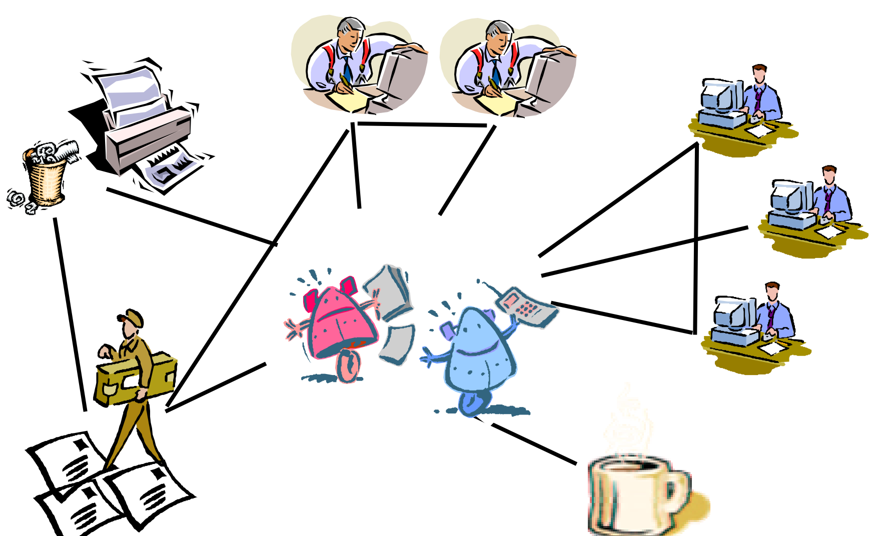
On the Risks and Rewards of Coordination in Multiagent Reinforcement Learning

Craig Boutilier
Department of Computer Science
University of Toronto
(joint work with Georgios Chalkiadakis)

A Multiagent Planning Problem



Learning & Coordination Problems

- Coordination of agent activities an important focus of multiagent learning (and RL)
  - (identical interest) stochastic games provide one useful model for studying such problems (multiagent? multi-agent? multi agent? MDPs)
- Known models:
  - Bayesian, FP, etc. models used to learn joint policies
  - in many cases, convergence to equilibrium assured
- Unknown models:
  - MARL techniques often used
  - convergence for some methods known, others seem to work reasonably well empirically

NIPS02, Multiagent Learning Workshop, Dec. 2002

3

The Curse of Multiple Equilibria

- One difficulty with “typical” MARL models
  - even if convergence to equilibrium assured, the equilibrium reached may be undesirable
  - influenced by structure of game

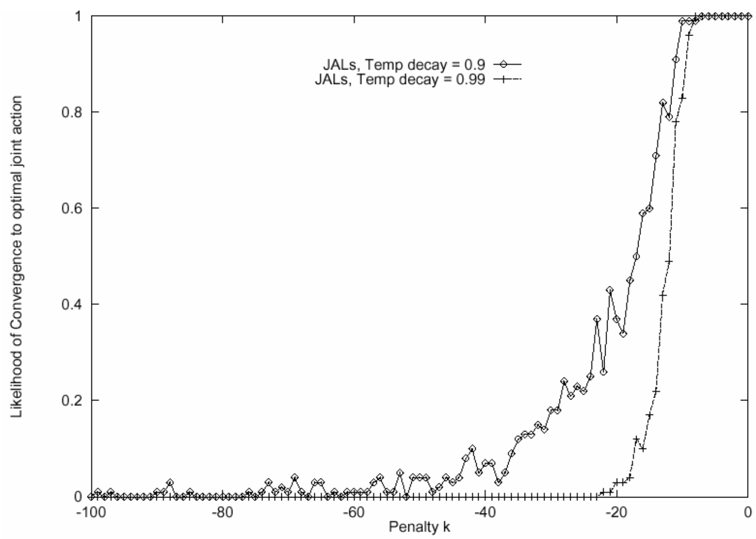
Penalty Game (Claus+Boutilier 97)

	a0	a1	a2
b0	10	0	k
b1	0	2	0
b2	k	0	10

NIPS02, Multiagent Learning Workshop, Dec. 2002

4

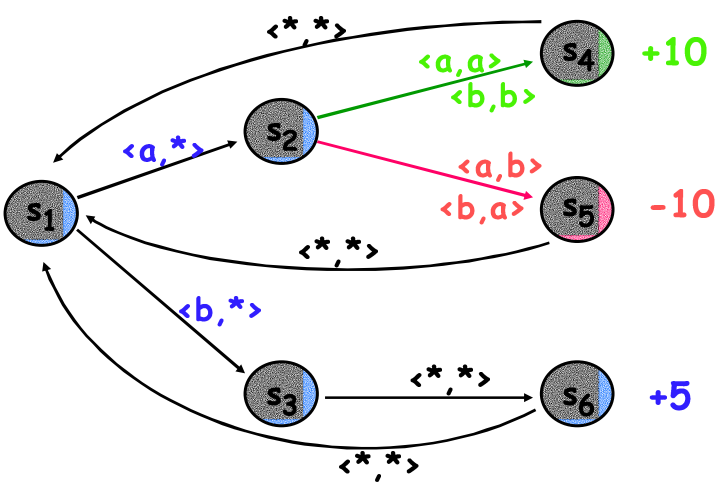
Convergence to Optimal Equil.?



NIPS02, Multiagent Learning Workshop, Dec. 2002

5

Opt In or Out Game



NIPS02, Multiagent Learning Workshop, Dec. 2002

6

Avoiding Suboptimal Equilibria

- A number of methods proposed to avoid convergence to suboptimal equilibria
  - CB96, LR00, KK02, WS02
  - generally, adopt an optimistic bias in exploration, ignoring the penalties or missed opportunities, in an effort to reach optimal equilibrium
  - e.g., [Penalty Game]: if player A persists in a0, this will eventually cause B to adopt b0 (optimal) using any standard RL algorithm
  - what price is paid? larger chance of accruing penalties before convergence…

"Optimal Exploration"

- Some heuristic methods may gaurantee convergence to optimal equil. (e.g., WS02)
- But what is the right performance metric?
  - common debate in (single agent) RL
- Tradeoff: is the price paid (penalties, lost opportunities) worth the gain offered by convergence to optimal (or better) equilibrium?
  - depends on discount factor, horizon, odds of converging to specific equilibrium, etc.
- Optimal exploration in MARL: *address explicitly!*

Bayesian Perspective on MARL

- Bayesian view of RL: optimal exploration easily formulated [Dearden et al., bandit problems]
- We have adopted this point of view for MARL
- However, several new components required
  - priors over models (incl. opponent strategies)
  - action selection as a POMDP
    - value of information (incl. what is learned about opponent strategies)
    - object level value (incl. how action choice impacts what opponent *will* do)

Basic Setup

- Assume a stochastic game
  - states S, fully observable
  - players $i \in \{1,\ldots,N\}$
  - action sets A_i , joint action set $A = X A_i$
  - dynamic Pr(s, **a**, t)
  - stochastic reward functions R_i
  - strategies σ_i , strategy profiles σ , σ_{-i}
- In MARL setting:
  - each agent experience has form <s, **a**, **r**, t>

Agent Belief State

- Each agent has *belief state* $b = \langle P_M, P_S, s, h \rangle$
  - P_M : density over space of possible models (games)
  - P_S : density over space of opponent(s) strategies
  - s : current state of the system
  - h : relevant history (i.e., that required to predict opponent moves given strategy beliefs)
- Update b' given experience tuple <s, **a**, **r**, t>
  - $P'_M(m) = \alpha \Pr(t, \mathbf{r} \mid \mathbf{a}, m)\, P_M(m)$
  - $P'_S(\sigma_{-i}) = \alpha \Pr(\mathbf{a}_{-i} \mid s, h, \sigma_{-i})\, P_S(\sigma_{-i})$
  - h' is suitable update of relevant history
  - combines Bayes RL and Bayes strategy learning

Simplifying Assumptions

- Factored local models $P_{(R)s}$ and $P_{(D)s,a}$
  - assume local densities are Dirichlet, independent
  - allows very simple updating of P_M
- Some convenient prior P_S
  - we use simple fictitious play-style beliefs (no history)
  - locally factored, independent at each state
  - more general models feasible
  - interesting question: what are reasonable, feasible classes of opponent models

Tradeoffs in Optimal Exploration

- Given belief state *b*, each action a_i :
  - has expected object level value
  - provides info. which can subsequently be exploited
- Object level value:
  - immediate reward
  - predicted state transition (expected value)
  - impact on future opponent action selection
- Value of Information:
  - what you learn about transition model, reward
  - what you could learn about opponent strategy
  - how this info impacts future decisions

POMDP Formulation

- Tradeoff can be made implicitly by considering long-term impact of action on belief state and associating value with belief states

$$Q(a_i, b) = \sum_{a_{-i}} \Pr(a_{-i} \mid b) \sum_t \Pr(t \mid a_i \circ a_{-i}, b) \sum_{\boldsymbol{r}} \Pr(\boldsymbol{r} \mid a_i \circ a_{-i}, b)\,[r + \gamma V(b(s, \boldsymbol{a}, \boldsymbol{r}, t))]$$

$$V(b) = \max_{a_i} Q(a_i, b)$$

Computational Apprxomations

- Solving belief state MDP intractable
- Myopic model (one step lookahead)
  - account for impact of action on next belief state
  - execute action with maximum myopic Q-value

$$V_m(b) = \max_{a_i} \int_m \int_{\sigma_{-i}} Q(a_i, s \mid m, \sigma_{-i}) P_M(m) P_S(\sigma_{-i})$$

- Sampling techniques used to evaluate integrals
  - sample games from P_M, solve corresponding MDP
  - can sample strategies (or use expectation if simple)
  - other tricks can be used (importance sampling, repair, sampling belief states, exploit repeated games, etc.)

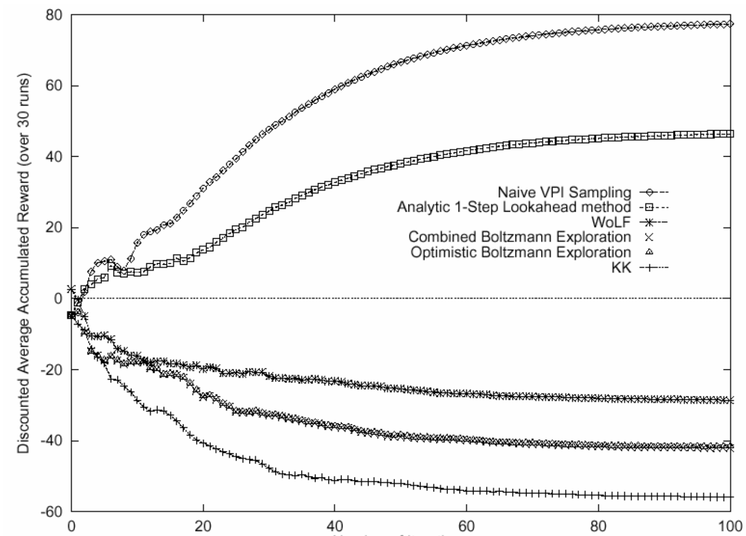
Computational Approximations

- Other approaches include using the rather different *Q-value sampling* approach to estimating EVOI [Dearden et al.]
  - see paper for details
  - approximates in very different way by sampling models, computing optimal Q-values, and determining whether these values are sufficient to change the optimal action choice at current state

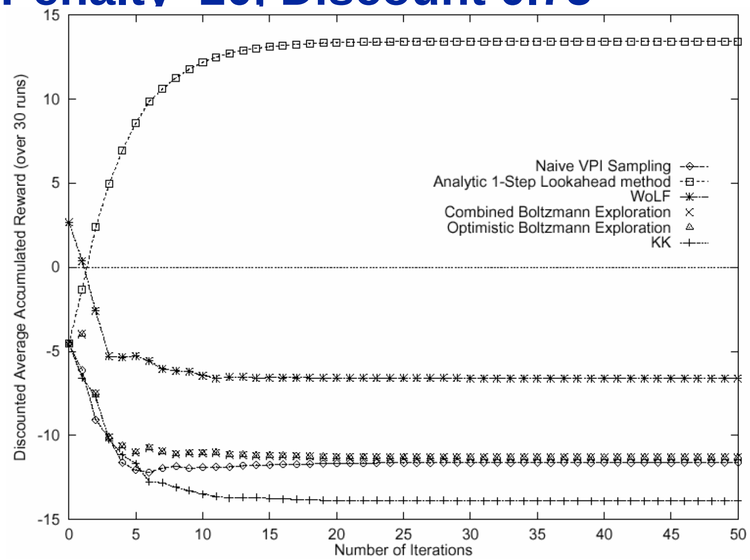
Empirical Results

- Tested the Bayesian approach using both:
  - one-step lookahead (BOL)
  - naïve VPI sampling (BVPI)
- Compared—on several repeated games and stochastic games—to several algorithms:
  - KK (Kapetanakis and Kudenko, AAAI-02)
  - OB, COB (Claus and Boutilier, AAAI-98)
  - WoLF-PHC (Bowling and Veloso, IJCAI-01)
    - much more general algorithm
- Compare using total discounted reward accrued

Penalty -20, Discount 0.95

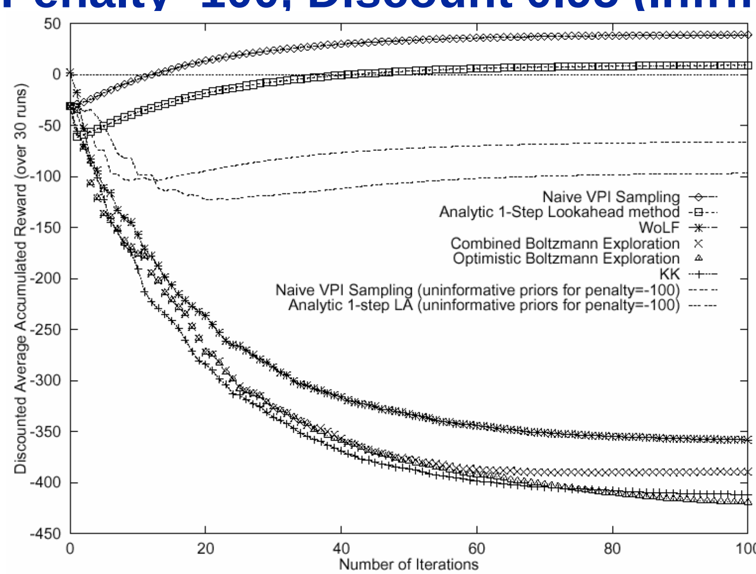


Penalty -20, Discount 0.75



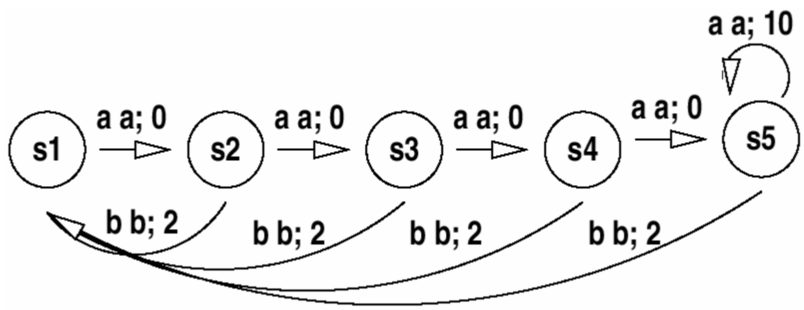
NIPS02, Multiagent Learning Workshop, Dec. 2002

Penalty -100, Discount 0.95 (Infrmd)

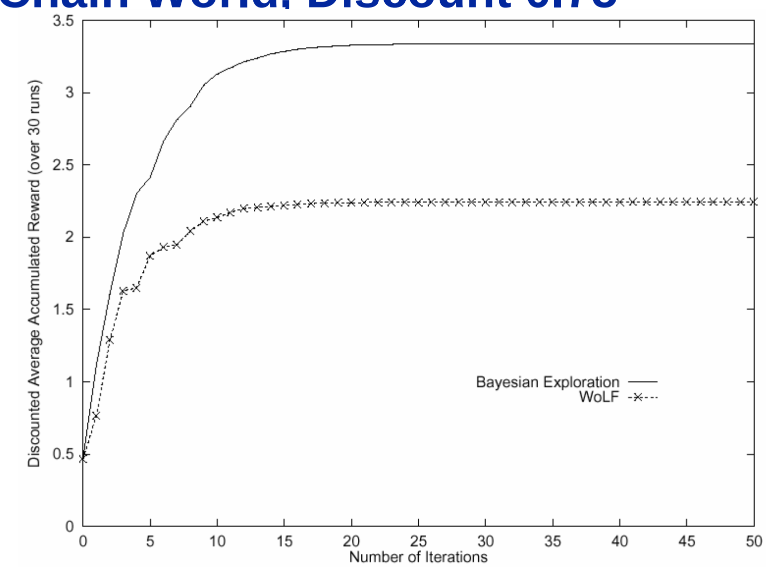


NIPS02, Multiagent Learning Workshop, Dec. 2002

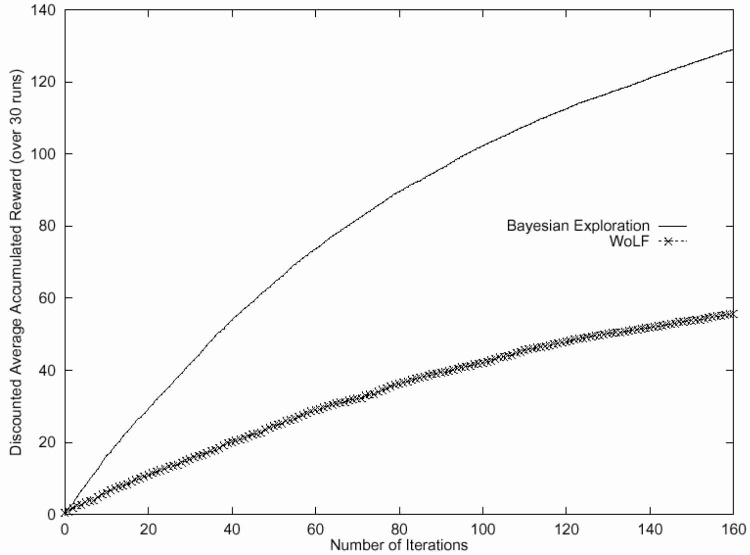
Chain World



Chain World, Discount 0.75



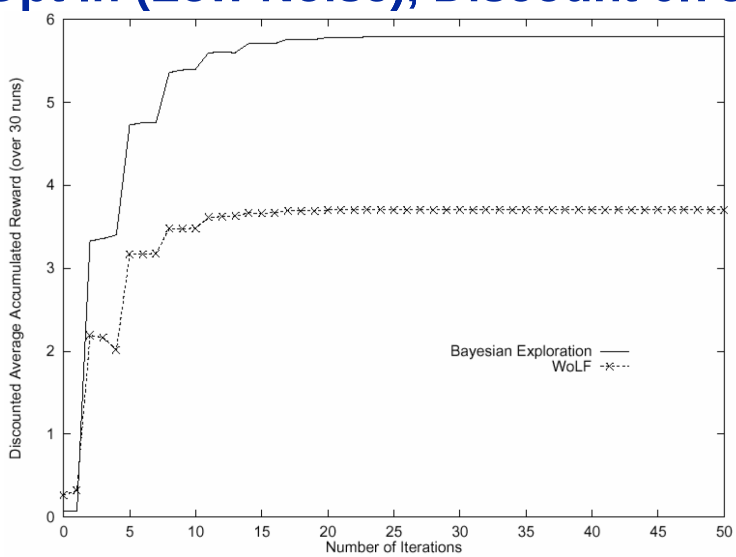
Chain World, Discount 0.95



NIPS02, Multiagent Learning Workshop, Dec. 2002

23

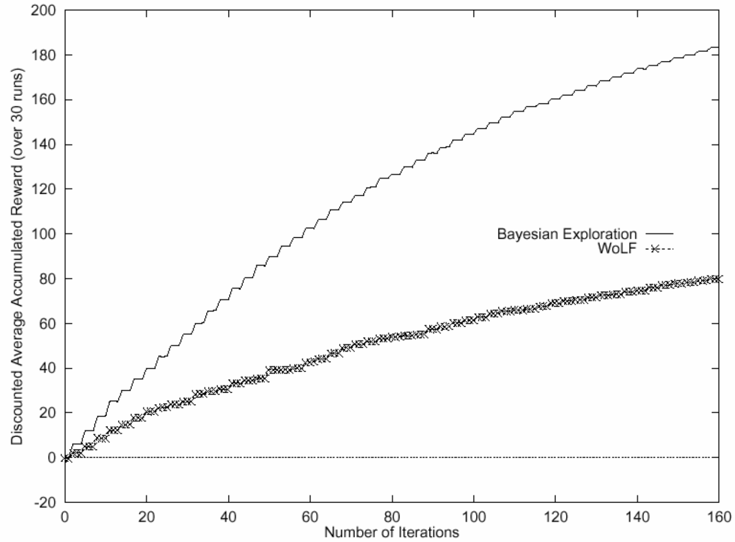
Opt In (Low Noise), Discount 0.75



NIPS02, Multiagent Learning Workshop, Dec. 2002

24

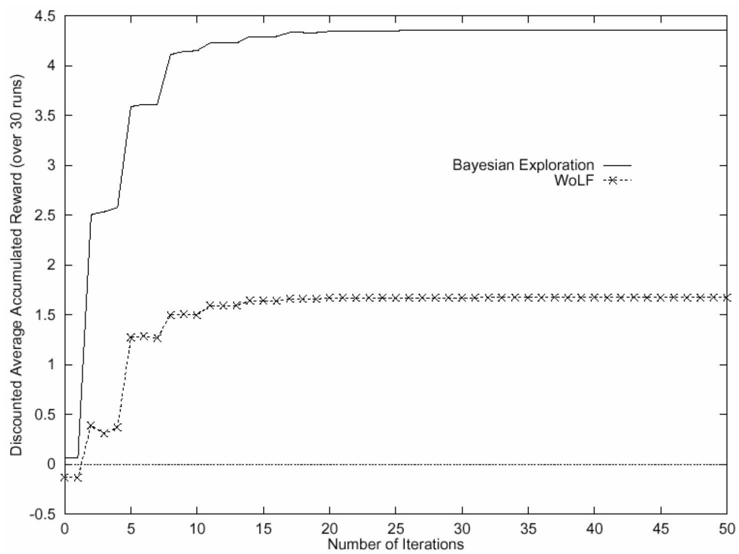
Opt In (Low Noise), Discount 0.99



NIPS02, Multiagent Learning Workshop, Dec. 2002

25

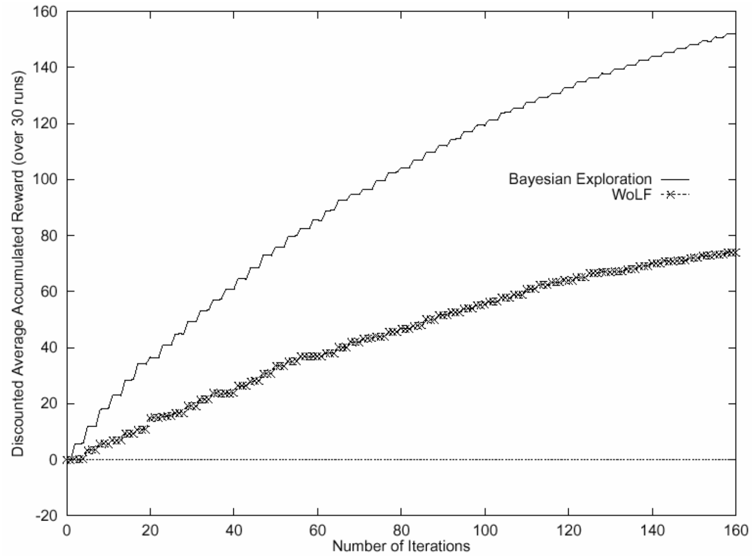
Opt In (Med. Noise), Discount 0.75



NIPS02, Multiagent Learning Workshop, Dec. 2002

26

Opt In (Med. Noise), Discount 0.99



Summary

- Bayesian method seems to perform well compared to other methods tested
  - algorithms designed to “force” convergence to optimal equilibrium pay a very large price
  - WoLF (which doesn’t force optimality) fares much better than algorithms designed for these problems!
- In sequential games:
  - BVPI shows better online performance than WoLF
  - CW: BVPI converges to optimal policy, WoLF doesn’t
  - OI (low): BVPI and WoLF converge to optimal policy sometimes sometimes not

Conclusions

- More thought needed on the objectives of MARL
- Bayesian technique explicitly addresses the tradeoff between exploration and exploitation
  - including "joint" exploration and exploitation
- Generally, performs better than other approaches wrt discounted reward
  - may sacrifice convergence to optimal (stochastically) if the cost outweighs the gain
  - but often does converge to optimal
  - very flexible model
    - *priors*; opponent models; discount/horizon; etc.