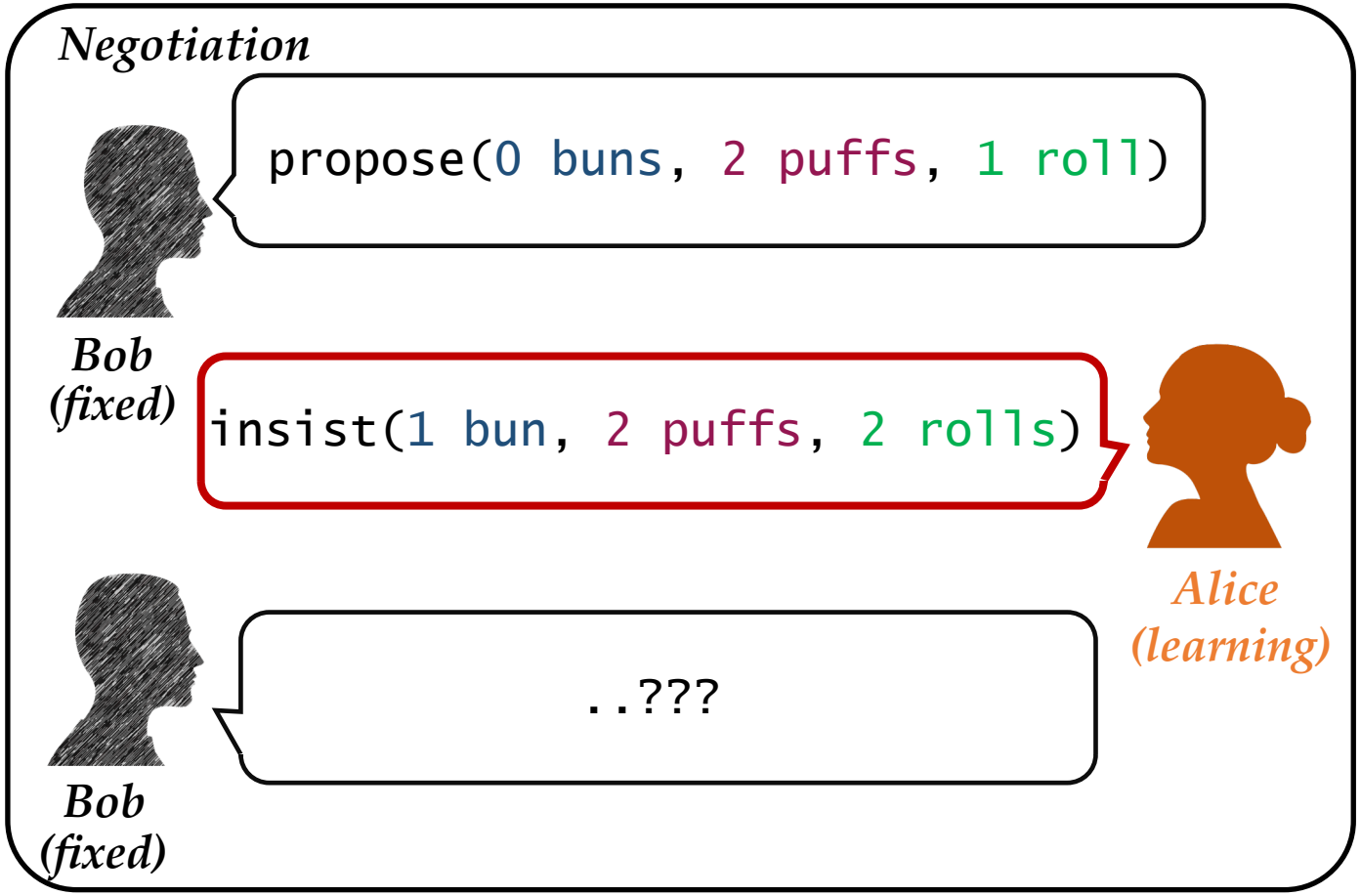
Alice's utterances

For $x_t \in X^A$

$$R_A(x_t) = \gamma^{T-t}(r_A - \mu_n)$$

running mean

Reinforcement Learning (RL)



Alice's utterances

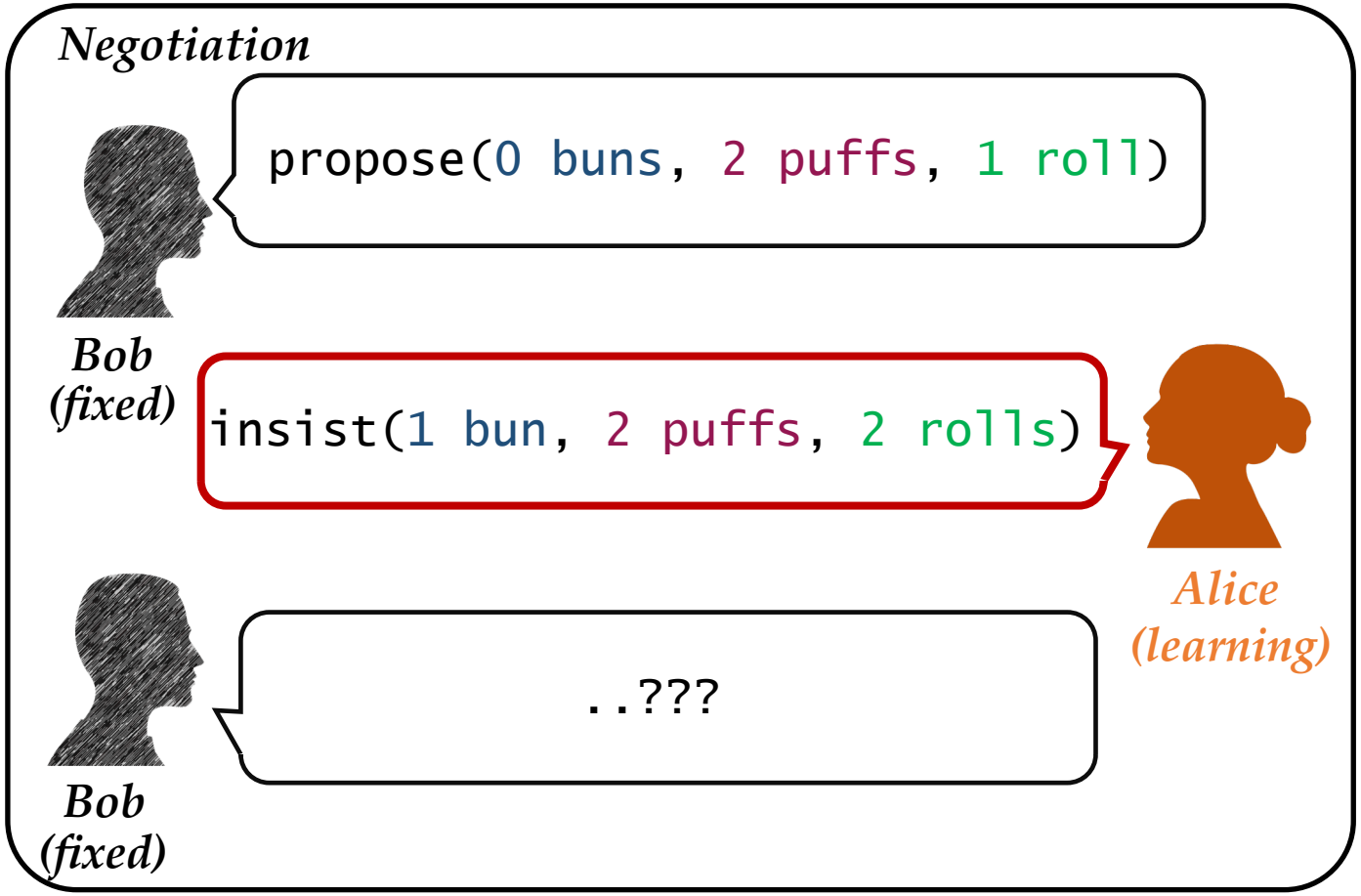
For $x_t \in X^A$

$$R_A(x_t) = \gamma^{T-t}(r_A - \mu_n)$$

running mean

Relationship to dataset: Alice inherits dataset biases through Bob

Reinforcement Learning (RL)



Alice's utterances

For $x_t \in X^A$

$$R_A(x_t) = \gamma^{T-t}(r_A - \mu_n)$$

running mean

Relationship to dataset: Alice inherits dataset biases through Bob

Mixed RL, SL (RL+SL)

Interleave SL training every nth timestep

- n=1: RL, SL, RL, SL …
- n=2: RL, RL, SL, RL, RL, SL …

Mixed RL, SL (RL+SL)

Interleave SL training every nth timestep

- n=1: RL, SL, RL, SL …
- n=2: RL, RL, SL, RL, RL, SL …

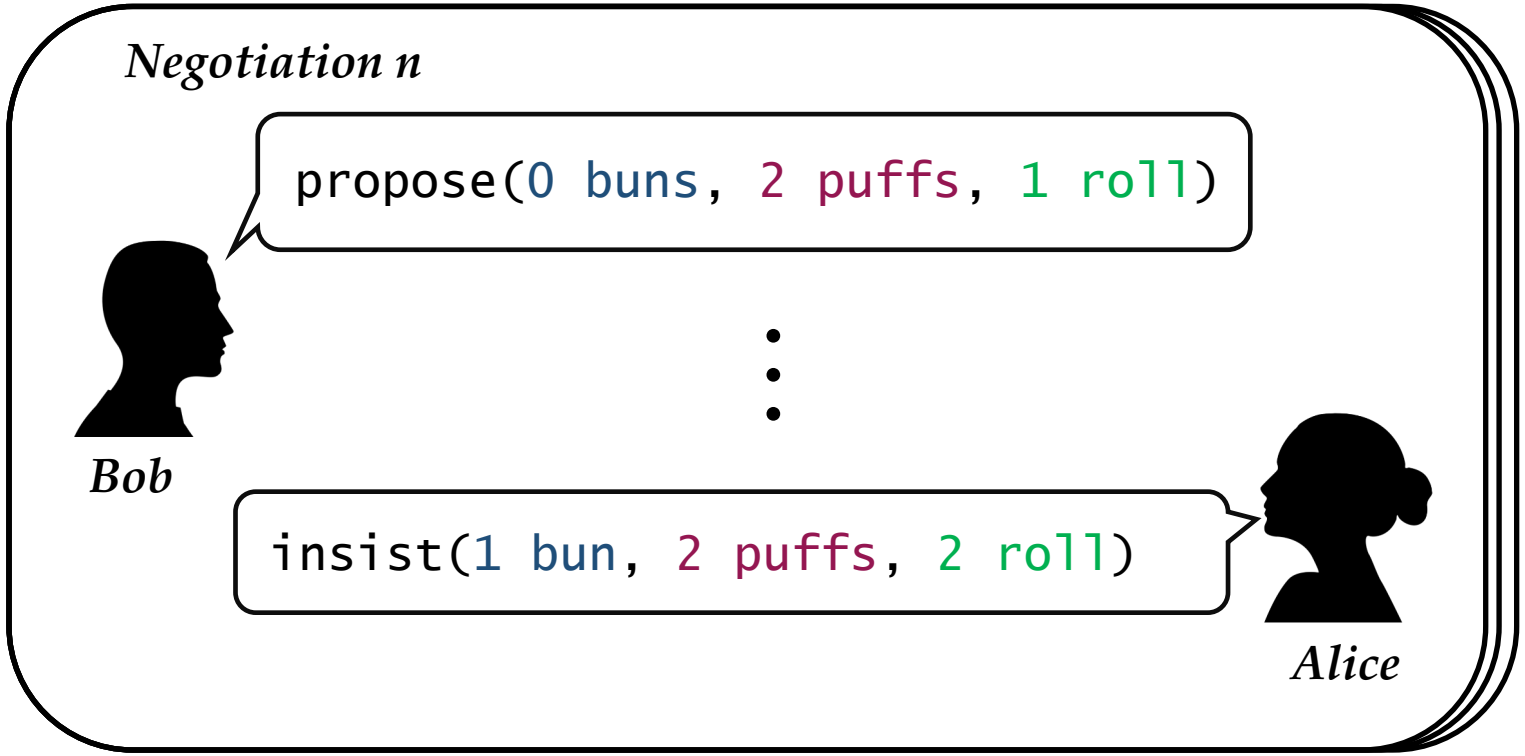
Relationship to dataset: same as SL, bias inherited from dataset

Problem: Low-quality, static datasets!

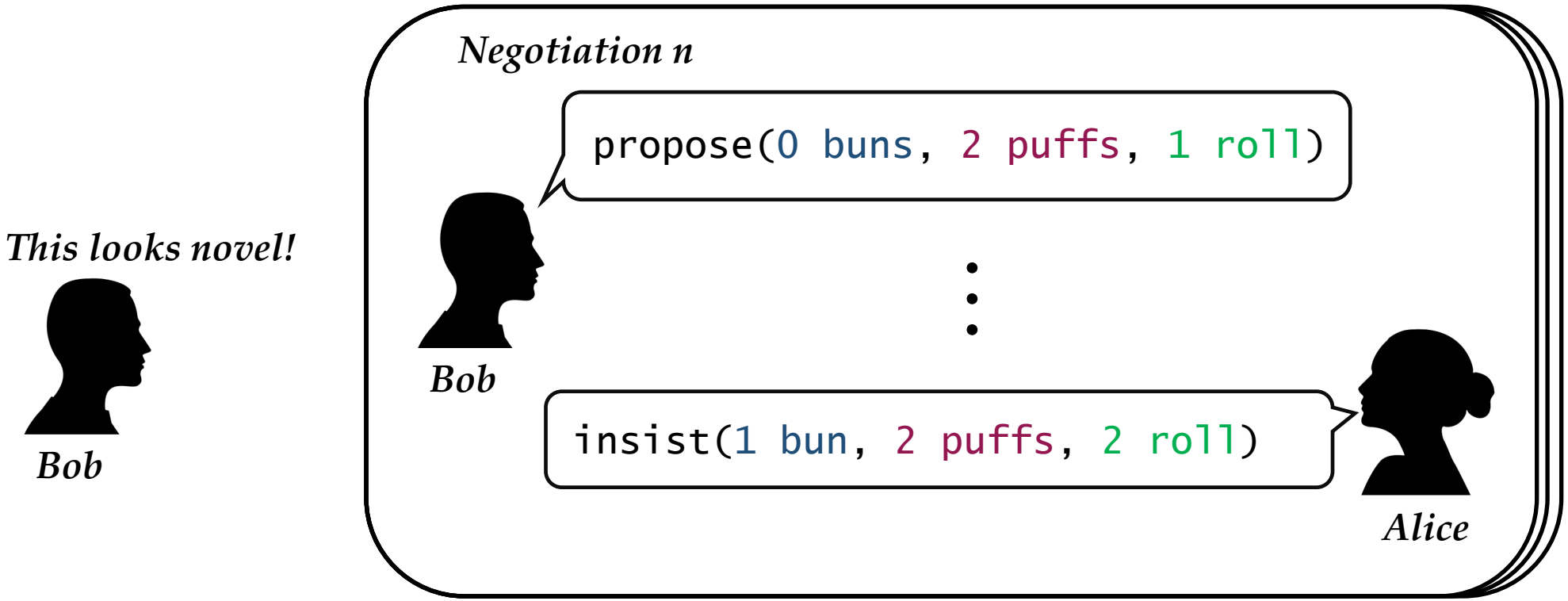
Problem: Low-quality, static datasets!

Key Insight:
Continually improve Bob with expert data!

Targeted Data Acquisition Framework

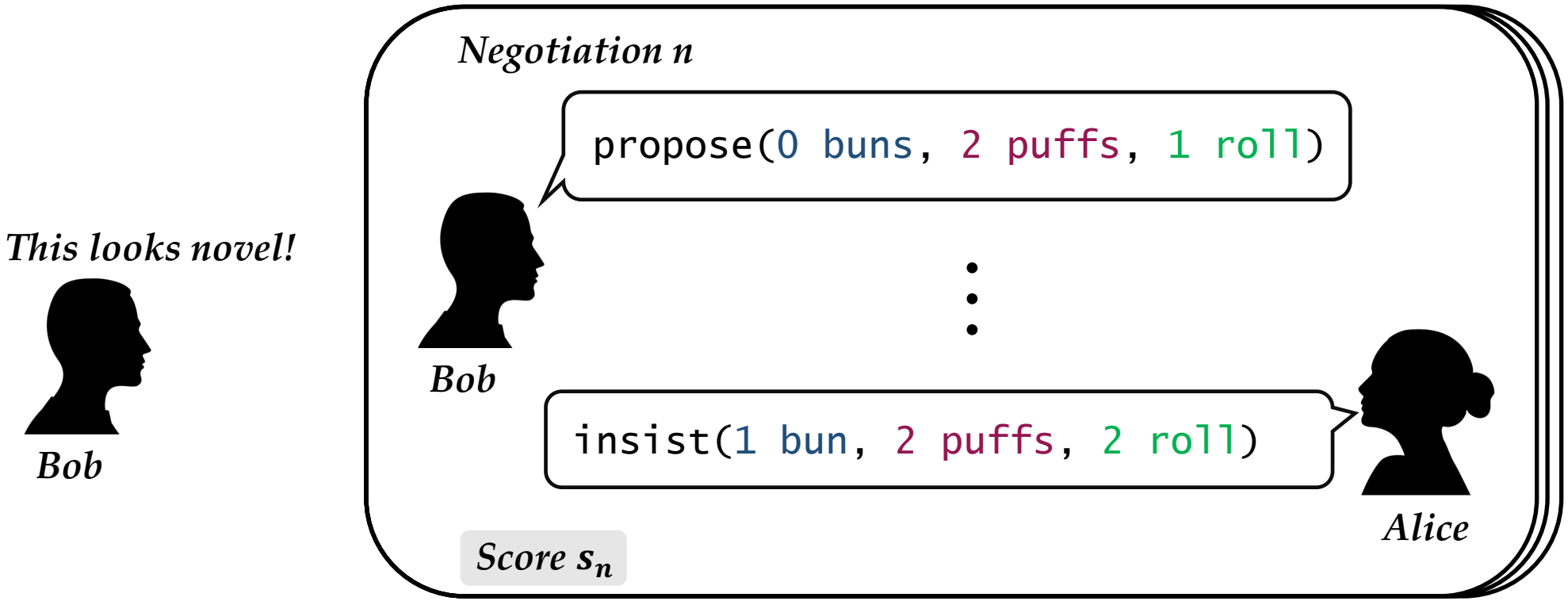


Targeted Data Acquisition Framework



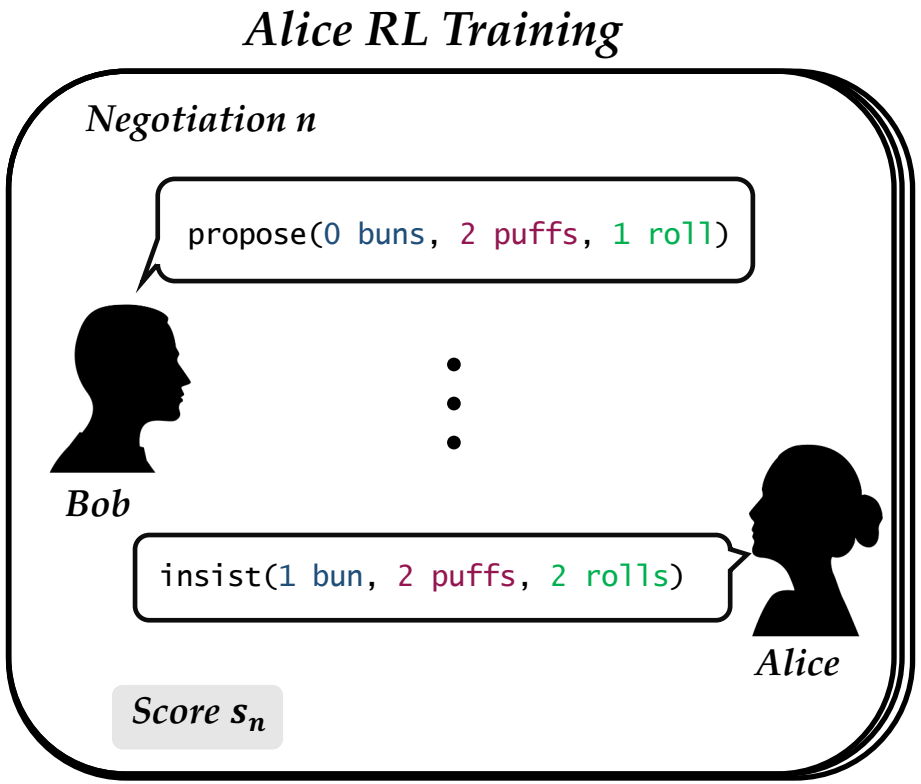
Novelty score: $s_n = \min_{x_t \in X^A} \log p_\theta(x_t | x_{0:t-1}, c^A)$

Targeted Data Acquisition Framework

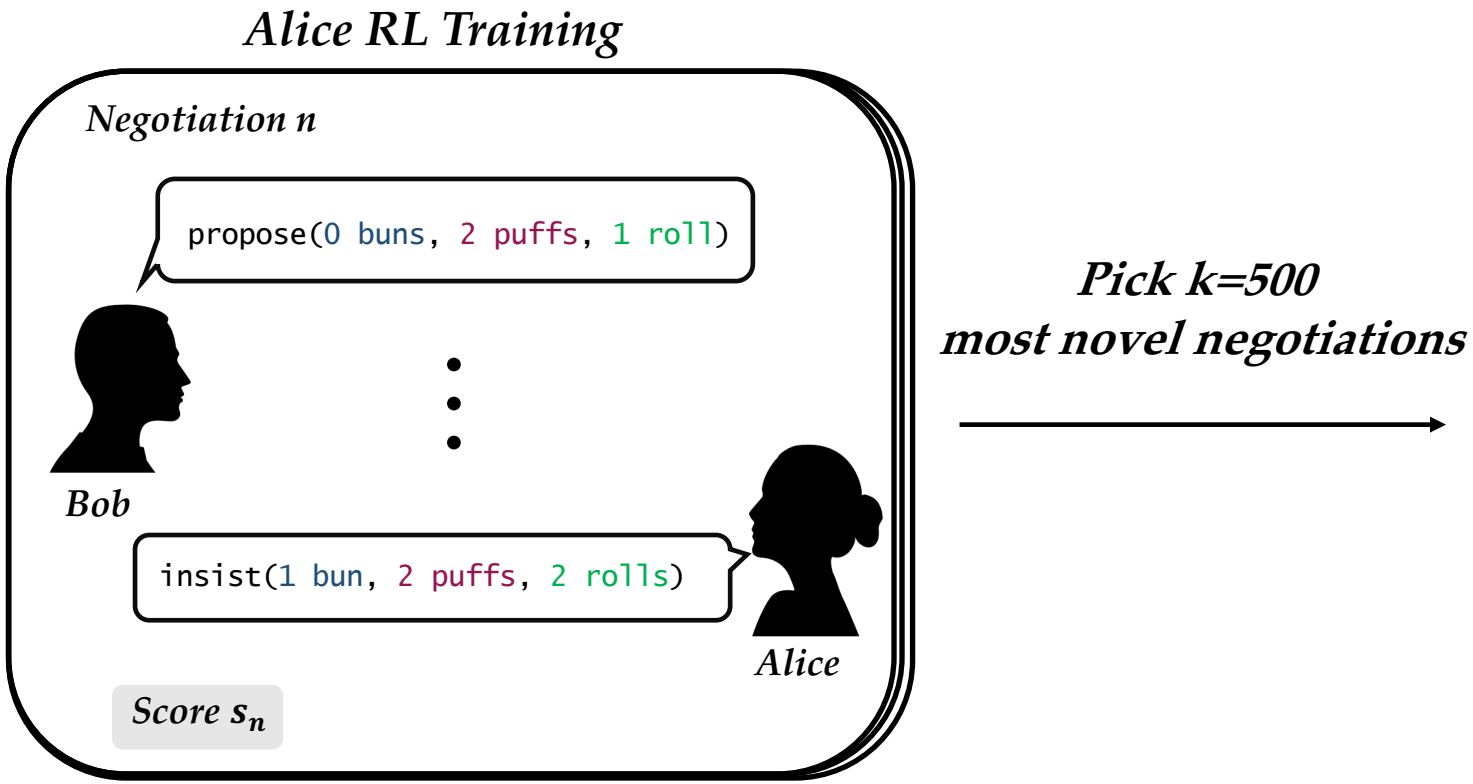


Novelty score: $$s_n = \min_{x_t \in X^A} \log p_\theta(x_t | x_{0:t-1}, c^A)$$

Targeted Data Acquisition Framework

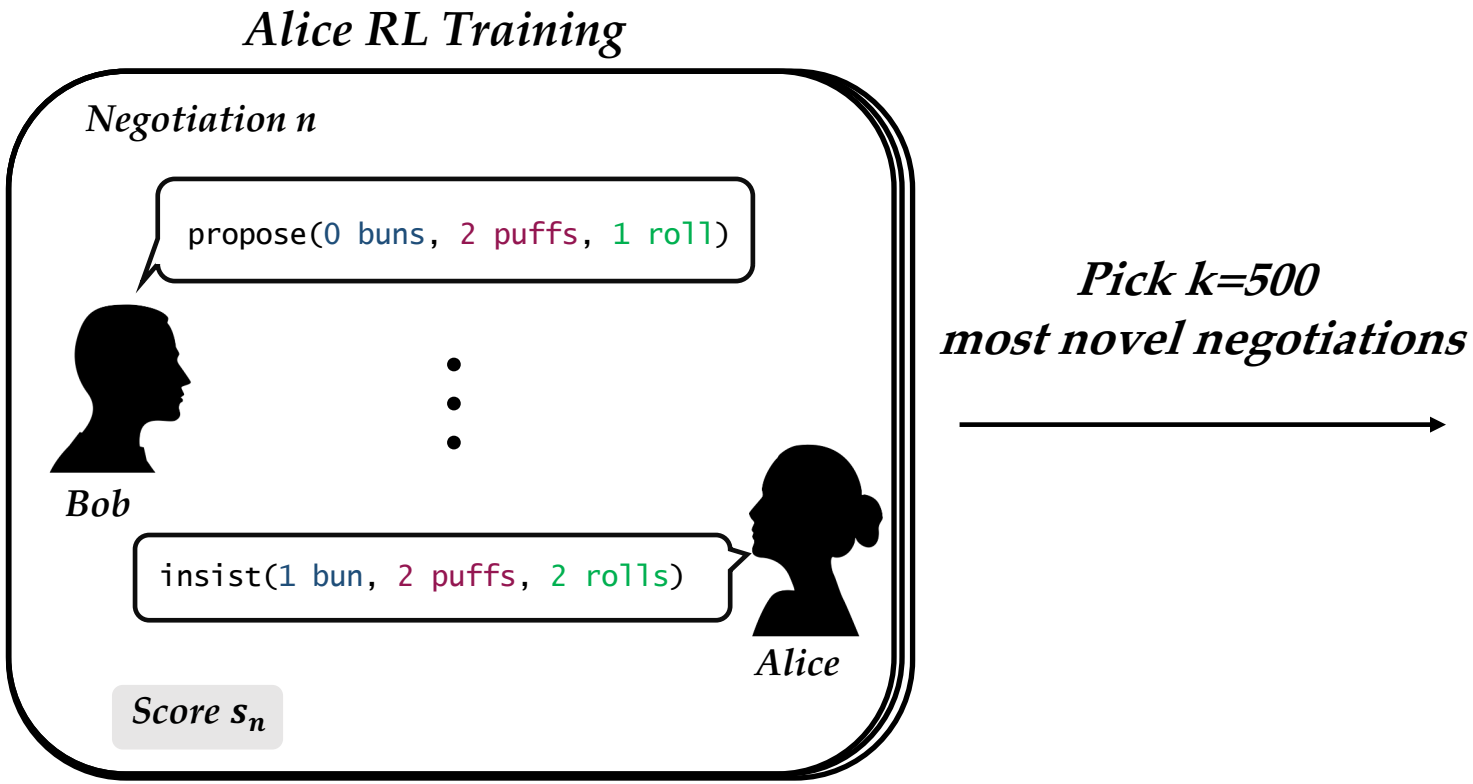


Targeted Data Acquisition Framework



Pick k=500 most novel negotiations

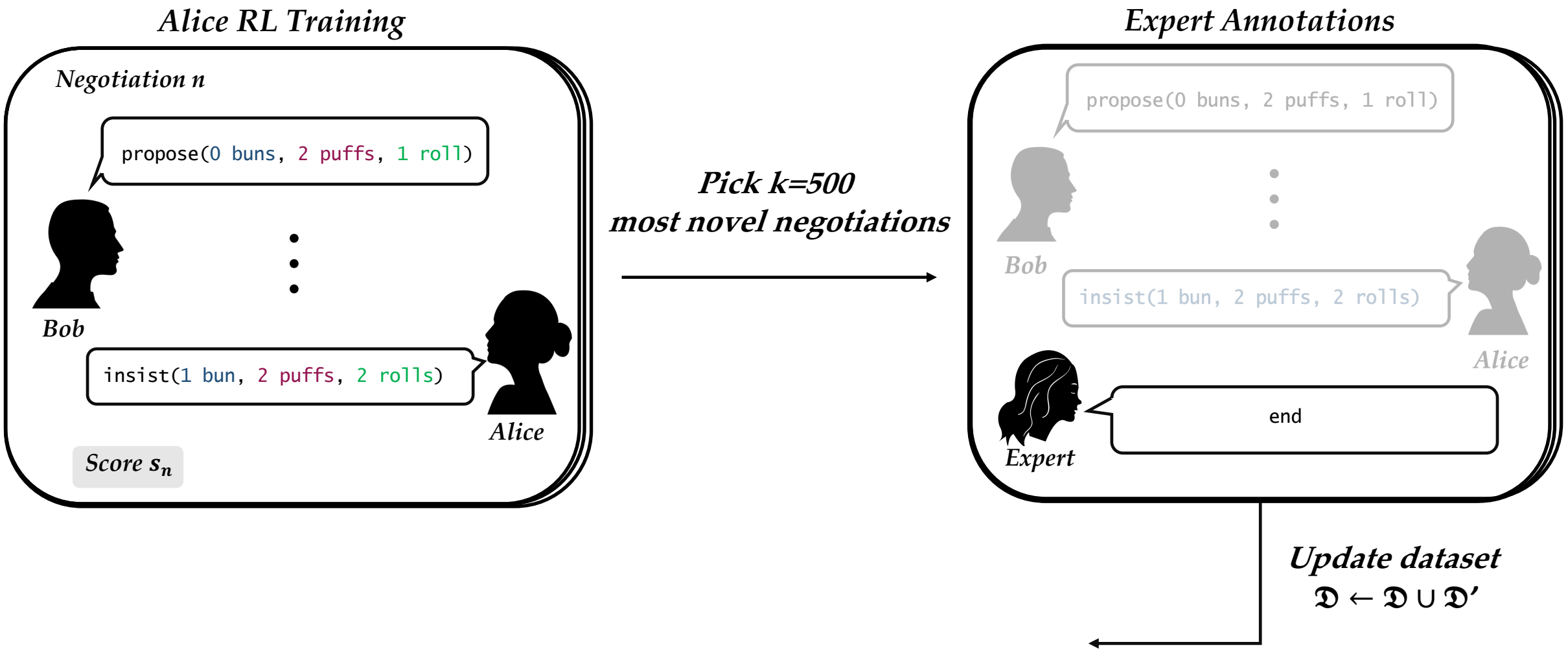
Targeted Data Acquisition Framework



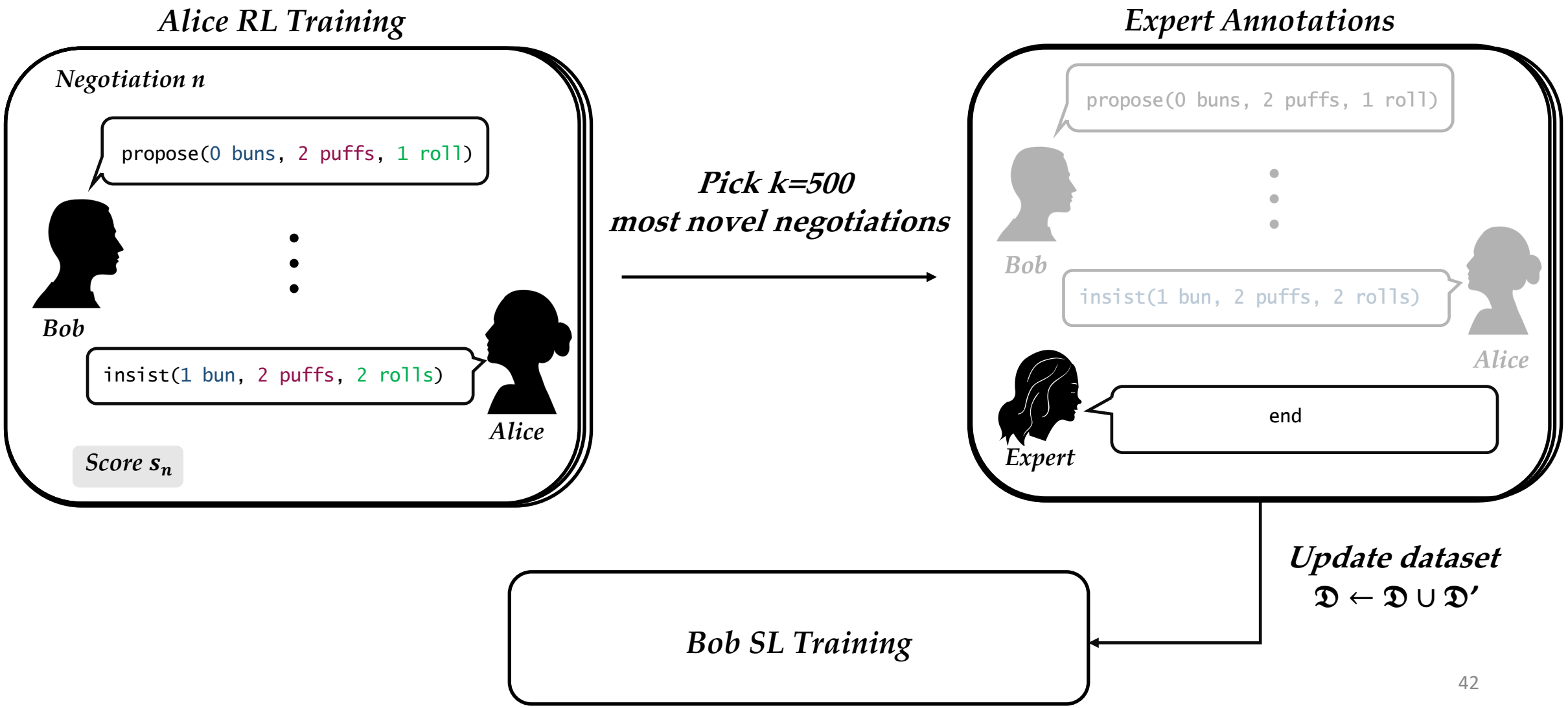
Pick k=500 most novel negotiations

Targeted Data Acquisition Framework

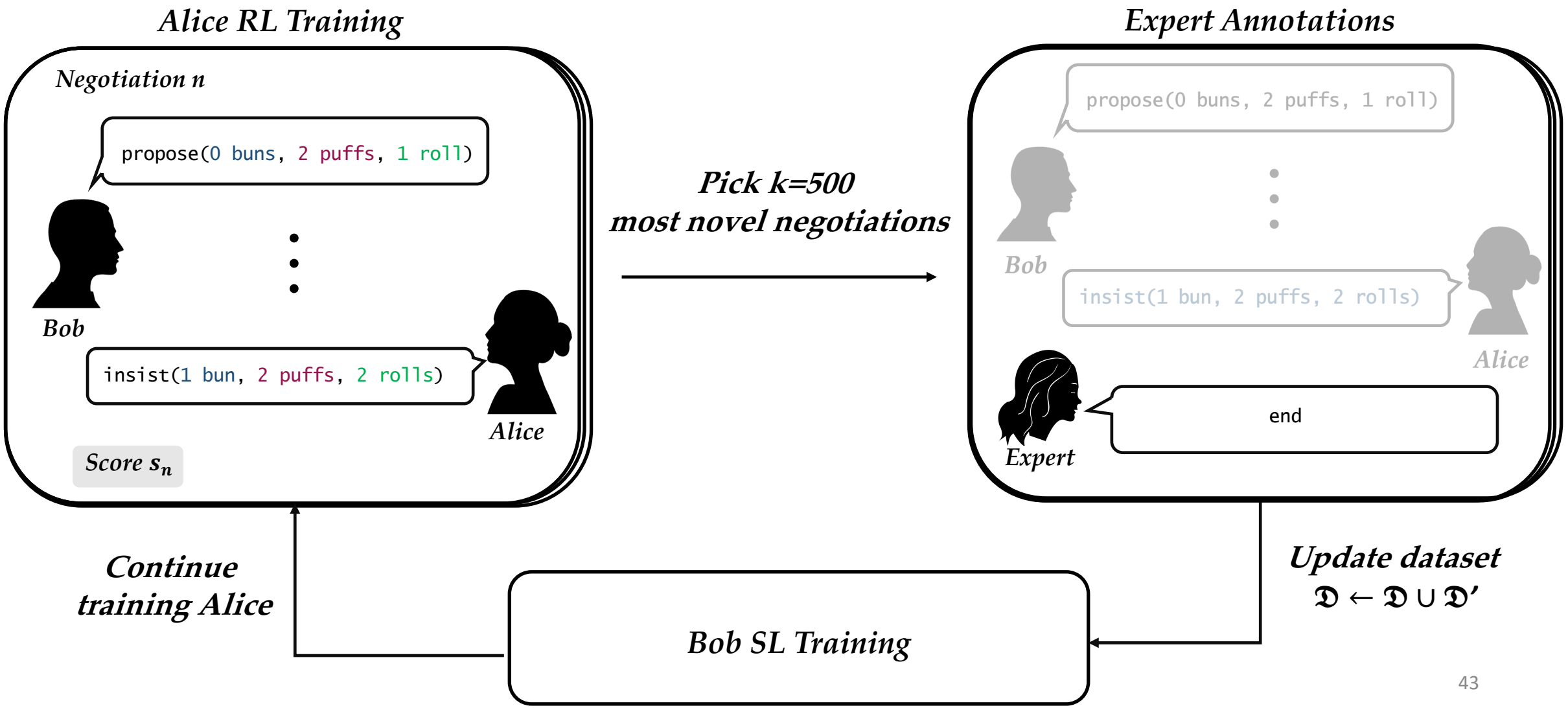
Targeted Data Acquisition Framework



Targeted Data Acquisition Framework



Targeted Data Acquisition Framework



Evaluation

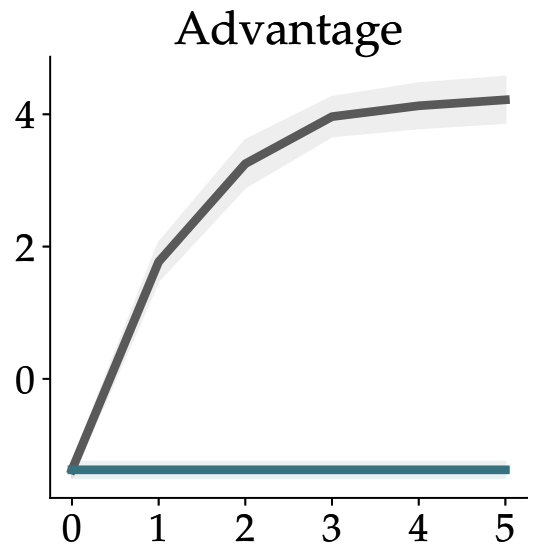
Can we balance self-interest and Pareto-optimality?

Results with a Simulated Partner

(higher is better)

Results with a Simulated Partner

(higher is better)

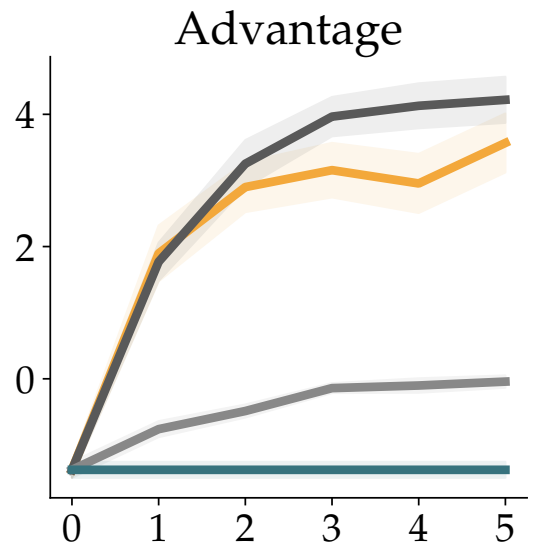


(D1) Self-interest

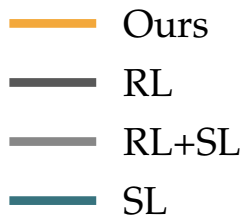
Ours
RL
RL+SL
SL

Results with a Simulated Partner

(higher is better)

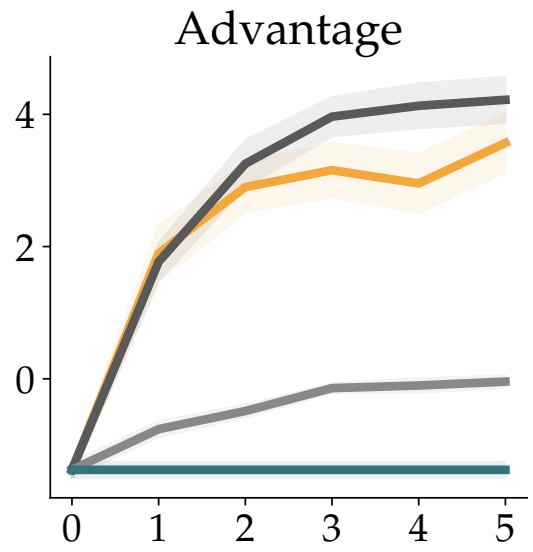


(D1) Self-interest



Results with a Simulated Partner

(higher is better)

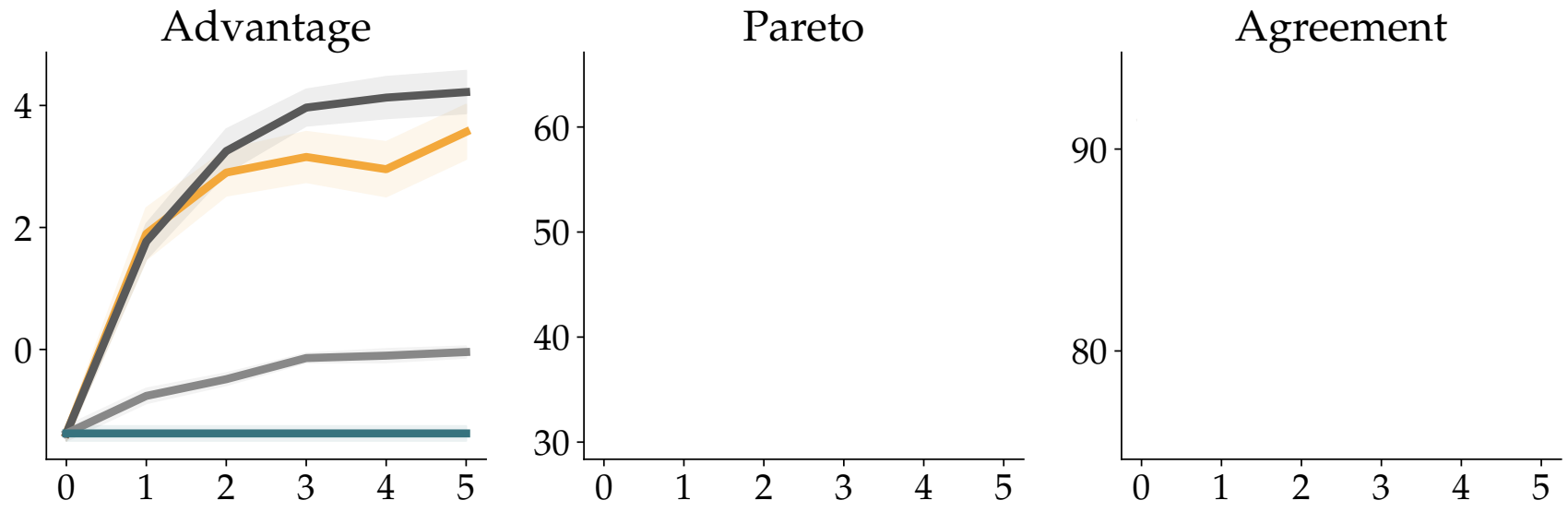


(D1) Self-interest✓

Ours
RL
RL+SL
SL

Results with a Simulated Partner

(higher is better)

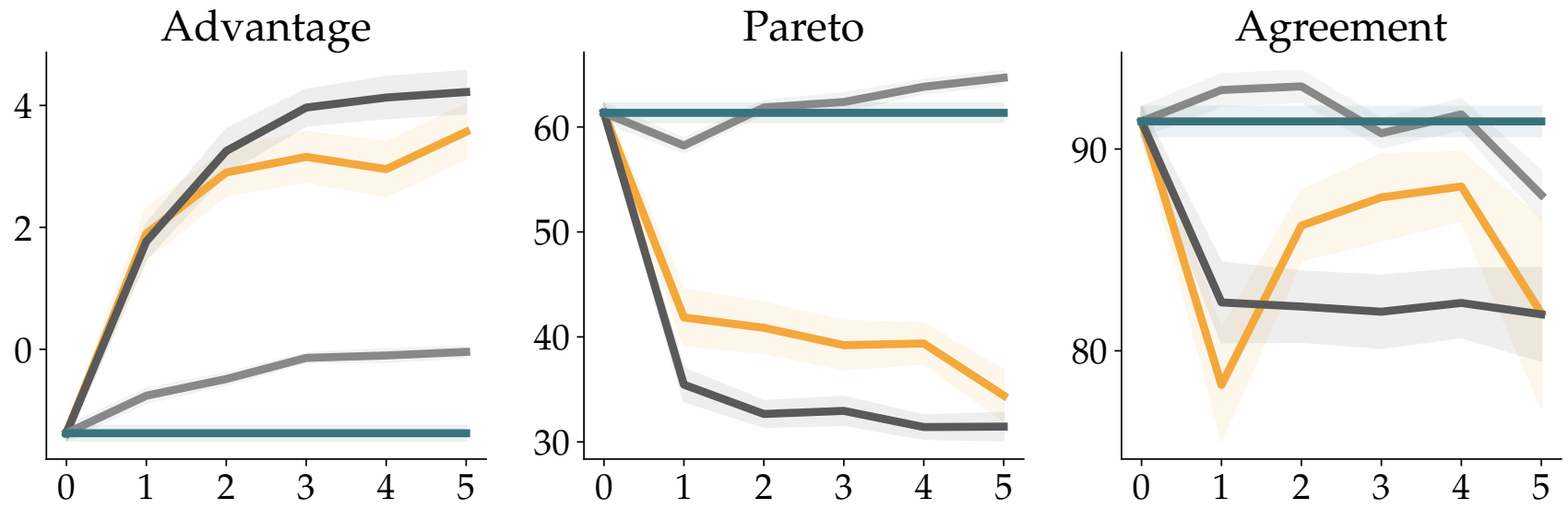


(D1) Self-interest✓

(D2) Pareto-Optimal

Results with a Simulated Partner

(higher is better)

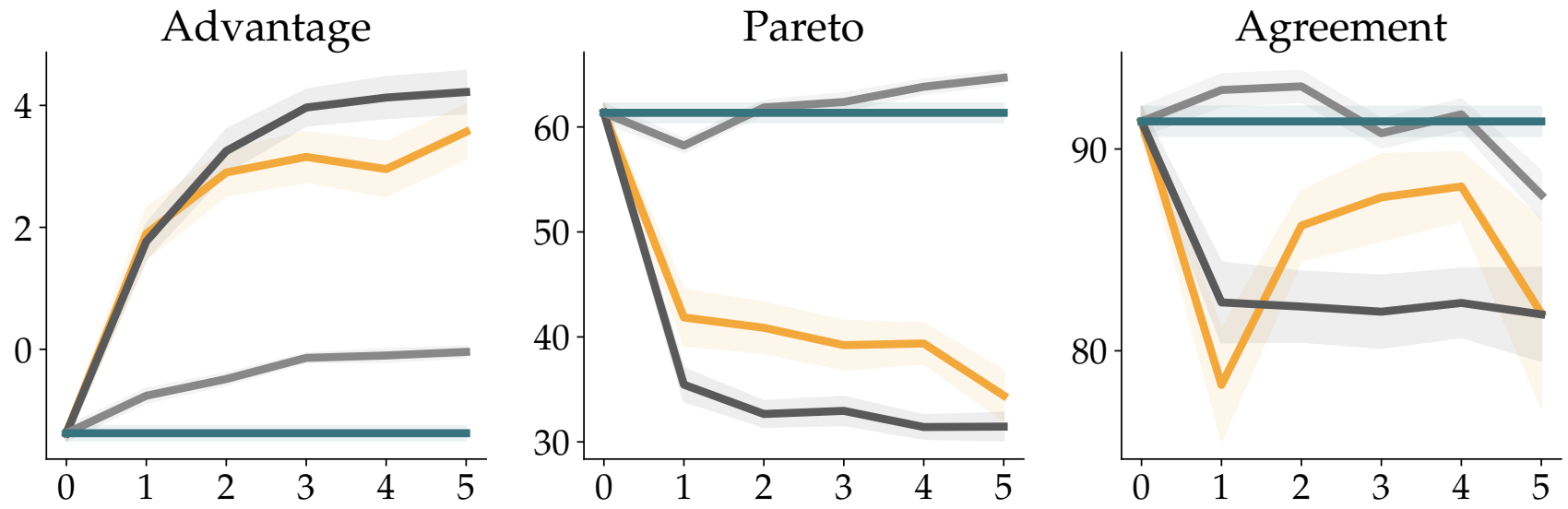


(D1) Self-interest✓

(D2) Pareto-Optimal

Results with a Simulated Partner

(higher is better)

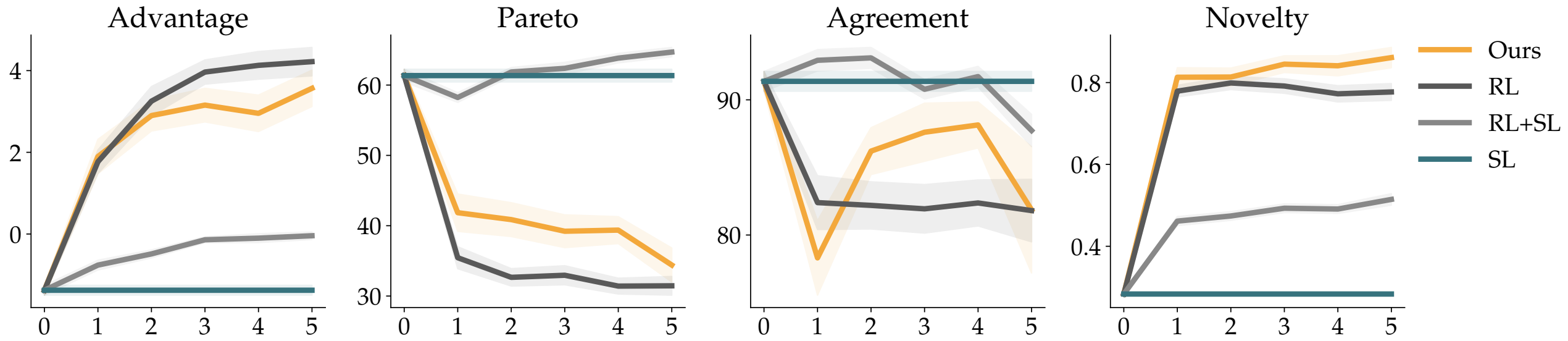


(D1) Self-interest✓

(D2) Pareto-Optimal ✓

Results with a Simulated Partner

(higher is better)



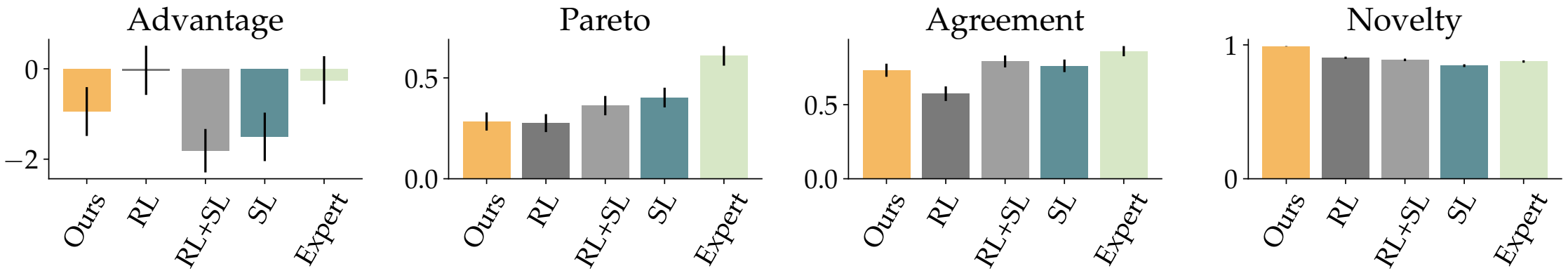
(D1) Self-interest✓

(D2) Pareto-Optimal ✓

Results with a Human Partner

N=101

(higher is better)



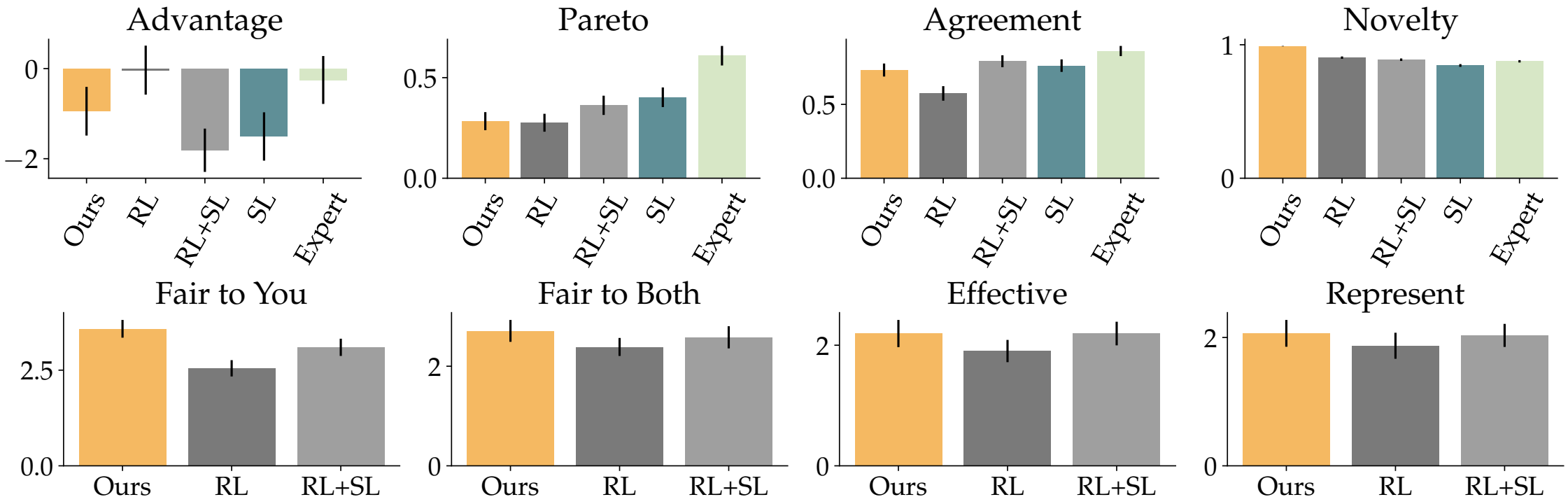
(D1) Self-interest ✓

(D2) Pareto-Optimal ✓

Results with a Human Partner

N=101

(higher is better)



Main Ideas

- Our approach balances self-interest and Pareto-optimality the best.
- This holds true against both simulated and human partners.