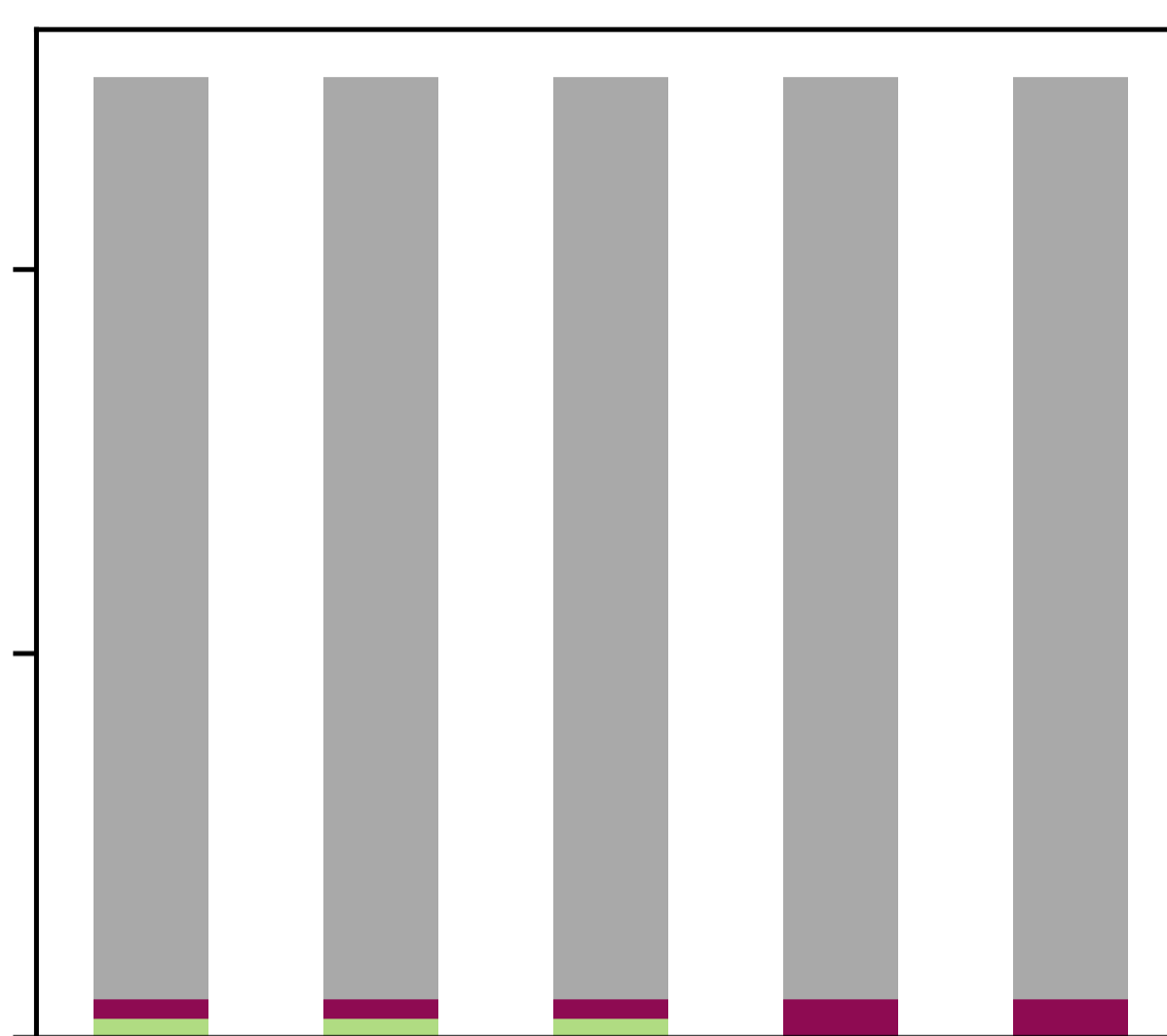


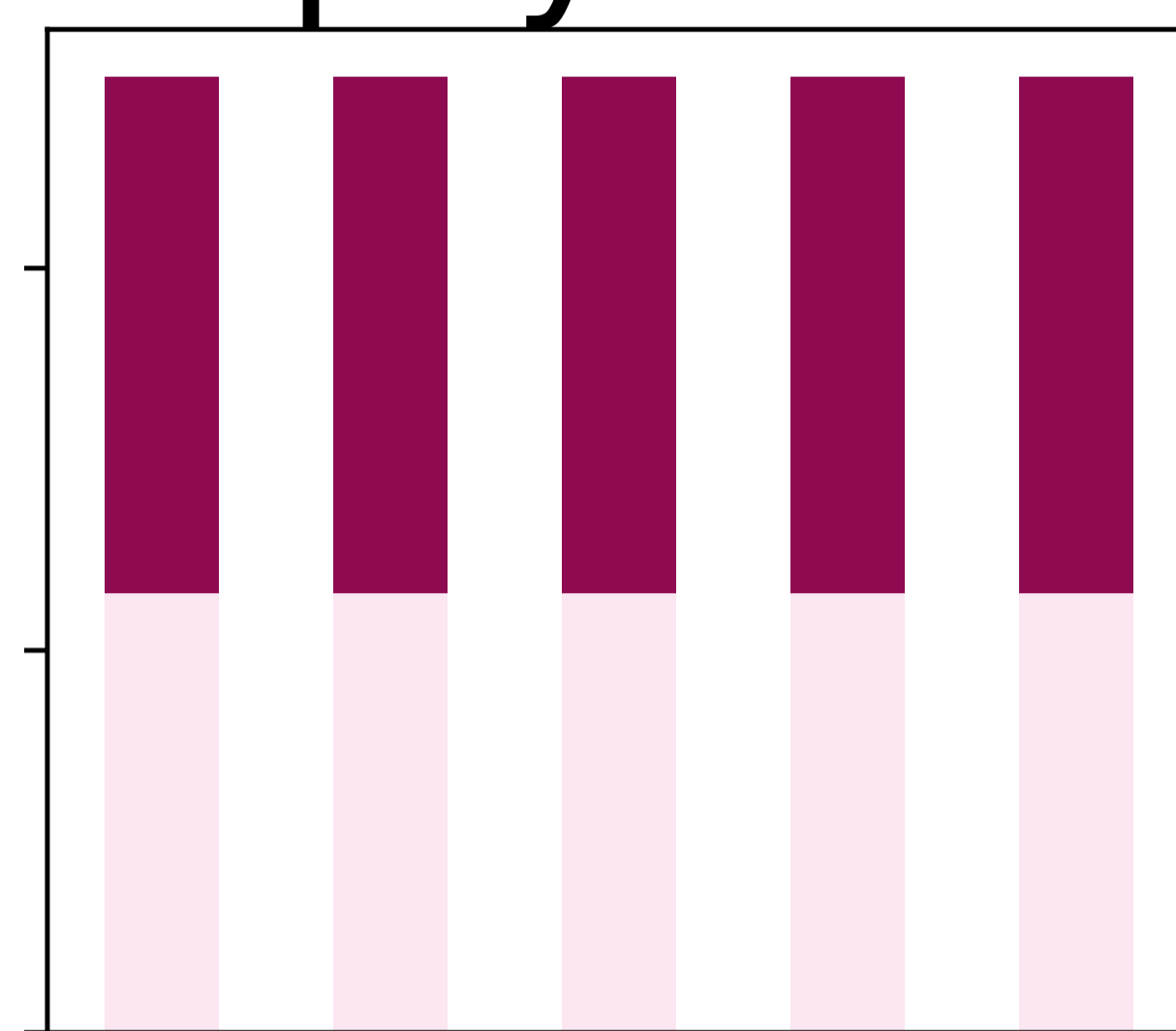
M's action | O's prev. move
(% of test time responses)

Action choices on five iterated matrix games
(all models trained vs TFT opponent)

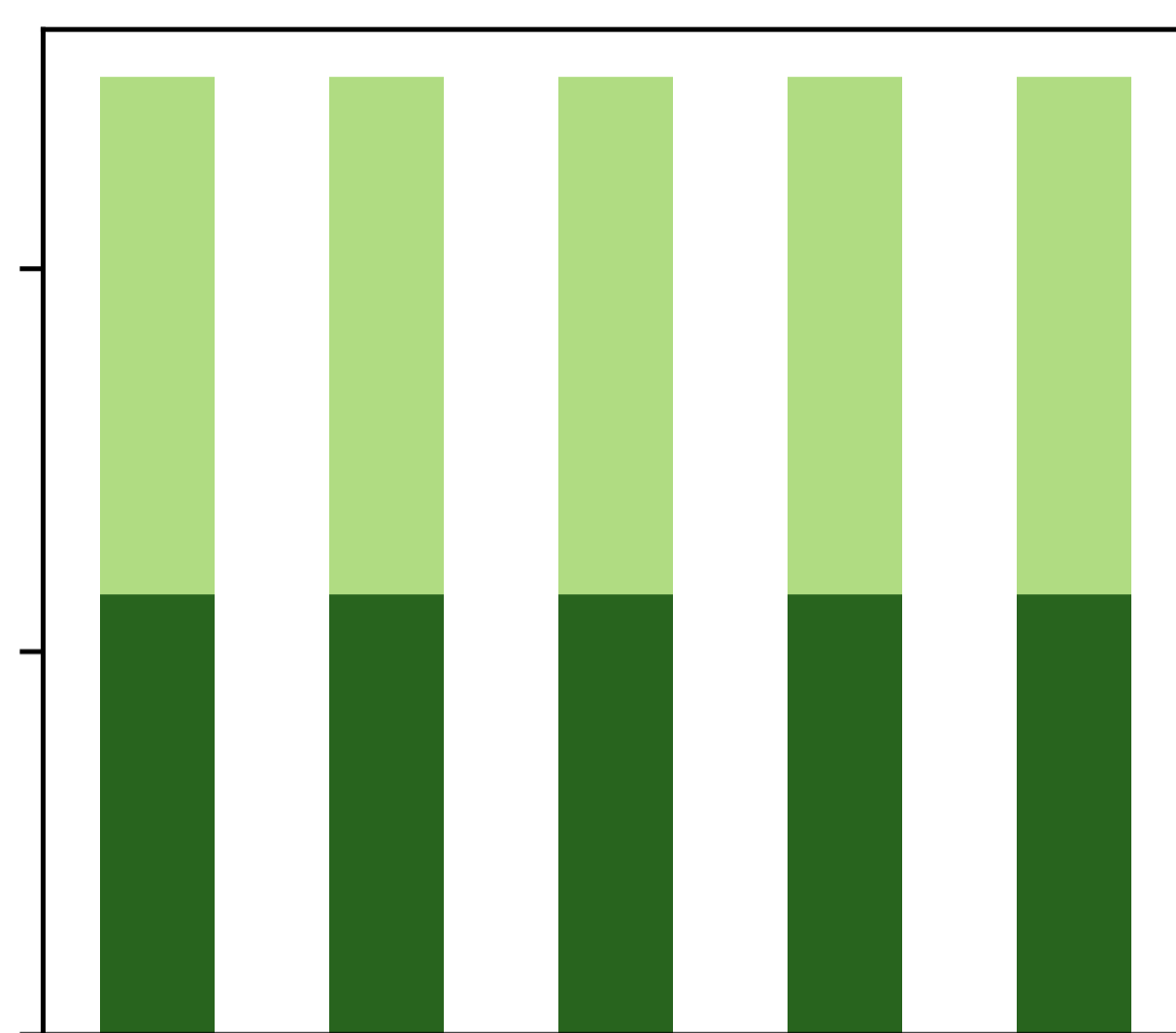
No fine-tuning



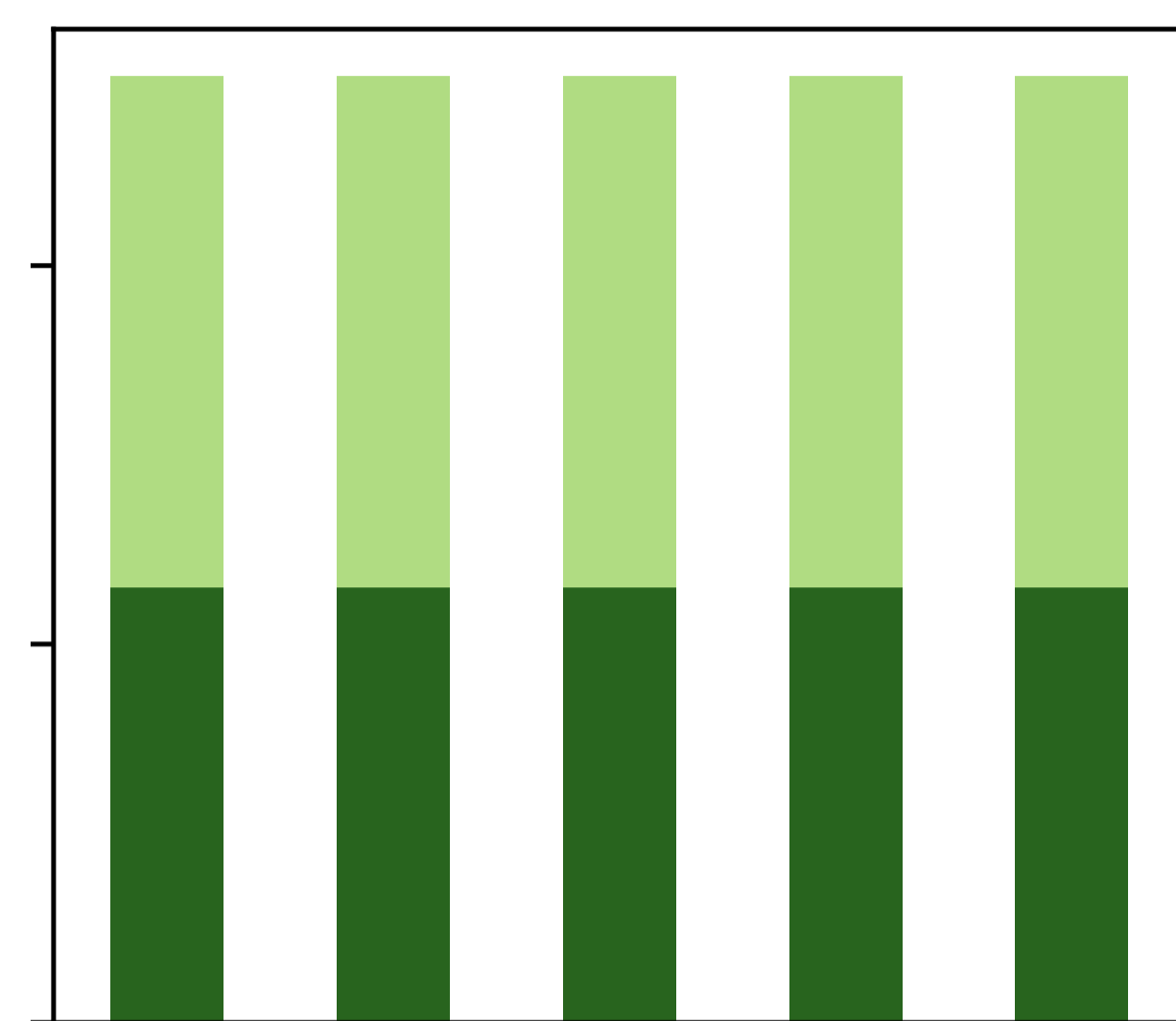
Game
payoffs



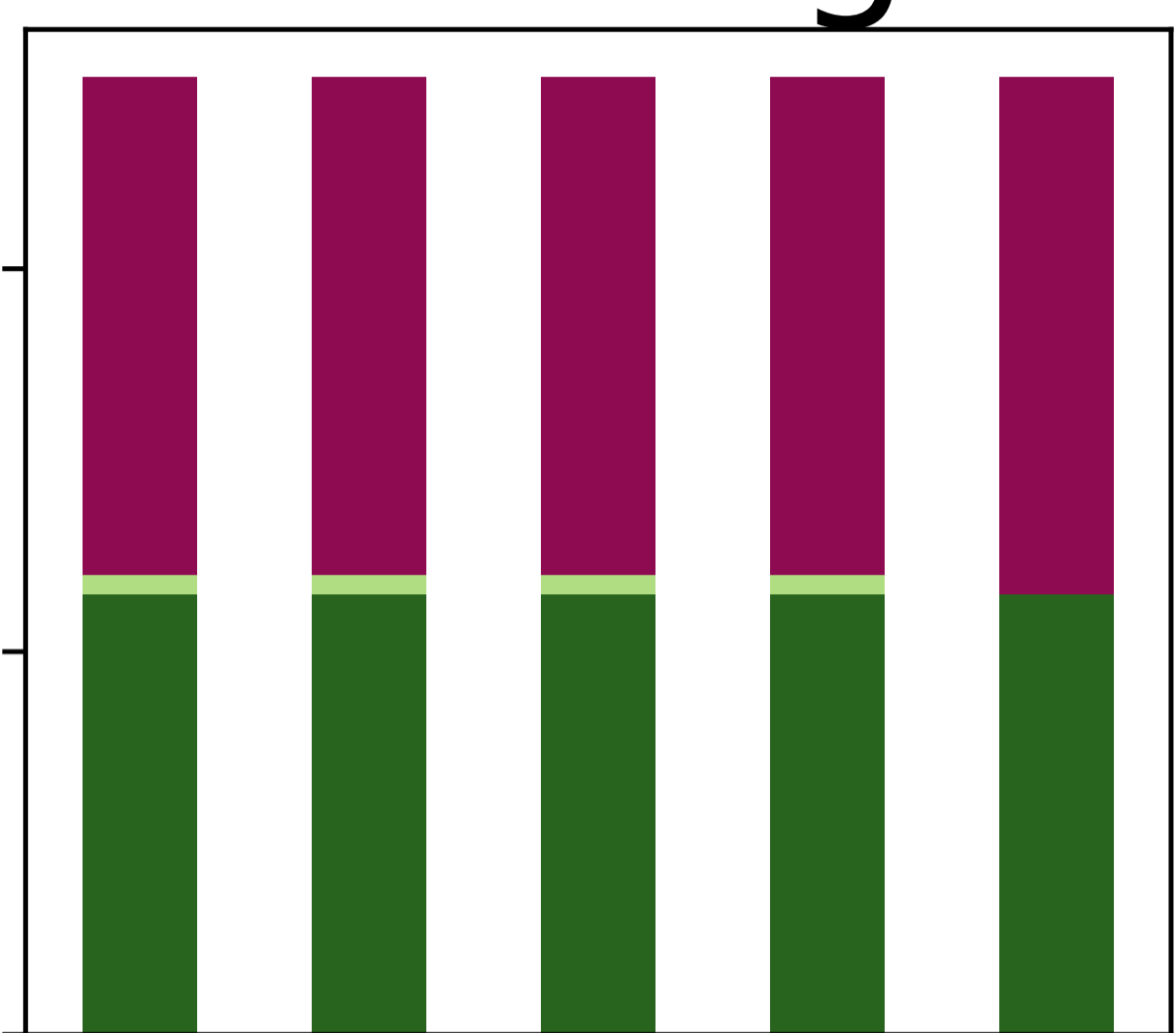
Deontological



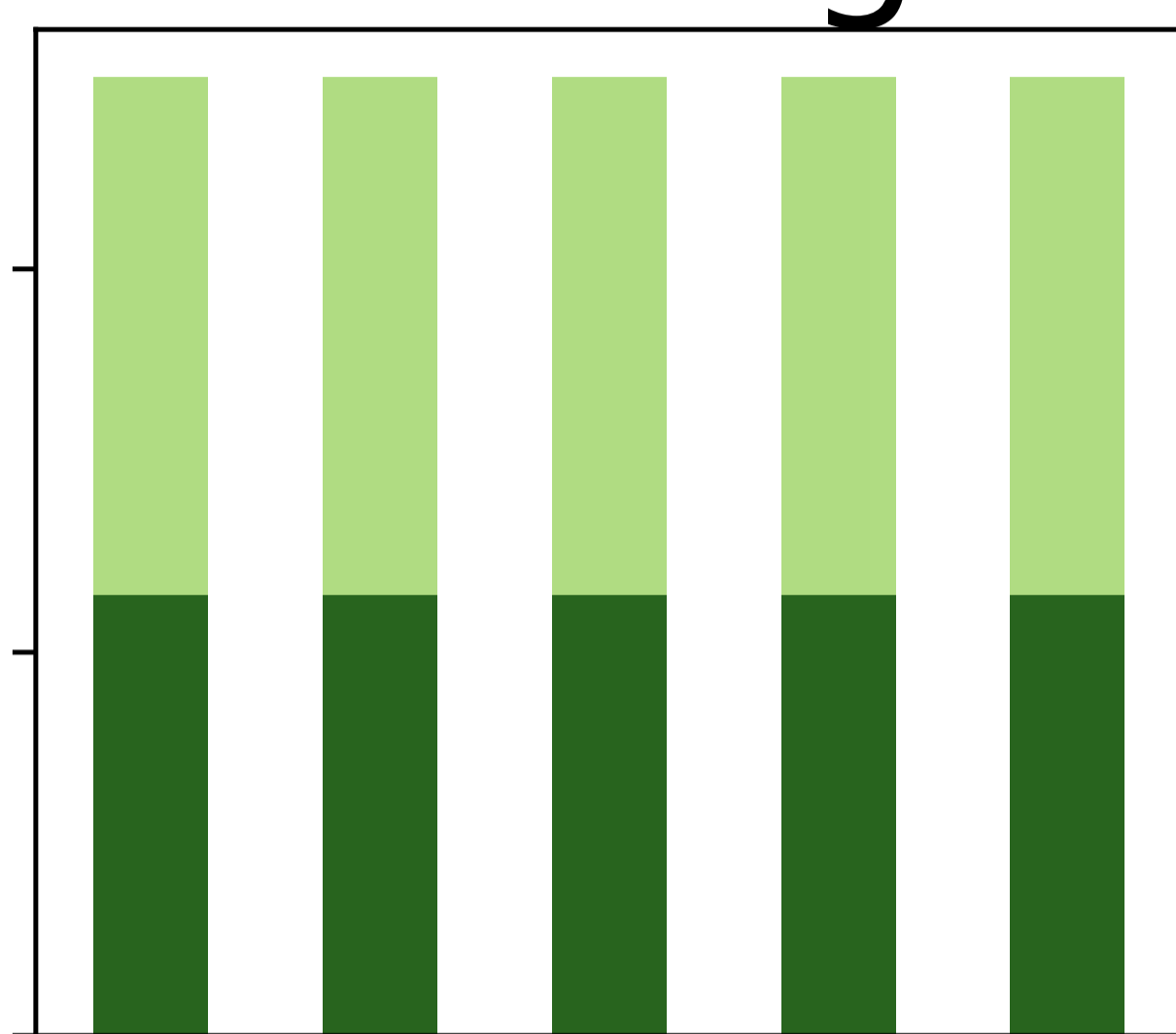
Utilitarian



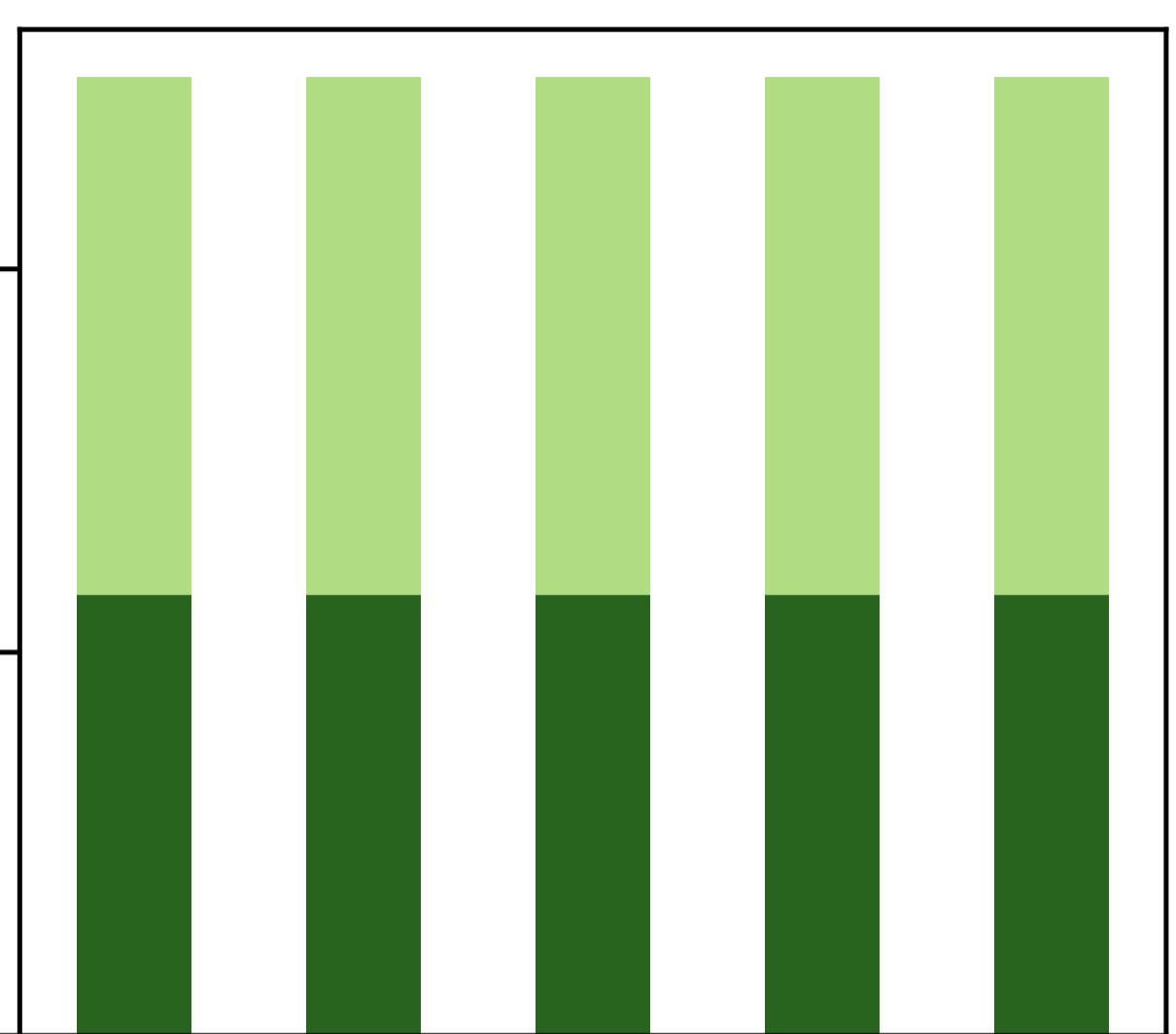
Game +
Deontological



Game, then
Deontological



Game, then
Utilitarian



Iterated Game