A Vision for Human-Machine Mutual Understanding, Trust Establishment, and Collaboration

Carlos R. B. Azevedo*, Klaus Raizer† and Ricardo Souza‡
Ericsson Research - Affiliated Brazilian Branch
* carlos.azevedo@ericsson.com, † klaus.raizer@ericsson.com, ‡ ricardo.s.souza@ericsson.com

***Abstract*—Human-machine interactions are likely to require synergistic multidisciplinary research efforts for supporting a paradigm shift towards collaborative-oriented use cases. An essential aspect of collaboration is trust and in order to establish it there is need for human-machine mutual understanding (HMMU). We argue that achieving HMMU will require evolving from an approach that reduces human factors as uncontrollable environmental elements, to one that repositions human emotions not only as a central part of an integrated control paradigm, but also as interpretable and steerable through appropriate information flows and mutual learning cycles. On the strategic decision-making side, we argue conflict resolution will require anticipating multiple trade-off situations that include human factors. On the operational level, symbiotic human-machine cognitive architectures should embed detected human emotions as inputs in shared machine control models. Trust measurements will play the role of mediating task coordination by pinpointing and dynamically composing appropriate situation-aware interaction protocols. In addition to a vision for HMMU, this paper proposes a multidisciplinary research strategy that attempts to unify the isolated efforts of different communities. The proposed vision is contextualized within a high-level research roadmap to support near and long-term activities in HMMU.**

***Keywords*—Human-Machine Interaction; Autonomous Systems; Situation-Awareness; Cognitive Agents; Trust; Shared Control.**

I. Introduction

The rapid deployment of autonomous systems in complex environments has brought about the need for interaction models that allow humans to shape and steer desired behaviors in such systems, in a safe manner. Modulating and sharing the control authority between human and computational agents is gaining notoriety in such cases, since it allows for collaborative systems that leverage the strengths and reduce the weaknesses of both humans and machines [1].

In order to build such systems there is need for establish trust among all parties. Trust helps agents embrace uncertainty as it obviates supervision, facilitates choice under risk, and can mediate conflict resolution in collective decision-making [2]. Furthermore, trust can enable decentralization and adaptive behavior in complex systems [3]. Multidisciplinary efforts are thus needed for enabling a collaborative society of humans and autonomous systems and for ensuring both parties can function synergistically to reach their full potential. Machines can effectively perform repetitive tasks, can sense and process information at faster rates, and can act on situations that are too complicated for human beings. Properly trained human operators, on the other hand, have better knowledge of situations and can internalize perceptions through emotions. However, there is still need for systems that learn from humans and vice-versa. Those learning cycles need to be mediated by explanation. DARPA's new research program on Explainable Artificial Intelligence corroborates this vision [4].

Fig. 1. Diagram for human-machine mutual understanding and collaboration.

Therefore, we argue that improved collaboration among humans and machines require three elements in sequence, namely, explanation capability, mutual-understanding and mutual trust, as illustrated in Figure 1.

If we want society to benefit from the synergistic effects of human-machine interaction, we thus need to interface human feelings and emotions with autonomous systems so that interactions can be perceived by users as more natural, timely, and effective in attaining collaboration towards a common goal.

II. From Affective Awareness to Trust-Guided Operations

A. The Role of Emotions

In the literature there has been two traditional ways to model emotions, namely, discrete models and dimensional models. A widely known example of the discrete type is the OCC emotion model proposed by Ortony, Clore and Collins [5], which describes a hierarchy of 22 emotions categories by evaluating the consequences of events, aspects of objects and the actions of agents. In dimensional models, on the other hand, each emotion is represented by a point in a multi-dimensional space. A widely known one is Russell's bipolar circumplex model [6], which has been used in a number of applications such as robotics [7] and emotion recognition [8]. Each model has its pros and cons, however Ron Sun et al. [9] have argued that emotions should not be viewed as a unitary thing, being rather emergent from the interactions among many cognitive processes. Therefore, nothing short of a

multi-purpose computational cognitive architecture [10] would be able to provide a comprehensive mechanistic interpretation of emotion.

In earlier works in the field of *affective computing*, Picard et al. [11] reported an automated process for recognizing eight different emotions, after collecting 30 days of data from multiple sensor modalities such as muscle tension, respiration, skin conductance, and heart variability, which achieved 81 percent accuracy in a lab. In more recent works they developed what she called the "first wearable affective computer", which was used to acquire real-world multi-modal driver stress data. The major lesson learned from these experiments was that a single modality - skin conductance - measured from the electrodermal activity (EDA) gave the highest correlation with multiple measures of stress, although heart rate and heart-rate variability were also sometimes helpful.

As the research community becomes aware of the need to enforce algorithmic accountability [12], designing interaction protocols that are able to explain how perceived affective elements are used in specific decisions becomes imperative. Such affective-enhanced explanations thus serve as a building block in the effort of establishing mutual understanding between humans and autonomous systems.

B. Trust in Autonomous and Automated Systems

Mayer et al. [13] defines interpersonal trust as the "willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the 'trustor', irrespective of the ability to monitor or control that party". We argue that trust between humans and autonomous systems should follow the same notion of interpersonal trust.

Most initiatives to evaluate trust of a user on autonomous/semi-autonomous systems adhere to a *performance-centric trust definition* that, according to authors, is naturally suited for supervisor-worker teams, where the artificial agent (worker) has no personal motivations as it collaborates with the humans (supervisor) towards a unified goal [14]. These strategies usually perform trust evaluation based on the human interactions [14] [15] (e.g. control inputs) or based on some evaluation of the efficiency of the agent itself [16] (e.g. how close can a robot follow a given trajectory). Different task delegation strategies are proposed in Burnett et al. [17], some of them based on monitoring effort levels applied by a trustee. Those strategies allows for utility-based models for decisions such as whether to delegate a task and to whom in scenarios requiring collaboration.

Our vision favors a trust-based model, where trust is defined not only by performance but also by mutual understanding between agents. Artificial agents should be able to perceive how their human counterparts are emotionally influenced by each decision while also being able to explain their actions.

Another common concept is that of *human-centered automation*, where there is an assumption that the human should always play the primary role in a task execution. Therefore, the human bears the ultimate responsibility for the system's safety [18]. This assumption is based on the notion that the user usually has more knowledge of the world state and of its implications than the artificial agent has. In our current technological state we disagree that the human will always have more knowledge than an artificial agent, although it may have complementary insights about certain situations. Therefore, we argue that responsibility should be shared among all parties based on each actor's strengths and weaknesses.

III. A Path to Mutual-Understanding

As seen in Figure 1, in order to reach the capability of *explanation* there is need for affective-guided operations and situation-aware interactions. Through *explanation* it is possible to reach *understanding* (mutual-understating protocols, in Fig. 2) among agents, which in turn allows for *trust* (trust-guided operations, in Fig. 2). In order to address this challenge, we present a roadmap for Human-Machine Mutual Understanding (HMMU) and cooperation, as seen in Figure 2. *Safety*, *resilience* and *adaptability* are the main drivers, which can be strengthened by the milestones devised in the roadmap.

Research in machine intelligence, biophysical signal processing, robotics, and communications enable the desired milestones, which can be achieved by employing the technologies and capabilities seen in the multi-disciplinary layers.

Emotions are powerful enhancers of performance and may spread rapidly among groups of people, whether in physical or virtual spaces [19]. Each emotional state can interfere on attention, disposition, reaction time, sometimes deviating people from safety and primary objectives. Proper implementation of cognitive and affective models are key for achieving affective-guided operations. Enhancing descriptors of system states with perceived user emotions and other human factors, and fusing them with contextual features can also augment human-machine joint situation awareness [20], leading to situation-aware interaction protocols (i.e. SIP).

Engineering machines that are able to *understand* users and operators, and to communicate back what they perceive will enable shared control and generate virtuous mutual learning cycles in which both parties can strengthen collaboration. The ability of communicating perceived situations (e.g. levels of urgency) in natural language is crucial in modes of interaction such as autonomous handover, where incipient studies on the effects of different messages on user emotions (e.g. annoyance) are being conducted [21]. Research in natural language interaction among autonomous systems and humans is expected to become more relevant as European citizens might soon have a right to request explanations of decisions taken by algorithms that affect them. This was concluded by examining a data protection law called the General Data Protection Regulation (GDPR), approved in April 2016. EU member states are expected to enforce the law in 2018.

Besides *mutual-understanding protocols*, in order to reach *trust-guided operations*, there is need to evaluate different types of existing interaction models and propose new ones, if necessary. Then it will be possible to propose models that represent *trust* levels between interacting agents. As a consequence of *trust*, we envision *collaborative operations*. In order to achieve this milestone, there is need for investigating affective analytics technologies and integrate them with multiple interaction modalities, such as haptic feedback, that

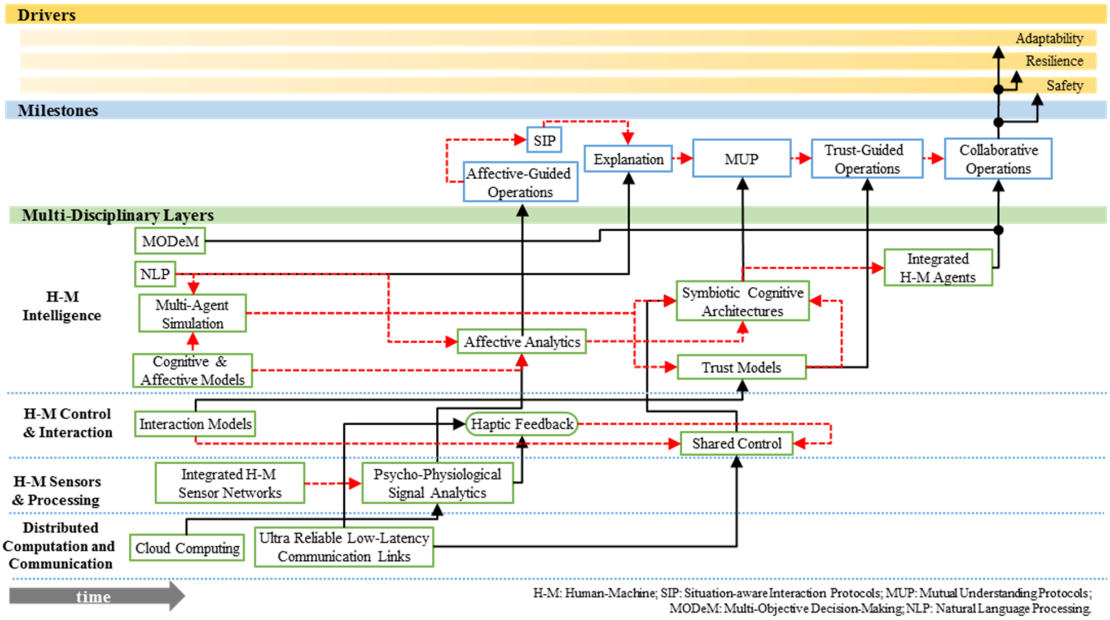

Fig. 2. A roadmap for human-machine mutual understanding and collaboration. For better clarity, full (black) arrows are connections between elements from different layers and dotted (red) arrows denote intra-layer connections. Boxes with rounded corners are non-mandatory.

can be integrated into and mediated by distributed *human-machine cognitive agents* that are capable of sharing the control responsibilities based on mutual-understanding principles.

Achieving the HMMU vision will lead to an integral and symbiotic relationship between humans and autonomous systems, thus unifying and amplifying natural and machine intelligences grounded on trust and collaboration, and therefore leading to better aligned values, goals, and purposes.

REFERENCES

[1] A. Broad, J. Schultz, M. Derry, T. Murphey, and B. Argall, "Trust adaptation leads to lower control effort in shared control of crane automation," IEEE Robotics and Automation Letters, vol. 2, no. 1, pp. 239–246, Jan 2017.

[2] T. L. Simons and R. S. Peterson, "Task conflict and relationship conflict in top management teams: the pivotal role of intragroup trust." Journal of Applied Psychology, vol. 85, no. 1, pp. 102–111, 2000.

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

[4] D. Gunning. (2017). Explainable artificial intelligence (xai). [Online]. Available: http://www.darpa.mil/program/explainable-artificial-intelligence

[5] A. Ortony, G. L. Clore, and A. Collins, *The cognitive structure of emotions*, New York, NY: Cambridge University Press, 1990.

[6] J. A. Russell, "A circumplex model of affect." Journal of Personality and Social Psychology, vol. 39, no. 6, pp. 1161–1178, 1980.

[7] G. Yoshioka, T. Sakamoto, and Y. Takeuchi, "Inferring affective states from observation of a robot's simple movements," in *Proc of the 24th IEEE Int. Sym. on Robot and Human Interactive Communication (RO-MAN)*, Aug 2015, pp. 185–190.

[8] V. H. Anh, M. N. Van, B. B. Ha, and T. H. Quyet, "A real-time model based support vector machine for emotion recognition through EEG," in *Proc of Int. Conf. on Control, Automation and Information Sciences (ICCAIS)*, Nov 2012, pp. 191–196.

[9] R. Sun, N. Wilson, and M. Lynch, "Emotion: A unified mechanistic interpretation from a cognitive architecture," Cognitive Computation, vol. 8, no. 1, pp. 1–14, 2016.

[10] A. L. Paraense, K. Raizer, S. M. de Paula, E. Rohmer, and R. R. Gudwin, "The cognitive systems toolkit and the cst reference cognitive architecture," Biologically Inspired Cognitive Architectures, vol. 17, no. 1, pp. 32–48, 2016.

[11] R. W. Picard, "Automating the recognition of stress and emotion: From lab to real-world impact," IEEE Multimedia, vol. 23, no. 3, pp. 3–7, July 2016.

[12] N. Diakopoulos, "Accountability in algorithmic decision making," Communications of the ACM, vol. 59, no. 2, pp. 56–62, 2016.

[13] R. C. Mayer, J. H. Davis, and F. D. Schoorman, "An integrative model of organizational trust," Academy of Management Review, vol. 20, no. 3, pp. 709–734, Jul. 1995.

[14] A. Xu and G. Dudek, "Maintaining efficient collaboration with trust-seeking robots," in *Proc of the 2016 IEEE/RSJ Int Conf on Intelligent Robots and Systems (IROS)*, Oct 2016, pp. 3312–3319.

[15] P. Kaniarasu, A. Steinfeld, M. Desai, and H. Yanco, "Potential measures for detecting trust changes," in *Proc of the 7th ACM/IEEE Int Conf on Human-Robot Interaction (HRI)*, 2012, pp. 241–242.

[16] P. Kaniarasu, A. Steinfeld, M. Desai, and H. Yanco, "Robot confidence and trust alignment," in *Proc of the 8th ACM/IEEE Int Conf on Human-Robot Interaction (HRI)*, March 2013, pp. 155–156.

[17] C. Burnett, T. J. Norman, and K. Sycara, "Trust decision-making in multi-agent systems," in *Proc of the 22nd Int Joint Conf on Articial Intelligence, ser. IJCAI11. AAAI Press*, 2011, pp. 115–120.

[18] T. Inagaki, "Smart collaboration between humans and machines based on mutual understanding," Annual Reviews in Control, vol. 32, no. 2, pp. 253–261, 2008.

[19] A. D. I. Kramer, J. E. Guillory, and J. T. Hancock, "Experimental evidence of massive-scale emotional contagion through social networks," in *Proc of the National Academy of Sciences*, vol. 111, no. 24, pp. 8788–8790, 2014.

[20] M. H. Martens and A. P. van den Beukel, "The road to automated driving: Dual mode and human factors considerations," in *Proc of the 16th Int IEEE Conf on Intelligent Transportation Systems (ITSC 2013)*, Oct 2013, pp. 2262–2267.

[21] I. Politis, S. Brewster, and F. Pollick, "Language-based multimodal displays for the handover of control in autonomous cars," in *Proc 7th Int Conf on Automotive User Interfaces and Interactive Vehicular Applications, ACM*, 2015, pp. 3–10.