# Within Epsilon of Optimal Play in the Cultaptation Social Learning Game

# (Extended Abstract)

Ryan Carr, Eric Raboin, Austin Parker, Dana Nau
University of Maryland
College Park, MD 20742
{carr2,eraboin,austinjp,nau}@cs.umd.edu

*Social learning*, in which members of a society learn by observing the behavior of others, is an important foundation for human culture, and is observed in many other species as well. It seems natural to assume that social learning evolved due to the inherent superiority of copying others' success rather than learning on one's own via trial-and-error innovation. However, there has also been substantial work questioning this intuition [3, 5, 1, 6, 4]. For example, blindly copying information from others is not useful if the information is wrong—or if it once was right but has since become outdated. Under what conditions does social learning outperform trial-and-error learning, and what kinds of social-learning strategies are likely to perform well?

One attempt to gain insight into these questions is an evolutionary simulation called The Social Learning Strategies Tournament [2],[1] which was created in order to study the conditions under which communication outperforms trial-and-error and vice-versa. More than 100 researchers worldwide have entered strategies in the tournament, vying for a €10,000 prize. To date, the tournament's organizers have not yet finished evaluating the strategies.

Moves in the social learning game are highly simplified analogs of the following real-world activities: spending time and resources to learn something new, learning something from another player, and exploiting learned knowledge. By developing a formal way of analyzing this set of activities, we hope it will allow us to perform case studies, and to identify how different patterns of behavior fare in different environments.

This extended abstract summarizes several contributions to knowledge about the social learning game:

1. We have derived a formula for approximating (to within any $\epsilon > 0$) the expected utility of a strategy in the social learning game.
2. We have produced an algorithm that incorporates a lookahead search to find near-optimal strategies.
3. We have shown that locally optimal moves are not necessarily optimal in the long term, but one can derive an upper bound on how many levels of lookahead are needed to find a globally optimal move.
4. We have given proofs of correctness and big-$O$ runtime analyses for our algorithms.

[1]NOTE: None of us is affiliated with the tournament in any way.

**Cite as:** Within Epsilon of Optimal Play in the Cultaptation Social Learning Game, (Extended Abstract), Ryan Carr, Eric Raboin, Austin Parker, Dana Nau, *Proc. of 8th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2009),* Decker, Sichman, Sierra and Castelfranchi (eds.), May, 10–15, 2009, Budapest, Hungary, pp. 1327–1328


## 1. THE SOCIAL LEARNING GAME

The social learning game is introduced in [2]. The game is an $n$-player round based game, where one move is made by each agent per round. The rules can be summarized as follows:

- At each step of the game, each player must choose one of three moves: *innovate*, *observe*, or *exploit*. How an agent does this constitutes the agent's "social learning strategy."
- Each *exploit* move provides the player an immediate numeric payoff, while *innovate* and *observe* moves help the player learn actions to exploit.
- Each agent has a fixed probability of dying at each round, to be replaced by the offspring of another agent. Agents with a higher average payoff are more likely to reproduce.

A player may only $exploit$ an action that has been learned through a previous $innovate$ or $observe$ move. Each exploitable action has an assigned utility drawn from a probability distribution function, which may be arbitrary, and is a parameter for the game. This utility changes on round $j$ with probability $c(j)$, called the *probability of change*. An action's payoff, and/or the accuracy of the information learned, may vary as the game goes on.[2]

An $observe$ move informs the agent of an exploit action made by another agent during the previous round, as well as the action's utility. In contrast, an $innovate$ move informs the agent of a *random* exploitable action and its utility. Innovation and observation are essential precursors to exploitation: without first learning an action, an agent is not able to gain any utility. Additionally, no agent knows of any other agents' behavior during the game except through observation.

Formally, all of the information each agent receives on each round can be described by a triple $\langle mv, act, v\rangle$, where $mv$ is whatever move the agent chose to perform on that round ($innovate$, $observe$, or $exploit$), $act$ is an exploitable action or a null value ($X[1], \ldots, X[m], \emptyset$), and $v$ is the utility value observed or received. A *history* is a sequence of such tuples.

[2]The Cultaptation tournament directors have deliberately avoided giving any information about how this variation may occur, as well as information about several other important game characteristics. Hence, in order to model these characteristics in as general a way as possible, we treat them as arbitrary parameters that may have different values at each time step.

Each round, each agent dies with probability $d$. Dead agents are removed from the game and replaced by the offspring of a living agent, chosen at random. Offspring use the same strategy as their parents, but do not inherit any knowledge of actions or utilities their parents may have. An agent's chance of being selected to reproduce is proportional to its *per round utility*, the utility it has earned (through exploit actions) divided by the number of rounds it has been alive.

The winning strategy is the one that is used by the most agents during the last quarter of the game. Hence, the goal of a strategy should be to maximize the number of times agents using it are selected to reproduce.

**Expected Per Round Utility**

In the paper, we develop a metric for evaluating any given strategy $S$, called the *expected per round utility* of $S$ or $EPU(S)$. We then prove that, if a strategy $S_{opt}$ exists such that, for any other strategy $S'$, $EPU(S_{opt}) \geq EPU(S')$, then $S_{opt}$ has the best possible chance of winning the game, and therefore it is the optimal strategy.

Let $H$ be the set of all possible histories. To calculate $EPU(S)$ we take the sum, for each history $h$ in $H$, of the probability that $S$ causes $h$ to occur, times the per round utility of $h$ (denoted $PRU(h)$:

$$EPU(S) = \sum_{h \in H} \underbrace{(1-d)^{|h|-1}}_{\text{Prob. of living |h| rounds}} \times \underbrace{P(h|S)}_{\text{Prob. } S \text{ causing } h} \times \underbrace{PRU(h)}_{\text{Per-round utility}} .$$

**Approximating EPU**

If we do not know when the game ends, then histories of any length are possible, so the set $H$ has infinite size and we cannot compute $EPU(S)$. However, note that the first term, $(1-d)^{|h|-1}$, decreases exponentially[3] as $|h|$ increases. This means that histories representing later rounds in the game contribute much less to the total $EPU$ than histories representing early rounds. In fact, if we let $EPU_k(S)$ be the portion of $EPU(S)$ contributed by histories of length $\leq k$, then we can show that for any $\epsilon > 0$, there exists a $k$ such that $EPU_k(S)$ is within $\epsilon$ of $EPU(S)$. In the paper, we prove that $k \leq \log_{(1-d)}(d\epsilon/v_{max})$, where $v_{max}$ is the value of the maximal-utility move.

**Finding $\epsilon$-Optimal Strategies**

In the paper, we present a simple search algorithm that, given the probabilities of innovating and observing each action and some search depth $k$, will find a strategy $S'$ that maximizes $EPU_k(S')$. If we choose $k$ so that $EPU_k(S)$ is within $\epsilon$ of $EPU(S)$, and if we let $S_{opt}$ be the strategy with the highest possible $EPU$, then $EPU(S')$ is within $\epsilon$ of $EPU(S_{opt})$, and $S'$ is $\epsilon$-optimal.

## 2. CONCLUSION

We have developed an algorithm for synthesizing near-optimal strategies in the social learning game.

To decide what move a strategy $S$ should make at each point in the game, our algorithm does a lookahead search to estimate each move's expected utility. The accuracy of this estimate relies on the fact that since the agent has a nonzero probability of death at each round, moves farther into the future have diminishing effects on the expected utility. The algorithm looks far enough ahead that that all further moves will change the expected utility by less than $\epsilon$. We have proved that this occurs within a lookahead depth of $\log_{(1-d)}(d\epsilon/v_{max})$, where $d$ is the probability of dying on each round and $v_{max}$ is the value of the maximal-utility move.

[3]Recall that $d$ is the probability that the agent dies on each round, so $(1-d)$ is always between 0 and 1

One limitation of our work is the algorithm's exponential running time—but we are confident that pruning techniques and approximation techniques can be developed to make the algorithm run much more quickly. Once the algorithm has been speeded up, this should make it possible to use the algorithm to analyze different parameter settings for the social learning game, to see which kinds of moves are optimal under what kinds of conditions. When is it, for instance, that innovation is always preferable to observations and vice-versa? Such investigations are left for future work.

Also left for future work is the examination of information gathering in the social learning game. One of our algorithm's inputs is the probability distributions from which the action utilities are drawn. We have kept the algorithm fully general by allowing these distributions to change from one time step to the next—but what the distributions are, and how/whether they change at each time step, is information that the game's authors have deliberately not revealed. Without access to such information, a game agent must either approximate the distributions or develop an algorithm that can do well without them. If we choose to approximate, should our agent be willing to sacrifice some utility early on, in order to gain information that will improve its approximation? Are there strategies that perform well in a wide variety of environments, that we could use until our agent develops a good approximation? Are some of these strategies so versatile that we can simply use them without needing to know the distributions? These remain open questions.

In conclusion, we hope that our results on the Cultaptation social learning game will help provide insight into the utility of inter-agent communication in evolutionary environments.

## Acknowledgments

This work was supported in part by AFOSR grant FA95500610405, NAVAIR contract N6133906C0149, DARPA's Transfer Learning and Integrated Learning programs, and NSF grant IIS0412812. The opinions in this paper are those of the authors and do not necessarily reflect the opinions of the funders.

## 3. REFERENCES

[1] C. Barnard and R. M. Sibly. Producers and scroungers: A general model and its application to captive flocks of house sparrows. *Animal Behavior*, 29:543–550, 1981.

[2] R. Boyd, M. Enquist, K. Eriksson, M. Feldman, and K. Laland. Cultaptation: Social learning tournament, 2008. `http://www.intercult.su.se/cultaptation`.

[3] R. Boyd and P. Richerson. Why does culture increase human adaptability? *Ethology and Sociobiology*, 16(2):125–143, 1995.

[4] L. A. Giraldeau, T. J. Valone, and J. J. Templeton. Potential disadvantages of using socially acquired information. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 357(1427):1559–1566, 2002.

[5] K. Laland. Social learning strategies. *Learning and Behavior*, 32:4–14, 2004.

[6] D. Nettle. Language: Costs and benefits of a specialised system for social information transmission. In J. Wells and et al., editors, *Social Information Transmission and Human Biology*, pages 137–152. Taylor and Francis, London, 2006.