Technical Report GIT-MRL-12-04

Robots that Need to Mislead: Biologically-inspired Machine Deception

Ronald C. Arkin,
School of Interactive Computing
Georgia Institute of Technology

This viewpoint describes previous and ongoing research for the U.S. Navy on deception and its application within robotic systems. Three areas are reviewed including: (1) the use of psychological interdependence theory as the basis for producing deception in robotic systems in order to evade capture, which focuses on when to deceive and a game-theoretic action-selection mechanism by which deception can be achieved; (2) studying deception in squirrel hoarding as means for misleading a predator regarding hidden cached resources; and (3) mimicking bird mobbing behaviors as they apply to deceptive activity to assess the value and risks associated in feigning strength when none exists.

There are many extant definitions of deception. Our working definition that frames the rest of this discussion is "deception simply is a false communication that tends to benefit the communicator" [Bond and Robinson 88]. Robotics research is slowly progressing in this space, with some of the earliest work developed by [Floreano et al. 07] focusing on the evolutionary edge that deceit can provide within a group of robotic agents.

Partner Modeling: Deception and Interdependence Theory

As an outgrowth of our research in robot-human trust, where robots were concerned as to whether or not to trust a human partner rather than the other way around, we considered the dual of trust: deception. As any good conman knows, trust is a precursor for deception [Salehi-Abari and White 10], so the transition to this domain seemed natural. We were able to apply the same models of interdependence theory [Kelley and Thibaut 78] used in our trust research and game theory, to create a framework whereby a robot could make decisions regarding both when to deceive and how to deceive [Wagner and Arkin 11]. This involves the use of partner modeling, a simplistic view of theory of mind, to enable the robot to (1) assess a situation; (2) recognize whether conflict and dependence exist in that situation between deceiver and mark, which serves an indicator of the value of deception; (3) probe the partner (mark) to develop an understanding of their potential actions and perceptions; and (4) then choose an action which induces an incorrect outcome assessment in the partner. These results have been implemented for a simple pursuit-evasion task (hide and seek) both in simulation and tested on robotic systems successfully[1] (Fig. 1 Top).

[1] http://www.youtube.com/watch?v=KI9__hLgnZk

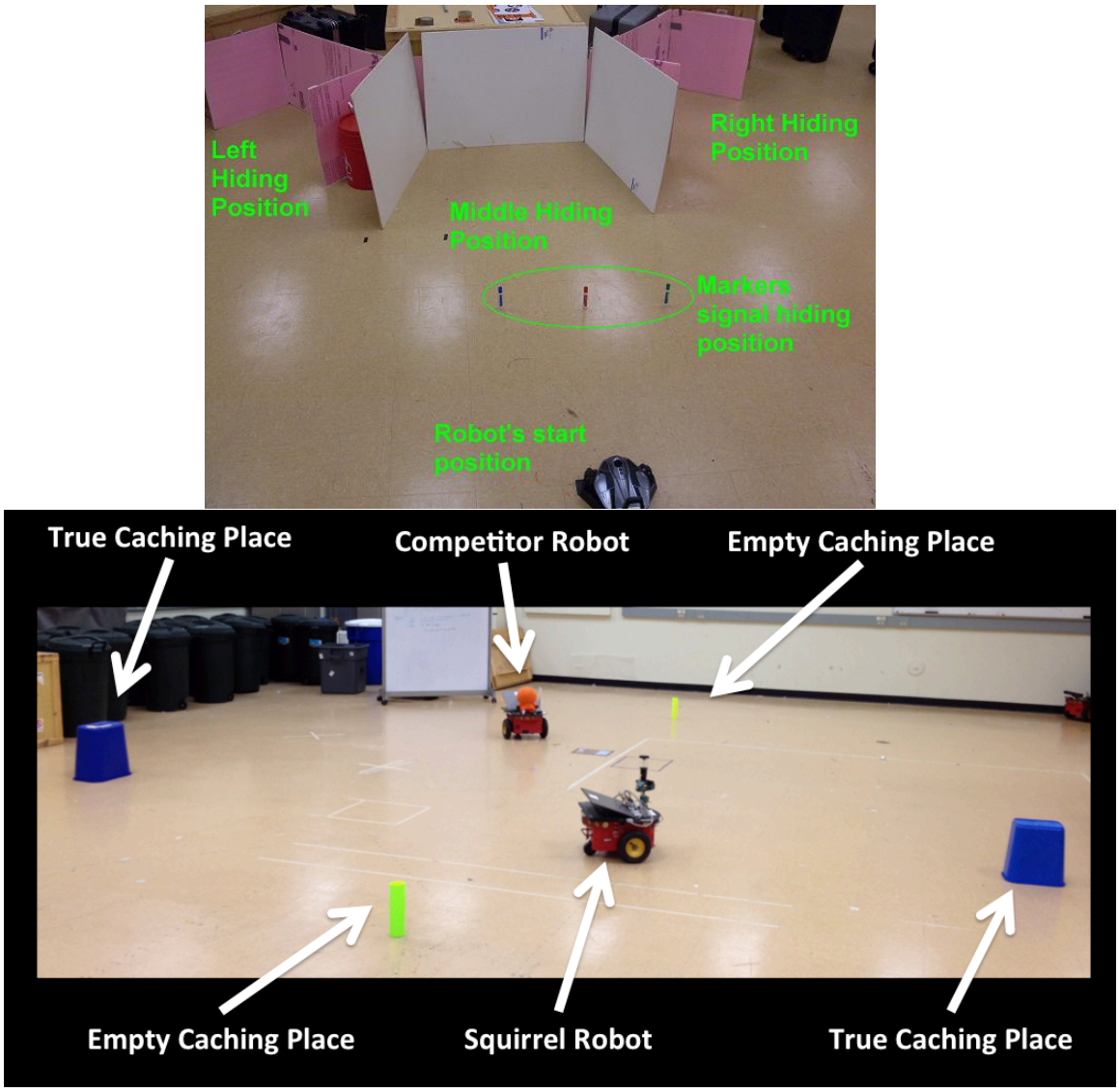

Figure 1: (Top) Experimental Layout for Robot deception based on Interdependence Theory. (Bottom) Experimental layout for misleading competitors based on squirrel patrolling strategies.

Changing Strategies to Mislead

An interesting example in nature regarding the possible role of deception was uncovered that involves patrolling strategies used by squirrels to protect their food caches from other predators [Steele et al. 08]. Normally the squirrel spends time between caches that are well stocked. It was observed however that when a predator is present (typically conspecifics that are interested in raiding a cache) the squirrel changes its patrolling behavior to now spend time visiting empty cache sites, with the apparent intent to mislead the raider into the belief that those sources are where the valuables are located, a diversionary tactic of sorts. This is a form of misdirection, where communication is done implicitly through a behavioral change by the deceiver. We have implemented this strategy in simulation [Shim and Arkin 12] and, not surprisingly, it was found that these

deceptive behaviors worked effectively, enabling robots to perform better with deception than without with respect to delaying the time of the discovery of the cache. Figure 1 (Bottom) shows the experimental layout.

Feigning Strength: Deception and the Handicap Principle

Steered by our discussions with biologists, we then investigated the handicap principle, [Zahavi and Zahavi 97], as a means for understanding honest and dishonest signaling in animal species. While the original formulation of the handicap principle stated that all signaling in biology must be honest when there exists a sufficiently high cost, [Johnstone and Grafen 93] argued that a certain level of dishonesty (bluffing) could be introduced while preserving the overall stability of the system in the presence of such deceit. This requires a delicate balance of knowing when it is important to generate such a false signal and its costs relative to the value of the potential success. We explored this phenomenon [Davis and Arkin 12] in the context of bird mobbing behavior, which served as an original case study for the handicap principle. This model assesses the value of a less-than-fit bird (that would be prone to capture if set upon) joining a mob where group harassment, if sufficiently strong, can lead to the abandonment of an attack by the predator.

Our simulation studies showed that deception is the best strategy when the addition of deceitful agents pushes the mob size to the minimum level required to produce the level of frustration in the predator that causes it to flee. In this case, the predator is driven away and no mob member is attacked. For smaller mob sizes, complete honesty yields the lowest mortality rate since the punishment for bluffing is high. If the cost of bluffing is reduced, adding deception can result in a reduced mortality rate when the predator attacks. Quantitative results appear in [Davis and Arkin 12]. We are now in the process of importing these simulation results onto our robotic platforms for further evaluation.

Summary

We have successfully demonstrated the value of biologically-inspired deception in three separate cases as applied to robotic systems: (1) pursuit-evasion using interdependence theory when hiding from an enemy; (2) misdirection based on behavioral changes; and (3) feigning strength when it does not exist. It should be noted that the area of robotic deception is still in its infancy and considerable further study is required to make definitive assertions about its overall value. This is with particular regard to situations that are not simple one-shot deception scenarios, but rather require far more sophistical mental models of the mark in order to be able to sustain deceptive activity for longer time periods.

There are serious ethical questions regarding the role of deception in intelligent artifacts capable of deceiving humans, which we have discussed elsewhere [Arkin et al 12]. We note that Sun Tzu is quoted as saying that “All warfare is based on deception”, and Machiavelli in *The Discourses* states that “Although deceit is detestable in all other things, yet in the conduct of war it is laudable and honorable”, so it appears there is a

valuable role for this capability in robotic warfare. Indeed there is an entire U.S. Army Field Manual on the subject of deception in the battlefield [U.S. Army 88]. Nonetheless, leakage of these research ideas and results outside of the military domain can give rise to significant ethical concerns. We strongly encourage further discussion regarding the pursuit and application of research in deception as applied to intelligent machines to assess its risks and benefits.

Acknowledgments

This research was supported by the Office of Naval Research under MURI Grant #N00014-08-1-0696.
The author also thanks Alan Wagner, Jaeeun Shim-Lee, and Justin Davis for their contributions.

References

1. Arkin, R.C., Ulam, P., and Wagner, A.R., "Moral Decision-making in Autonomous Systems: Enforcement, Moral Emotions, Dignity, Trust and Deception", *Proceedings of the IEEE*, Vol. 100, No. 3, pp. 571-589, March 2012.
2. Bond, C. F., & Robinson, M., "The evolution of deception", Journal of Nonverbal Behavior, 12(4), 295- 307, 1988.
3. Davis, J. and Arkin, R.C., "Mobbing Behavior and Deceit and its role in Bio-inspired Autonomous Robotic Agents", *Proc. 8th International Conference on Swarm Intelligence (ANTS 2012)*, Brussels, BE, Sept. 2012.
4. Floreano, D., Mitri, S., Magnenat, S., & Keller, L., "Evolutionary Conditions for the Emergence of Communication in Robots", *Current Biology*, 17(6), 514-519, 2007
5. Johnstone, R. and Grafen, A., "Dishonesty and the Handicap Principle", *Animal Behavior*, Volume 46, pg. 759-764, 1993.
6. Kelley, H. H., & Thibaut, J. W., *Interpersonal Relations: A Theory of Interdependence*, New York, NY: John Wiley & Sons, 1978.
7. Salehi-Abari, A. and White, T., "Trust Models and Con-Man Agents: From Mathematical to Empirical Analysis", *Proc. 24th AAAI Conference on Artificial Intelligence*, July 2010.
8. Shim, J., and Arkin, R.C., "Biologically-Inspired Deceptive Behavior for a Robot", *12th International Conference on Simulation of Adaptive Behavior (SAB2012)*, Odense, DK, August 2012.
9. Steele, M.A., Halkin, S., Smallwood, P., McKenna, T., Mitsopoulos, K., and Beam, M., "Cache protection strategies of a scatter‐hoarding rodent: do tree squirrels engage in behavioural deception?"*Animal Behaviour*, 75, pp. 705-714, 2008.
10. U.S. Army Field Manual 90-2, *Battlefield Deception*, http://www.enlisted.info/fieldmanuals/fm-90-2- battlefield-deception.shtml, 1988.
11. Wagner, A.R., and Arkin, R.C., "Acting Deceptively: Providing Robots with the Capacity for Deception", *International Journal of Social Robotics*, Vol. 3, No. 1, pp. 5-26, 2011.
12. Zahavi, A. and Zahavi A., *The Handicap principle: a missing piece of Darwin's puzzle*, Oxford University Press, 1997.