Assurances and Machine Self-Confidence for Enhanced Trust in Autonomous Systems

Matthew Aitken, Nisar Ahmed, Dale Lawrence, Brian Argrow, and Eric Frew
Department of Aerospace Engineering Sciences, University of Colorado at Boulder
Boulder, CO 80309
email: [Matthew.Aitken, Nisar.Ahmed, Eric.Frew]@colorado.edu

This work investigates a model-based approach to understanding how user trust evolves in systems consisting of a supervising user and an autonomous agent. This model consists of a multivariate model for user trust, and a feedback connection between user and agent. Feedback information is termed assurance, which is also shown to consist of multiple aspects concerning the state of the autonomous agent. It is argued that the closed loop interactions between user and agent can and should be designed to calibrate user trust. In order to develop design principles, it is first necessary to define the terms and salient components of these models to provide a logical framework for their interconnection. Although elements such as trust and assurance are essential in a usable autonomous system, they are also nebulous concepts with multiple meanings [1]. We provide definitions and structure that enables a systematic study of the problem. Our user trust model implies that better assurances can be designed by providing users with better insight into the 'competency boundaries' for key decision-making components of the autonomy. One potentially important assurance is a report of the *self-confidence* (i.e. self-trust) the autonomy has in its own process. We are currently developing formal computational mechanisms for assessing machine self-confidence as an assurance in the context of a probabilistic autonomous route planning.

Model Overview

Trust is an important topic that is of interest to many different communities. An overview of literature on the subject reveals work that spans fields such as psychology [2, 3], sociology [3, 4], business [1, 5], and engineering [6, 7, 8, 9, 10, 11]. A key point is that trust is not one dimensional— trust models that will be used for design must be developed to capture many intricate aspects of human trust. The focus of trust research in the business and engineering worlds are most similar. It is driven by not only understanding the user trust (as in psychology) or relationships with autonomous systems (as in sociology), but how feedback can be carefully chosen to ensure proper development of trust when human agents are replaced by autonomous machine agents. An important question in e-commerce, for instance, is whether people are willing to trust legal advice from a website if another human is not also speaking with them in person [1]. Trust in engineering design has been well-studied for automation systems [6, 7, 8], and has formed much of the basis for understanding autonomous systems, which present new challenges that are non-existent with automation [9, 12, 10, 11]. Certification processes are also an embodiment of the idea of developing an appropriate level of trust in a system. Acknowledging the feedback of assurances to influence user trust in autonomy is, however, a fairly new idea. The central theme of this work is that an autonomous system can and should be designed to help properly calibrate user trust. To develop design principles, it is necessary to define the parameters of the design space and provide a rigorous framework for analysis and synthesis. Our basic model of the user trust relationship is seen in Figure 1. Key definitions are:

- **Autonomous System**: an agent or system comprised of a machine being driven or controlled by some form of autonomy. An autonomous system always interacts with a human user.
- **Autonomy**: the ability to perform complex mission-related tasks with substantially less human intervention for more extended periods of time, sometimes at remote distances, than current systems [13].
- **Machine**: the physical system and supporting low-level electronics and/or software in an autonomous system commanded by the autonomy.
- **Trust**: a user's willingness to depend on the autonomous system [1]. Trust depends on a particular system or situation, as well as a user's background dispositions and beliefs. Trust always leads to an action.
- **Assurance**: the autonomous system's ability to affect the user's trust. As used here, the term is not intended to have a positive or negative connotation - assurances can decrease trust.

This model provides a structured framework for understanding how an individual's trust will be used for interaction and how trust can develop over time. Looking first at the 'User Trust' block, there are several different inputs. Past experiences assist in the initial development of trust with a system. Similar to the idea of a Markov Chain, past experiences are what help initialize the new state of trust about a particular autonomous system. Another input, "Environmental Factors", describes how outside variables can affect the user's trust in a system. These variables relate to how the autonomous system might perform in the given scenario, but are independent from the autonomous system's properties. For example, bad weather might influence the user's trust in a self-driving car.

Trust actions are the only output of the 'User Trust' block.

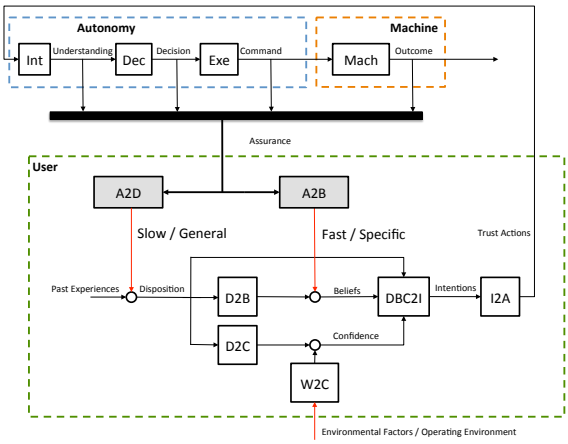

Fig. 1. Closed loop model of a user trust and an autonomous system

These actions are labeled as such because they demonstrate the user's trust or lack thereof. The actions of the user influence the behavior and performance of the autonomous system. The resulting properties of the autonomous system to this trust action input are then fed back to the user trust block as assurances. Assurances affect the user trust states, causing a dynamic change in the user's trust and therefore the user's trust actions. This model has several important implications. First, the model makes clear that trust is only relevant in so far as it leads to a trust action — therefore the autonomous system should include an interface with affordances that are rich enough to allow the user to make various decisions to reflect her trust intentions. Second, this model emphasizes that trust is a property of the user, whereas the assurances determines the characteristics of the autonomous system. Third, the model provides insights into how to certify autonomous systems. Certification attempts to quantify the necessary levels of trust that must be prevalent in the system before using it in a way that might harm others or cause failure. The model suggests that we can develop trust in an autonomous system and then license it through the same process used to license trusted human pilots, drivers, etc. Finally, the model implies that better assurances can be designed by providing better insight into the Understanding and Decision components of the Autonomy. One interesting and potentially important assurance is a report of the *self-confidence* (i.e. self-trust) that the autonomy has in its own process.

Machine Self-Confidence as an Assurance

Reporting self-confidence would mimic the way in which human drivers and pilots are assessed for the purposes of obtaining a driver's or pilot's license: an evaluator puts the person through a series of tasks and evaluates outcomes together with the process/rationale of the person. By providing a report of self-confidence, the user is able to change their trust based on comparison between the actual outcome and the system's (predictive) confidence that the outcome would be achieved. Given the growing complexity and sophistication of autonomous systems and tasks to which they are assigned, the essential idea behind machine self-confidence is to generate a computable 'shorthand' metric that easily allows users to gain insight into the actual capabilities/limitations of autonomous systems, thus enabling proper calibration of trust.

Machine self-confidence can be formally defined as an autonomous agent's perceived ability to achieve assigned goals (within a defined region of autonomous behavior) after accounting for: (1) uncertainties in its knowledge of the world, (2) uncertainties of its own state, and (3) uncertainties about its reasoning process and execution abilities. Self-confidence, while certainly strongly coupled to uncertainty, goes beyond simple assessment of probabilities of whether assigned tasks can be successfully accomplished. As discussed in [14], the concept of self-confidence is closely tied to an autonomous agent's self-awareness of its 'competency boundaries'.

Several recent works have attempted to translate this definition into algorithms for computing and reporting machine self-confidence [15, 16]. In our ongoing work, we are developing mechanisms to evaluate machine self-confidence for sophisticated policy-based decision-making algorithms based on partially observable Markov decision processes (POMDPs). POMDPs are powerful and popular tools for solving complex optimal control and planning problems under uncertainty, but require sophisticated approximations in real-world systems. For a robot to understand the 'competency boundaries' of its POMDP planner, it must somehow recognize when and where such approximations may break down. This directly leads to consideration of 'self-confidence factors'. These factors can then be combined into a single 'self-confidence score', which can serve as a simple assurance that cues users on when to adjust their trust in the system (e.g. on a scale of -1 to 1, where -1 indicates complete lack of confidence in ability to complete a task and 1 indicates complete confidence).

We are currently considering quantitative metrics for scoring 5 factors that can be applied to self-confidence assessment in POMDP planning (as well as possibly other planning approaches): (1) model validity: is the model used for decision-making a reasonable representation of the real world?); (2) expected outcome assessment: does the distribution of expected rewards under a given policy indicate robustness and desirable interim behavior during task execution?; (3) solver quality: is the approximation being used to find a solution (policy) appropriate for the given problem and model?; (4) interpretation of user commands: did the autonomy understand the user's intentions and translate these into appropriate tasks?; (5) past performance: how well did the autonomy do on previous instances of the same problem or similar problems?

To connect these ideas with the user trust model in practice, we are developing a ROS simulation testbed involving user interaction with an autonomous unmanned ground vehicle, which must navigate in a highly uncertain environment using very limited information and very limited high-level interaction with a human. We are designing this testbed with human user studies in mind to deploy and study the impact of self-confidence as a possible assurance for properly calibrating trust in autonomy.

References

[1] D. H. Mcknight and N. L. Chervany, "What trust means in e-commerce customer relationships: An interdisciplinary conceptual typology.," *International Journal of Electronic Commerce*, vol. 6, no. 2, pp. 35–59, 2001.

[2] J. B. Rotter, "Interpersonal trust, trustworthiness, and gullibility.," *American Psychologist*, vol. 35, no. 1, pp. 1–7, 1980.

[3] D. M. Rousseau, S. B. Sitkin, R. S. Burt, and C. Camerer, "Not so different after all: A cross discipline view of trust," *Academy of Management Review*, vol. 23, no. 3, pp. 393–404, 1998.

[4] L. G. Zucker, "Production of trust: Institutional sources of economic structure, 1840-1920," *Research in Organizational Behavior*, vol. 8, no. 1, pp. 53–111, 1986.

[5] R. M. Morgan and S. D. Hunt, "The commitment-trust theory of relationship marketing," *Journal of Marketing*, vol. 58, no. July, pp. 20–38, 1994.

[6] J. D. Lee and N. Moray, "Trust, self-confidence, and operators' adaptation to automation," *International Journal of Human-Computer Studies*, vol. 40, no. 1, pp. 153–184, 1994.

[7] J. D. Lee and K. a. See, "Trust in automation: designing for appropriate reliance.," *Human factors*, vol. 46, pp. 50–80, 1 2004.

[8] R. Parasuraman, T. Sheridan, and C. Wickens, "A model for types and levels of human interaction with automation," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 30, pp. 286–297, 5 2000.

[9] F. Gao, A. S. Clare, J. C. Macbeth, and M. L. Cummings, "Modeling the impact of operator trust on performance in multiple robot control,," in *AAAI Spring Symposium: Trust and Autonomous Systems,*, AAAI, 3 2013.

[10] P. Kaniarasu, A. Steinfeld, M. Desai, and H. Yanco, "Robot confidence and trust alignment," in *ACM/IEEE International Conference on Human-Robot Interaction*, pp. 155–156, 2013.

[11] Y. Wang, Z. Shi, C. Wang, and F. Zhang, "Human-robot mutual trust in (semi)autonomous underwater robots," *Studies in Computational Intelligence*, vol. 554, pp. 115–137, 2014.

[12] P. Kaniarasu, A. Steinfeld, M. Desai, and H. Yanco, "Potential measures for detecting trust changes," *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction - HRI '12*, p. 241, 2012.

[13] C. on Autonomy Research for Civil Aviation; Aeronautics, S. E. B. D. on Engineering, and P. S. N. R. Council, *Autonomy Research for Civil Aviation: Toward a New Era of Flight*. The National Academies Press, 2014.

[14] A. Hutchins, M. Cummings, M. Draper, and T. Hughes, "Representing autonomous systems self-confidence through competency boundaries," in *Proc. of the 2015 Human Factors and Ergonomics Society Meeting (HFES 2015)*, 2015.

[15] U. Kuter and C. Miller, "Computational mechanisms to support reporting of self confidence of automated/autonomous systems," in *Proc. of the AAAI 2015 Fall Symposium on Self-confidence in Autonomous Systems*, pp. 18–21, 2015.

[16] N. Sweet, N. Ahmed, U. Kuter, and C. Miller, "Towards self-confidence in autonomous systems," in *Proc. of AIAA 2016 InfoTech@Aerospace, SciTech 2016*, 2016.