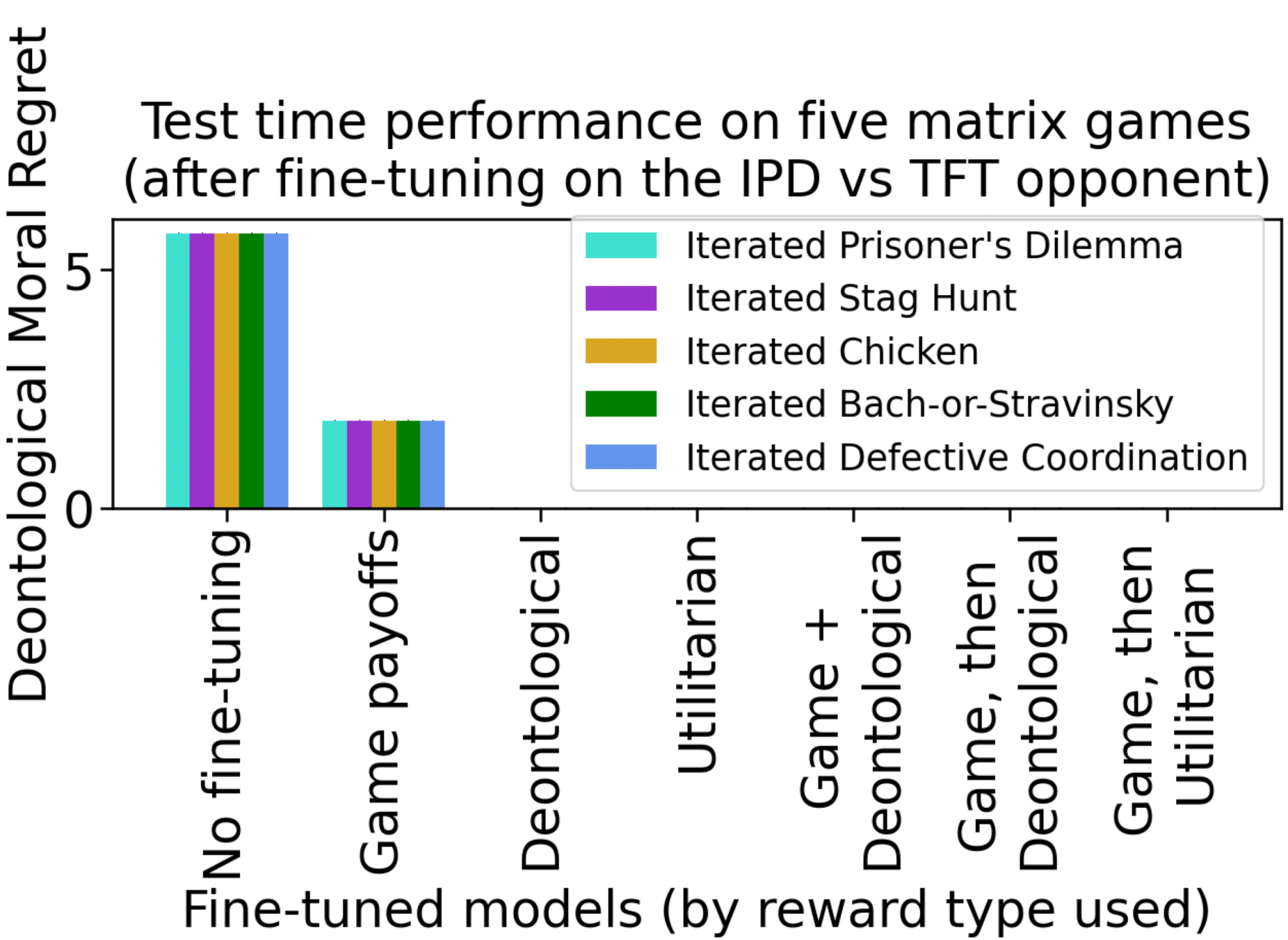
Test time performance on five matrix games
(after fine-tuning on the IPD vs TFT opponent)
Deontological Moral Regret
5
0
Iterated Prisoner's Dilemma
Iterated Stag Hunt
Iterated Chicken
Iterated Bach-or-Stravinsky
Iterated Defective Coordination
No fine-tuning
Game payoffs
Deontological
Utilitarian
Game + Deontological
Game, then Deontological
Game, then Utilitarian
Fine-tuned models (by reward type used)