Human Interactive Machine Learning for Trust in Teams of Autonomous Robots

Robert S. Gutzwiller & John Reeder
Space and Naval Warfare Systems Center Pacific
San Diego, USA
gutzwill; jreeder@spawar.navy.mil

Abstract— Unmanned systems are increasing in number, while their manning requirements remain the same. To decrease manpower demands, machine learning techniques and autonomy are gaining traction and visibility. One barrier is human perception and understanding of autonomy. Machine learning techniques can result in "black box" algorithms that may yield high fitness, but poor comprehension by operators. However, Interactive Machine Learning (IML), a method to incorporate human input over the course of algorithm development by using neuro-evolutionary machine-learning techniques, may offer a solution. IML is evaluated here for its impact on developing autonomous team behaviors in an area search task. Initial findings show that IML-generated search plans were chosen over plans generated using a non-interactive ML technique, even though the participants *trusted* them slightly less. Further, participants discriminated each of the two types of plans from each other with a high degree of accuracy, suggesting the IML approach imparts behavioral characteristics into algorithms, making them more recognizable. Together the results lay the foundation for exploring how to team humans successfully with ML behavior.

Keywords— Human automation interaction; machine learning; supervisory control; unmanned vehicles; robots

I. Human Automation Interaction

Methods of collaborative human automation interactions are becoming more crucial in many domains, and especially in human-automation and human-robotic command and control [1]–[3]. At the same time, the methods of invoking control are advancing rapidly, and new solutions and methods are being found using machine learning techniques. Unlike standard programming, machine learning results in behaviors that are not always explainable; in fact this problem is robust enough that a new DARPA program, XAI, is being stood up to make these brilliant techniques easier to understand.

The importance of understanding what algorithms are capable of doing is obvious when you are co-located with a potentially dangerous device. Thus for human-robot interaction, physical proximity creates a demand for high trust between the humans and the machines [4]–[6]. Less intuitively, trust in unmanned systems and autonomy is still needed when these systems are operated from a distance through command abstractions, such as supervisory control. Moreover, supervisory control is precisely where machine-learning algorithms should be leveraged in helping to determine the best mixtures of tasks, vehicles, and operator performance for mission success.

This work was funded by ARPI IMPACT project and an ILIR funded by the Office and Naval Research. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the Department of Defense or the United States. We thank Daniel Barber and Julien Abich, and their research assistants for help in data collection.

Military command and control (C2) is traditionally a human-dominated area. Agents within a C2 framework (humans) are routinely given their autonomy via the commander. Such autonomy is in relation to achieving some underlying mission, and one can call this autonomy the provision of "commander's intent", which conveys high-level goals. Through the communication of intent, the commander and the supervised agents (whether human or computer) both develop expectancies. The commander expects to (a) not be asked to define individual actions toward achieving his goals, but also (b) that any actions taken can still be reasonably assessed within the mission context as adhering to intention or not, so that the commander can maintain mission alignment [7]. Opaque agents and behavior, such as those created by machine learning, then, by definition impede C2 by disrupting the ability to assess alignment. If the commander cannot develop expectancies, it is unlikely that trust will form with machine learning algorithms, which otherwise may be excellent performers. How can we bridge that gap? Interactive Machine Learning may provide one such method, to be described later. But first, we describe some of the major challenges facing human-automation interaction.

II. Automation Failures, Trust And Reliance

If implemented, machine learning will be difficult to supervise [8], and calibrated trust will be nearly impossible to achieve as it relies critically on understanding the intentions and behaviors of the system (transparency – see [9]). Trust in automation is a complex research area, well summarized across several reviews [10], [11]. Lee and See [10] outlined three general bases for development of trust for automation in humans: performance of the automation (does it fail unexpectedly), process (whether the automation is understandable and fits well into the users workflow), and purpose (the automation functions as intended). Though purpose, process and performance can form the basis for trust, trust is still different from reliance (the choice to use the agent or automation.) For example, one can choose not to use a robot to perform a task, even though it could be very trustworthy; or vice versa, distrust a system but have no choice but to rely on it under certain circumstances, such as cognitive overload [12].

Many different failures in human-automation interaction can be traced back to faulty trust calibration (how well users align their trust with the actual capability of an automated system). Calibration in this sense is one of several ideal states of cooperation then. Calibrated users should then have a more accurately informed decision process, avoid misuse

(because one will not over-rely on, nor fail to monitor the system), and so calibration promotes appropriate use [13].

Often humans must rely on their perception of an automated system or robots ability and behavior. The more obvious these abilities and behavioral intentions are, the more obvious failure states become. It is not that a system has to be perfect in order to be trusted, but it must be somewhat predictable; trust is more calibrated if one can "trust" automation to make certain kinds of mistakes (e.g. [14]) but not others.

With an opaque system, the operator cannot compensate for these faults (risking mission performance), in part because the expectancies surrounding failure conditions are not obvious. Calibrated and high-resolution trust is less likely because automation mistakes are not observable. Many have suggested increasing automation transparency is needed to improve teaming here; but the tradeoff with transparency in this case is that opaque systems may provide more optimal solutions. Neuroevolutionary computation [15]–[17] is one such method; the serious downside to neuroevolutionary computation is that it can result in "black boxes" from the human operator's point of view, which can make its application unsuitable for the real world. When applied to robotic plans, it may have the user asking questions like "What is this robot doing? What is it going to do? Why did it do that?"

The focus on increasing the optimality of these systems, largely performed in the domains of computer science and mathematics, generally ignores the need for user interaction. We attempt to mitigate the notable downside of generating black box solutions with new methods, as explained below, seeking to make their behavior more tolerable to the human supervisors and commanders who might oversee their operation.

III. Interactive Machine Learning

A new method to improve the amount of comprehension between the user and an evolved agent may be incorporating the user into the evolution of the system. Such inclusion may be achieved by allowing humans to define goal states, and offer guidance through user input along an evolutionary path. Past efforts have used human input in evolutionary mechanisms with positive results. However, the evolutionary goals have been only been idiosyncratic, such as beauty [18], [19]. We seek to explore similar techniques in the development of autonomous vehicle team behaviors, which have much more objective goals such as movement to locations and coverage of definable areas, in order to provide more understandable and predictable behaviors. These are key aspects of developing and maintaining situation awareness during operation [20], [21], and build on the growing movement to convey "transparent" information from autonomous systems [22]. Our central hypothesis is that IML will develop behaviors (plans in this experiment) that adhere more closely to user goals and expectations. Plans should be more identifiable and trustworthy as a result. We focused on three questions: (1) does the incorporation of humans in deriving ML algorithms, through IML, lead to more human trust in the plans that are generated? (2) Do participants, who helped generate plans, recognize, and be able to differentiate between IML and black box plans (which used neuroevolution, but no human involvement). Finally, (3) does the amount of neuroevolution that occurs, represented as steps, affect either trust or plan recognition?

IV. Experiment

Sixty participants were recruited from the University of Central Florida, and received payment ($15/hr) as compensation, in compliance with all IRB statutes. Participants completed a trust pre-experiment survey [23]; then performed in three phases: training, comparison, and labeling.

A. Training Phase

Participants were taught about the goal of three robots trying to search two areas effectively, and that the human role was to help train automated behaviors to maximize the amount of the area searched. A set of robot search agents in a virtual environment were shown exploring a space. Agents were autonomous and left signal decay trails in their wake, allowing participants to view how much of the targeted area had been searched. Participants responded by choosing from these options a good behavior to evolve further. Participants were counter-balanced across the frequency of user input in IML (a decision prior to every 10 or every 25 steps of evolution). With fewer steps of evolution, the human has more "say" in the outcome. After making a selection, the algorithm evolved, and then new "plans" were presented. Participants responded through approximately 410 steps of evolution due to time constraints (about 40 points of interaction for 10-step, and only about 16 points of interaction for 25-step).

B. Comparison Phase

After training, participants were shown *two* teams in action. One of the two teams was IML and the other was black box, with the location of each team on the screen randomized (left or right). Plan pairs were chosen on the backend to equate fitness between them. When plans stopped participants selected the plan they believed would best cover the designated areas, and then made a response, 1-100 on a sliding trust scale indicating 1 for no trust and 100 for complete trust in the chosen plan.

C. Labeling Phase

Following comparison, participants were shown a single team in action, and asked whether the team was IML, or black box. The interactive evolution teams were drawn from the specific individual's set of IML plans. Approximately 50% of each type of plan was shown randomly over 80 trials. Participants were given immediate feedback on their answers. At the end of the phase, participants were asked for their decision criteria for determining whether the teams in action had human IML, or were the evolved plans. Responses to the last question ended the experiment. Following, participants were debriefed and thanked for their participation.

V. Initial Results

A. Comparison Results

Overall, participants chose the IML plans more often (M= .66, SD= .16), and their trust in these choices overall was moderate to high (M= 61). At least on this initial level we have

found users believed that IML-generated behaviors were better than black box, when given the choice between them. Considering no fitness difference between them, users must be basing their choice on other characteristic imparted by the participants in the IML training phase. There was no interaction between plan choice and the amount of user involvement (10 versus 25 steps). However participants completing the 25 step condition first choose IML plans more often (71.5% versus 60% of the time; $F(1,44)= 5.93$, $p= .02$, $n_p^2= .12$).

After eliminating participants with missing data, trust was rated higher for plans with less (25-step) human involvement ($F(1,34)= 4.9$, $p= .04$, $n_p^2= .13$). Participants trusted black box plans ($M= 63$) more than IML ($M= 60$) plans ($F(1,34)= 6.5$, $p= .02$, $n_p^2= .16$). Such a difference may not be operationally significant, as both effects were small and with no interaction. Yet the data are interesting because they conflict with the choice data above - **IML plans were chosen more, but trusted less**.

B. Labeling Results

Participants labeled plans accurately, regardless of counterbalance condition or whether the behaviors were pulled from 10 or 25 steps training phases ($M= .77$). When wrong, participants were equally likely to mistake a behavior as IML as black-box, a check on whether our participants were biased in responding to all trials as one or the other type. To be accurate, participants had to be detecting or recognizing some difference between the plan types.

VI. Discussion

Involving humans in generating neuroevolutionary behaviors for teams of agents (IML) resulted in behaviors that participants choose more often, and were able to recognized. The importance of this first step is key, as it suggests that IML imparts traits to ML behaviors, which could be tuned to increase the expectancy and alignment of teams of machines. As mentioned in the introduction, this is a key limitation to employment. Despite their preferences, participants trusted IML plans slightly less than black-box plans, despite generally good trust of plans ($M= 61$).

From a methodological standpoint, IML appears to have been effective even with small amounts of user involvement. Users may be imparting traits, correcting early, common "odd" behaviors of the algorithms, or possibly, it was their active involvement in the behavior development that made it familiar to them. No matter the explanation our work shows a hopeful avenue for exploration toward making otherwise opaque algorithms useful, and creating expectancies or familiarity for the user.

Although machine learning offers required advantages, it can be opaque to users and reduce their awareness, confounding C2. We have shown there is promise in interactive machine learning techniques that increase user selection of team behaviors compared to pure evolution alone.

References

[1] M. Cummings, J. How, A. Whitten, and O. Toupet, "The impact of human–automation collaboration in decentralized multiple unmanned vehicle control," Proc. IEEE, vol. 100, no. 3, pp. 660–671, 2012.

[2] J. Hoc, "Towards a cognitive approach to human–machine cooperation in dynamic situations," Int. J. Hum. Comput. Stud., vol. 54, no. 4, pp. 509–540, Apr. 2001.

[3] J. Hoc and C. Chauvin, "Cooperative implications of the allocation of functions to humans and machines," Unpublished paper, 2011.

[4] V. Groom and C. I. Nass, "Can robots be teammates? Benchmarks in humanrobot teams," Inter. Stud., vol. 8, no. 3, pp. 483–500, 2007.

[5] J. R. Lee and C. I. Nass, "Trust in computers: The computers-are-social-actors (CASA) paradigm and trustworthiness perception in human-computer communication," in Trust and technology in a ubiquitous modern environment: Theoretical and methodological perspectives, D. Latusek and A. Gerbasi, Eds. Information Science Reference, 2010, pp. 1–15.

[6] C. I. Nass, B. J. Fogg, and Y. Moon, "Can computers be teammates?," Int. J. Hum. Comp. Stud., vol. 45, pp. 669–78, 1996.

[7] R. Willard, "Rediscover the art of command and control," Proc. United States Nav. Inst., vol. 128, no. 10, pp. 52–54, 2002.

[8] T. B. Sheridan and R. Parasuraman, "Human-automation interaction," Rev. Hum. Factors Ergon., pp. 89–129, Jan. 2005.

[9] T. L. Sanders, T. Wixon, K. Schafer, J. Y. C. Chen, and P. A. Hancock, "The influence of modality and transparency on trust in human-robot interaction," In Proc. of the IEEE CogSIMA conference, pp. 156–159, 2014.

[10] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," Hum. Factors, vol. 46, no. 1, pp. 50–80, 2004.

[11] T. Sanders, K. E. Oleson, D. R. Billings, J. Y. C. Chen, and P. A. Hancock, "A model of human-robot trust: Theoretical model development," In Proc. Hum. Factors Ergon. Soc. Annu. Meet., vol. 55, no. 1, pp. 1432–1436, 2011.

[12] C. D. Wickens, J. Hollands, S. Banbury, and R. Parasuraman, Engineering psychology and human performance, 4th ed. Upper Saddle River, NJ: Pearson, 2013.

[13] R. Parasuraman and V. Riley, "Humans and automation: Use, misuse, disuse, abuse," Hum. Factors, vol. 39, no. 2, pp. 230–253, 1997.

[14] A. Freedy, E. DeVisser, G. Weltman, and N. Coeyman, "Measurement of trust in human-robot collaboration," In Proc. Int. Conf. Collab. Technol. Syst., 2007.

[15] J. Gauci and K. O. Stanley, "Generating large-scale neural networks through discovering geometric regularities," In Proc. 9th Annu. Conf. Genet. Evol. Comput. - GECCO '07, pp. 997–1004, 2007.

[16] K. O. Stanley, D. B. D'Ambrosio, and J. Gauci, "A hypercube-based encoding for evolving large-scale neural networks.," Artif. Life, vol. 15, no. 2, pp. 185–212, Jan. 2009.

[17] K. O. Stanley and R. Miikkulainen, "Evolving neural networks through augmenting topologies," Evol. Comput., vol. 10, no. 2, pp. 99–127, May 2002.

[18] J. Clune, K. O. Stanley, R. T. Pennock, and C. Ofria, "On the performance of indirect encoding across the continuum of regularity," IEEE Trans. Evol. Comput., vol. 15, no. 3, pp. 346–367, Jun. 2011.

[19] J. Secretan, et al., "Picbreeder: evolving pictures collaboratively online," In Proc. of the SIGCHI Conference, pp. 1759–1768.

[20] M. R. Endsley, "Situation awareness misconceptions and misunderstandings," J. Cogn. Eng. Decis. Mak., vol. 9, no. 1, pp. 4–32, 2015.

[21] M. R. Endsley, "Toward a theory of situation awareness in dynamic systems," Hum. Factors, vol. 37, no. 1, pp. 32–64, 1995.

[22] J. Y. C. Chen, K. Procci, M. Boyce, J. Wright, A. Garcia, and M. Barnes, "Situation Awareness – Based Agent Transparency," US Army Research Laboratory, Aberdeen Proving Ground, MD, Rep. ARL-TR-6905, 2014

[23] J. Jian, A. Bisantz, & C. Drury, "Foundations for an empirically determined scale of trust in automated systems," Int. J. Cog. Ergo., vol. 4, no. 1, 53-71.