Toward Transparent Communication of Machine Self-Confidence for Uncertain Decision-Making

Jamison McGinley*, and Nisar Ahmed*
*Smead Aerospace Engineering Sciences
University of Colorado Boulder, Email: {jamison.mcginley, nisar.ahmed}@colorado.edu

I. INTRODUCTION

Autonomous robots will be used in a variety of challenging environments to accomplish complex tasks at human behest. These tasks will be delegated and be supervised by human users, who (despite not necessarily being robotic experts) must be able to establish a basis for correctly using and depending on robotic autonomy for success. The willingness to depend is referred to as *trust*, a subjective measure which, in part, is a function of human belief in an agent's *competency*, as well as belief in the predictability and the 'normality' of the tasking situation (among many other factors) [8]. In operational contexts, it is well-known that the trust developed by a user towards an autonomous system could potentially lead to an inaccurate understanding of that system's capabilities [4]. Such misunderstanding can lead to improper tasking of the agent, and subsequent misuse, abuse, or disuse of autonomy.

One strategy to address these issues is for an autonomous robot to report its own perspective on tasks at hand [10, 11]. If done correctly, a user could better judge whether the robot is sufficiently capable of completing a task within desired delegation parameters, thus adjusting user expectations of performance in a manner suitable to the situation at hand. This idea lies at the core of a wide spectrum of algorithmic strategies for generating *soft assurances*, which are collectively aimed at 'trust management' [8].

This work focuses on a particular class of algorithmic soft assurances that are based on assessments of 'machine self-trust', i.e. *machine self-confidence*. Formally, machine self-confidence can be defined as an autonomous agent's own perceived competency to execute tasks within desired parameters while accounting for uncertainties in its environment, states, and limited reasoning/execution capabilities [12, 5]. We use here the *Factorized Machine Self-confidence (FaMSeC)* computational framework developed in refs. [1, 6, 7], which considers computation of several interrelated (and non-exhaustive) factors that enable autonomous decision-making agents to generate machine self-confidence assessments in the context of solving problems described by Markov decision processes (MDPs):

- *command interpretation (CI)*: to what extent does the agent understand user intent for achieving some objective, and translate this into a context appropriate task?
- *model validity (MV)*: are the agent's learned/assumed models and training data used for decision-making sufficient for operating in the real world?
- *solver quality (SQ)*: are the approximations used by the system for solving decision-making problems appropriate for the given task and model?
- *outcome assessment (OA)*: do the sets of possible events, rewards, costs, utilities, etc. for decisions governed by a policy lead to a desirable landscape of outcomes?
- *past performance (PP)*: what can be inferred from the system's own experience and other available information for past or similar problem instances?

Computed scores for each factor can be mapped to notional scales with upper/lower bounds U_i/L_i, where L_i gives a shorthand indication of 'complete lack of confidence' (i.e. some aspect of task, environment, or operation context falls completely outside the agent's competency boundaries), and U_i indicates 'complete confidence' (i.e. all aspects are well within system's competency boundaries). To date, computable metrics have been investigated for 2 factors. For OA, ref. [1] proposed using a logistic transform of the upper/lower partial moment (UPM/LPM) of the non-discounted cumulative reward pdf $p(R_\infty)$ generated by a given policy π, to measure the 'expected margin of success' that an agent expects to achieve for completing the MDP task via π (versus a minimum total reward bound R_*, which is assumed to reflect user performance expectations and assumes total confidence in all other factors),

$$\frac{UPM}{LPM}|_{R_*} = \frac{\int_{R_*}^{\infty}(R_\infty - R_*)\cdot p(R_\infty)dR_\infty}{\int_{-\infty}^{R_*}(R_* - R_\infty)\cdot p(R_\infty)dR_\infty}. \quad (1)$$

For SQ (assuming 'floating' OA and total confidence for all other factors), refs. [6, 7] proposed learning-based surrogate modeling strategies to compare the $p(R_\infty)$ pdfs of approximate MDP policies $\hat{\pi}$ derived from 'live candidate' solvers (e.g. online MCTS that must be run aboard limited hardware [9]) to those of 'trusted' MDP policies derived offline π^* (e.g. which require more resources than available on a live platform). Refs. [6, 7] also performed a large user study, which determined that reporting OA and SQ factors via exceedingly simple user interfaces had significant and substantial impacts on supervisory delegation of adversarial navigation tasks. We build here on the insights of this prior work to formulate a follow-up user study, with the dual aims of: (1) validating the utility of self-confidence reporting in uncertain tasking contexts relevant to scientific exploration, and (2) determining the extent to which additional elements of *transparent reasoning* are necessary and useful for self-confidence reporting interfaces.

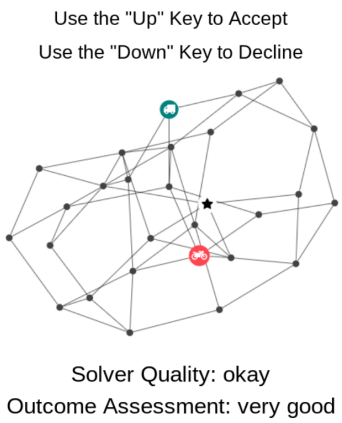

Fig. 1: Simple self-confidence reporting UI from study in [6].

II. PRIOR AND PROPOSED USER STUDIES

Figure 1 shows the simple interface used for the MDP-based navigation task studied in refs. [6, 7], where a human user (supervisor) must decide whether to dispatch an autonomous delivery truck to a goal location in a complex road network. The truck uses an approximate policy given by MCTS to find its way to the goal, while also avoiding a mobile bandit that can intercept the truck and prevent delivery. This task was used as part of a large ($N = 255$ person) study of compensated MTurk users to assess whether FaMSeC-based OA and SQ reporting (provided as Likert values mapped from numerical scores) could improve supervisory performance (measured by total delivery score and time spent on multiple tasking instances) and self-reported trust (measured by follow up surveys). By allowing users to decide only whether to dispatch the truck ('go') or not ('no go') on 30 different tasking instances, confounding influences on trust could be mitigated and non-expert users could more easily grasp the supervision task (as measured by pre-task quizzes and follow up surveys). This allowed for clearer assessments on the value of self-confidence reports.

The hypotheses tested in this scenario were that reporting of OA, SQ, or both would improve the combined delivery score (versus the control baseline of having no self-confidence reports at all), increase the operator's self-reported trust, and allow for faster decision making by the human. Using a between subjects design for the different self-confidence reporting cases, the delivery score awarded +1 point for each successful 'go' call (reached goal), -1 point for each failed 'go' call (captured by bandit), and -0.25 points for each 'no go' call (declined dispatch). The hypotheses stem from the idea that the score can only be maximized if the user understands when it is appropriate to dispatch the truck given the stochastic nature of the task and approximate nature of the policy – the impacts of which on agent performance are (in principle) conveyed by OA and SQ reporting, respectively. Each subject's score, decision times and responses to a questionnaire related to trust were collected. Analysis of the data revealed that the presence of either/both FaMSeC factors had strong positive effect on cumulative score, a weaker positive impact on self-reported trust, and a negligible impact on decision making time.

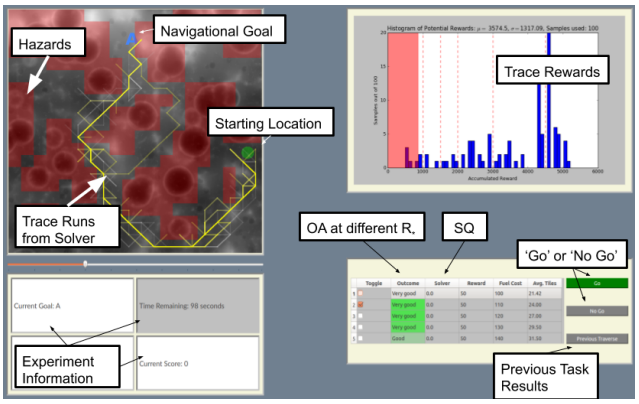

Fig. 2: Transparent reporting UI for lunar navigation task.

The aim of the proposed user study is two-fold: (1) to validate the findings from refs. [6, 7] in a similar but non-adversarial uncertain decision-making domain that is more akin to scientific exploration missions, and (2) to examine how additional transparent reporting functionality might enable improvements in self-reported trust. For (1), we consider a rover driving on the moon that has limited navigational capacity and a mission that requires it to traverse the landscape, i.e. to extract and analyze rock samples at specific locations. This task has a openly traversable terrain with hazardous areas (craters, boulders, etc.). to be avoided, so that the rover does not get stuck, damaged, or run out of fuel. As such, more real world depth is available in this task for validating the utility of OA and SQ reporting through a structurally similar user study.

With regards to (2), it is hypothesized that allowing users to 'drill-down' into FaMSeC values can improve self-reported trust in the agent over multiple task instances. For instance, if the rover fails a task when it had low confidence and the operator can contextualize the lack of confidence (e.g. presence of several hard to avoid obstacles), then loss of faith by the operator can be potentially curbed [3]. Leveraging the same scoring scheme and questionnaire approach from the original study, we can evaluate the potential to augment the effectiveness of these OA and SQ reporting while also improving self-reported operator trust in autonomy [2]. Specifically, we will formulate and present a user study to systematically assess the hypothesis that increased transparency and contextualization of OA and SQ reports will allow operators to better understand the system's competency and achieve a higher cumulative task score, while reporting increased self-reported trust.

Figure 2 shows a prototype UI with transparent reasoning features that allow users to 'drill down' into FaMSeC assessments, before deciding whether to dispatch the rover. The estimated $p(R_\infty)$ pdf is shown in the upper right with a number of relevant R_* values marked for reference to indicate various possible performance margins. This feature allows the user to investigate how different performance margin expectations influence reported confidence and to connect R_∞ values along the pdf to expected task execution outcomes, overlaid as navigation on the left portion of the interface. User interactions with interface elements will be recorded to analyze statistics on which are actually used to inform self-confidence analysis.

ACKNOWLEDGMENT

This work is supported by the DARPA Competency Aware Machine Learning (CAML) program and CU Boulder Discovery Learning Apprenticeship (DLA).

REFERENCES

[1] Matthew Aitken. Assured human-autonomy interaction through machine self-confidence. Master's thesis, University of Colorado Boulder, 2016.

[2] Gershon Weltman Amos Freedy, Ewart DeVisser. Measurement of trust in human-robot collaboration. 2007. Proceedings of the Symposium on Collaborative Technologies Systems.

[3] M. Desai, M. Medvedev, M. Vazquez, S. McSheehy, S. Gadea-Omelchenko, C. Bruggeman, A. Steinfeld, and H. Yanco. Effects of changing reliability on trust of robot systems. In *Proceedings of the 2012 7th ACM/IEEE International Conference on Human-Robot Interaction*, pages 73–80. IEEE, 2012.

[4] Mary T Dzindolet, Scott A Peterson, Regina A Pomranky, Linda G Pierce, and Hall P Beck. The role of trust in automation reliance. *International journal of human-computer studies*, 58(6):697–718, 2003.

[5] Andrew R Hutchins, Mary L Cummings, Mark Draper, and Thomas Hughes. Representing autonomous systems' self-confidence through competency boundaries. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 59, pages 279–283. SAGE Publications Sage CA: Los Angeles, CA, 2015.

[6] Brett Israelsen. *Algorithmic Assurances and Self-Assessment of Competency Boundaries in Autonomous Systems*. PhD thesis, University of Colorado at Boulder, Boulder, 2019.

[7] Brett Israelsen, Nisar Ahmed, Eric Frew, Dale Lawrence, and Brian Argrow. Machine self-confidence in autonomous systems via meta-analysis of decision processes. In *International Conference on Applied Human Factors and Ergonomics*, pages 213–223. Springer, 2019.

[8] Brett W Israelsen and Nisar R Ahmed. "Dave... I can assure you... that it's going to be all right..." a definition, case for, and survey of algorithmic assurances in human-autonomy trust relationships. *ACM Computing Surveys (CSUR)*, 51(6):1–37, 2019.

[9] Mykel J Kochenderfer. *Decision making under uncertainty: theory and application*. MIT Press, 2015.

[10] Steve McGuire, P Michael Furlong, Christoffer Heckman, Simon Julier, Daniel Szafir, and Nisar Ahmed. Failure is not an option: Policy learning for adaptive recovery in space operations. *IEEE Robotics and Automation Letters*, 3(3):1639–1646, 2018.

[11] Steve McGuire, P Michael Furlong, Terry Fong, Christoffer Heckman, Daniel Szafir, Simon J Julier, and Nisar Ahmed. Everybody needs somebody sometimes: Validation of adaptive recovery in robotic space operations. *IEEE Robotics and Automation Letters*, 4(2):1216–1223, 2019.

[12] Nicholas Sweet, Nisar R Ahmed, Ugur Kuter, and Christopher Miller. Towards self-confidence in autonomous systems. In *AIAA Infotech@ Aerospace*, page 1651. 2016.