On The Critical Role of Conventions in Human-AI Collaboration

Andy Shih, Arjun Sawhney, Jovana Kondic, Stefano Ermon, Dorsa Sadigh

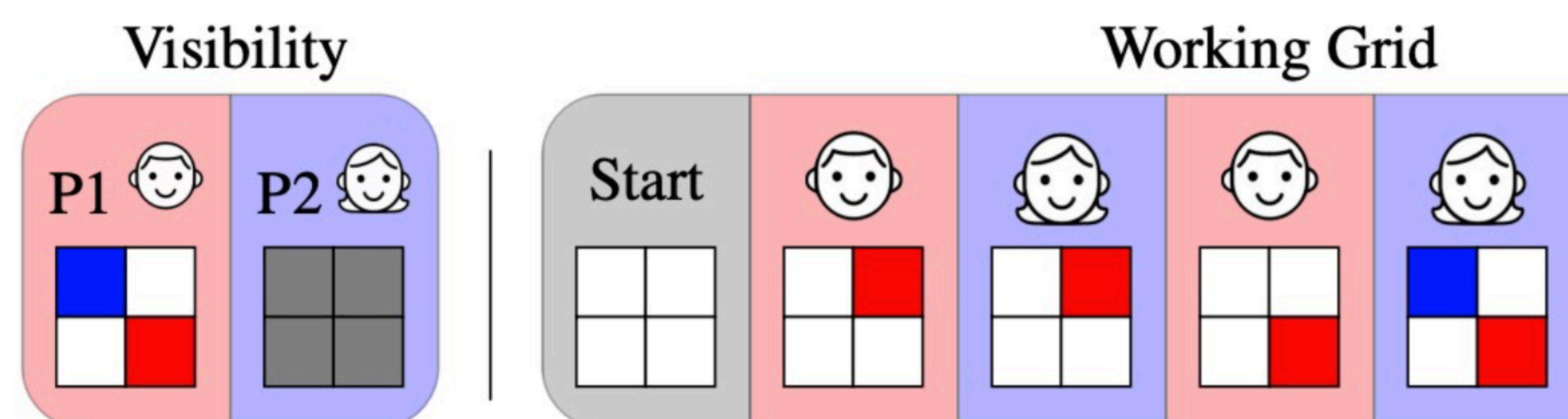andyshih@cs.stanford.edu

Conventions: **Shared representation that breaks symmetry between equally optimal strategies (developed over repeated interactions)**
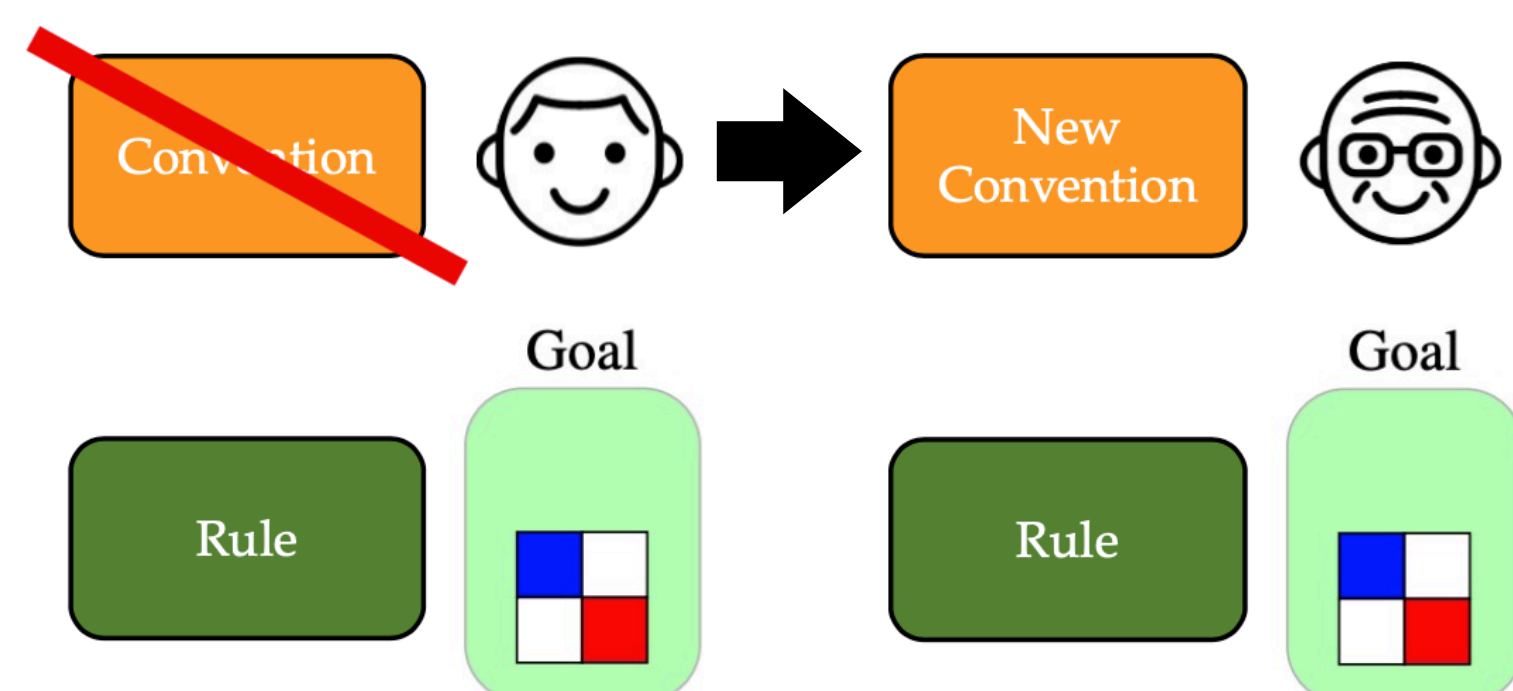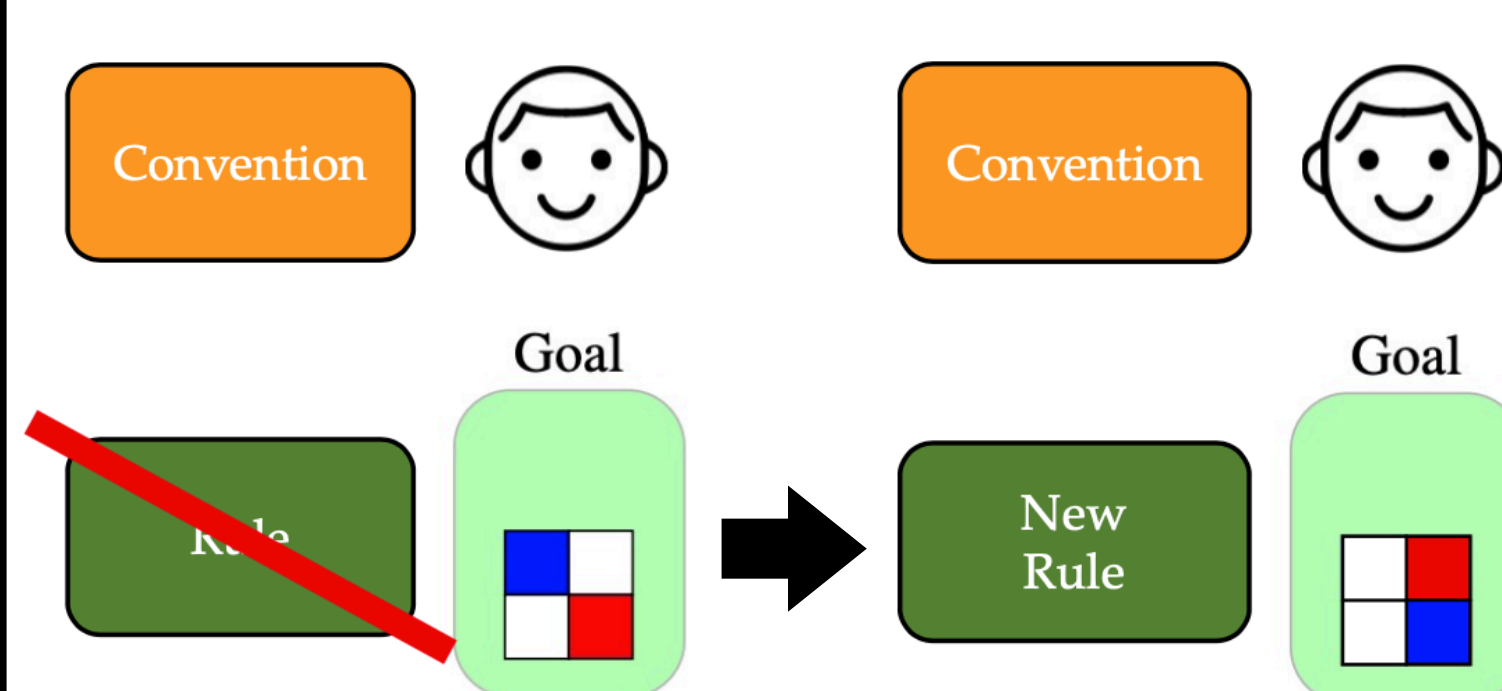
Collaborative Block Placing

Visibility | Working Grid

Conventions: Signaling location of blue block

Rule: Placing red block correctly

New Partner

Convention → New Convention

Goal | Goal

Rule | Rule

New Task

Convention | Convention

Goal | Goal

Rule → New Rule

Modular Architecture

Multiply the action distributions

State s

task module

partner module

g^t

$g^t(a|z)$

g_1^p ... g_n^p

$g_1^p(a|z)$ ... $g_n^p(a|z)$

$$\pi_1(a|s) = \frac{1}{Z_1} g^t(a|z) g_1^p(a|z)$$

$$\pi_n(a|s) = \frac{1}{Z_n} g^t(a|z) g_n^p(a|z)$$

$$D(s) = \sum_{a \in A} |g^t(a|s) - \frac{1}{n}\sum_i \pi_i(a|s)|$$
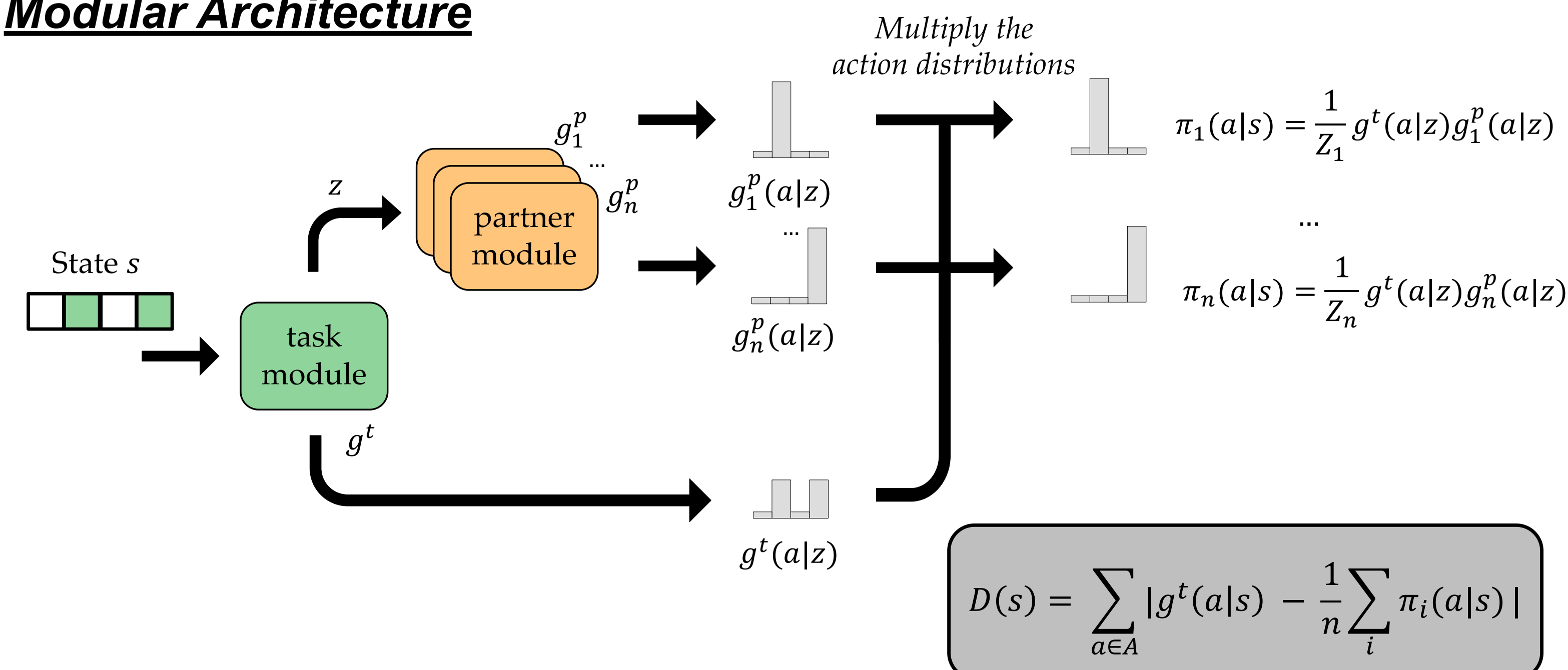
Experiments

Our Method — BaselineAgg — BaselineAgg EarlyStopping — BaselineModular — FOMAML

arms

blocks

Hanabi