Misdirection in Robot Teams:: Methods and Ethical Considerations

Ronald C. Arkin
Georgia Institute of Technology
Atlanta, GA USA

EXTENDED ABSTRACT

Trust, dependability, cohesion, and capability are integral to an effective team. These attributes are the same for teams of robots. When multiple teams with competing incentives are tasked, a strategy, if available, may be to weaken, influence or sway the attributes of other teams and limit their understanding of their full range of options. Such strategies are widely found in nature and in sporting contests such as feints, misdirection, etc. This talk focuses on one class of higher-level strategies for multi-robots, i.e., to intentionally misdirect using shills or confederates as needed, and the ethical considerations associated with deploying such teams. As multi-robot systems become more autonomous, distributed, networked, numerous, and with more capability to make critical decisions, the prospect for intentional and unintentional misdirection must be anticipated.

This NSF funded project currently underway studies strategies to enable robots, multi-robots and teams of multi-robots to model, generate, and cope with misdirection in various situations. This research direction in robotic control offers a novel approach to resilience in and among these teams to these forms of possible disruption. Computational models, drawn particularly from studies of human endeavors and group behaviors, provide a general framework for understanding, producing, and countering misdirection in robotic systems. A framework of computational models will be designed using recursive schema-theoretic models of behaviors at the individual and team levels, building on decentralized methods of control and communication.

While benefits are clearly apparent to the team performing the deception, ethical questions surrounding the use of misdirection or other forms of deception are quite real and we have published earlier on this topic [1,2]. The IEEE Global Initiative on the Ethics of Autonomous and Intelligent Systems has produced a preliminary set of recommendations on agent deception among other topics. The author served as co-chair of the Affective Computing Committee for this initiative that was responsible for these guidelines regarding deception, ensuring that consideration of the ethical aspects of this work and its social implications will reach a large audience, and they will be discussed here.

The relationship of misdirection and its relationship to intelligence is well documented. Indeed, the Turing test, a hallmark measure of artificial intelligence, is based on confusing a human with a computer. Deception is believed to play a significant role in Human-Human interaction, and thus has a place in Human-Robot Interaction: *"The development of deception follows the development of other skills used in social understanding" [3].* The philosopher Dan Dennett stated: *"another price you pay for higher-order intentionality is the opportunity [for] ... deception "* [4].

Our research group has conducted prior work on robot deception for individual robots, including using these agents to feign strength where there is none [5], feint [6] or mislead [7,8], and provide

support for those in distress [9,10] among others. We have experience drawing on interdisciplinary models, for example psychological interdependence theory, small animal misdirection, a dishonesty model drawn from biology and criminology. This research has led to the development of the first taxonomy of human-robot deceptive activities, including misdirection [11]. Here we consider *team* misdirection, from organizational models drawn from sports, the military, biology and other relevant disciplines. As stated earlier ethical considerations to date have played an important role [1] and continue to do so in our ongoing research.

Robot teams that use misdirection provide the ability to confuse, to obscure, and execute other novel behaviors that no single agent could provide. Heterogeneity arises from the presence not only of a deceiving agent, but also shills, which can support misdirection indirectly. Bluffing where a team's strength lies, group distraction, feigning group movements, emulating many robots with a small number, false displays, feints, or demonstrations, deceptive logistical movements, etc., are examples for group deception drawn from military operations. The goal of such activity may include: inducing in the mark a misperception of intent; masking the movement of the overall team; or a miscalculation of the numbers of the team, dispositions and intentions. For coordinated activity against an opposing organization (e.g., sports teams) multiple agents are required, and heterogeneous robot teams can be tasked for these purposes. To our knowledge, limited, if any, research has been conducted on coordinated robot team misdirection, especially when using robots of differing capabilities, let alone the ethical aspects of group deception.

Other researchers have studied deception in a robotics context, but few have considered the ethical consequences. Floreano demonstrated robots evolving deceptive strategies in an evolutionary manner, learning to protect energy sources; Terada demonstrated that a robot was able to deceive a human by producing a deceptive behavior contrary to the human subject's expectations; Work at Yale illustrated increased engagement with a cheating robot in the context of a rock-paper-scissors game; Research at CMU showed an increase of users' engagement and enjoyment in a multi-player robotic game in the presence of a deceptive robot referee; University of Tsukuba showed a deceptive robot assistant can improve the learning efficiency of children; Brewer et al. shows that deception can be used in a robotic physical therapy system; A robot sheepdog, can be categorized as robot deception, since the robot aims to deceive sheep so that it can control the sheep flock automatically; Generating deceptive robot motion to convey intentionality in a robot to a human observer at CMU was performed; A form of deception for adversarial patrols where the team is capable of recognizing that the adversary can only see part of the team and a form of heterogeneous teaming called scarecrow deception where individuals appear to have sensing ability but do not. There are other examples of course, but these are representative examples.

References [abridged due to space limitations]

[1] Arkin, R.C., "Ethics of Robot Deception", *IEEE Technology and Society Magazine*, Sept. 2018.

[2] Arkin, R.C., Ulam, P., and Wagner, A.R., "Moral Decision-making in Autonomous Systems: Enforcement, Moral Emotions, Dignity, Trust and Deception", *Proc. IEEE*, 100(3):571-589, March 2012.

[3] Vasek, M.E., "Lying: The development of children's understanding of deception," Clark University, 1984.

[4] Dennett, D. C., "When Hal kills, who's to blame? computer ethics," in *HAL's Legacy: 2001's Computer as Dream and Reality*, MIT Press, 1997.

[5] 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*, pp:276-283, Sept. 2012.

[6] Shim, J., and Arkin, R.C., "Biologically-Inspired Deceptive Behavior for a Robot", *12th International Conference on Simulation of Adaptive Behavior*, pp. 401-411, August 2012.

[7] Wagner, A.R., and Arkin, R.C., "Acting Deceptively: Providing Robots with the Capacity for Deception", *International Journal of Social Robotics*, 3(1): 5-26, 2011.

[8] Wagner, A. and Arkin, R.C., "Robot Deception: Recognizing when a Robot Should Deceive", *Proc. IEEE International Symposium on Computational Intelligence in Robotics and Automation*, Dec. 2009.

[9] Shim, J. and Arkin, R.C., "Other-oriented Robot Deception: How can a robot's deceptive feedback help humans in HRI?", *Eighth International Conference on Social Robotics*, Nov. 2016.

[10] Shim, J. and Arkin, R.C., "The Benefits of Robot Deception in Search and Rescue: Computational Approach for Deceptive Action Selection via Case-based Reasoning", *IEEE International Symposium on Safety, Security, and Rescue Robotics*, Oct. 2015.

[11] Shim, J. and Arkin, R.C., "A Taxonomy of Robot Deception and its Benefits in HRI", *Proc. IEEE Systems, Man and Cybernetics Conference*, Oct. 2013.

- Provide a short abstract of 150-250 words suitable for inclusion in a program.

SHORT ABSTRACT

Trust, dependability, cohesion, and capability are integral to an effective team. These attributes are the same for teams of robots. When multiple teams with competing incentives are tasked, a strategy, if available, may be to weaken, influence or sway the attributes of other teams and limit their understanding of their full range of options. Such strategies are widely found in nature and in sporting contests such as feints, misdirection, etc. This talk focuses on one class of higher-level strategies for multi-robots, i.e., to intentionally misdirect using shills or confederates where needed, and the ethical considerations associated with deploying such teams. As multi-robot systems become more autonomous, distributed, networked, numerous, and with more capability to make critical decisions, the prospect for intentional and unintentional misdirection must be anticipated.

While benefits are clearly apparent to the team performing the deception, ethical questions surrounding the use of misdirection or other forms of deception are quite real and we have published earlier on this topic. The IEEE Global Initiative on the Ethics of Autonomous and Intelligent Systems has produced a preliminary set of recommendations on agent deception among other topics. The author served as co-chair of the Affective Computing Committee for this initiative that was responsible for these guidelines regarding deception, ensuring that consideration of the ethical aspects of this work and its social implications will reach a large audience, and they will be discussed here.