The National Academies Board on Human System Integration (BOHSI) Panel: Explainable AI, System Transparency, and Human Machine Teaming

BOHSI Director: Toby Warden, National Academies

BOHSI Chair: Pascale Carayon

Panel Chair: Emilie M. Roth, Roth Cognitive Engineering

Panelists: Jessie Chen, Senior Research Scientist for Soldier Performance US Army Research Laboratory, William J. Clancey, Florida Institute of Human and Machine Cognition, Mica Endsley, SA Technologies, Robert Hoffman, Florida Institute of Human and Machine Cognition, and Marc L. Steinberg, Office of Naval Research

The National Academies Board on Human Systems Integration (BOHSI) has organized this session exploring the state of the art and research and design frontiers for intelligent systems that support effective human machine teaming. An important element in the success of human machine teaming is the ability of the person on the scene to develop appropriate trust in the automated software (including recognizing when it should not be trusted). Research is being conducted in the Human Factors community and the Artificial Intelligence (AI) community on the characteristics that software need to display in order to foster appropriate trust. For example, there is a DARPA program on Explainable AI (XAI). The Panel brings together prominent researchers from both the Human Factors and AI communities to discuss the current state of the art, challenges and short-falls and ways forward in developing systems that engender appropriate trust.

INTRODUCTION

The National Academies Board on Human Systems Integration (BOHSI) has organized this session exploring the state of the art and research and design frontiers for intelligent systems that support effective human machine teaming. The panel will begin with an introduction by Toby Warden, Board Director, and Pascale Carayon, Chair of BOHSI.

Over the last decade we have seen dramatic successes in the ability of machine learning Artificial Intelligence (AI) software to recognize images, understand and translate speech, play complex games such as Chess and Go at an expert level, and operate vehicles somewhat autonomously. These advances have captured the imagination of the general public and are stimulating a surge of investment by government and industry. Evidence of this is a recent Presidential executive order on maintaining American leadership in AI. The executive order includes a directive to NIST to support development of "reliable, robust, and trustworthy systems that use AI technologies" (Executive Order issued Feb. 11, 2019). Envisioned commercial applications include intelligent digital assistants, robotic agents for healthcare and home applications, and self-driving cars. Envisioned military applications include image and video processing systems for intelligence analysis, ground robots working collaboratively with dismounted infantry, and operators remotely supervising multiple heterogeneous unmanned vehicles.

Among the key factors to the success of these AI applications will be their ability to support the people on the scene in achieving their goals. The effectiveness of these AI software cannot be judged based on their performance operating in isolation, but rather the joint performance of the individuals on the scene working with the support of the AI software. Recently the concept of human machine teaming has been coined to capture this work systems approach. There is consensus that an important element in the success of human machine teaming is the ability of the person on the scene to develop appropriate trust in the automated software (Hoffman, 2017). This includes understanding the conditions under which the software is likely to perform well and the conditions that are likely to be beyond its competence envelope. Some researchers in the human factors community have adopted the term 'calibrated trust' to refer to this need to foster appropriate trust that includes knowing when not to trust the software (e.g., Atkinson, Clancey & Clark, 2014; Lee and See, 2004; Schaefer et. al, 2016).

Research is being conducted in the Human Factors community and the AI community on the characteristics that software systems need to display in order to foster appropriate trust. One characteristic that has been suggested and actively researched in the Human Factors community is the need for AI software to display transparency. The concept of 'transparency' also sometimes called 'observability' as well as 'apparency' is intended as a metaphor to convey the ability of AI software to communicate its actions and plans and the rationale behind them (e.g., its reasoning process, its projection of outcomes and associated uncertainties) so as to foster appropriate trust (Chen, 2018; Woods and Hollnagel, 2006). It is argued that transparent systems enable individuals on the scene to develop an accurate mental model of the software system that allows them to understand not only what it is currently doing, but also why it is doing it and what it will do next (Endsley, 2017).

In the AI community, the recent focus has been on developing 'explainable' software systems. For example, there is a large ongoing DARPA research program on Explainable AI (XAI). The focus of this program is to develop methods that allow machine learning systems (e.g.,

deep learning and neural network systems), which are currently largely "inscrutable" black boxes, to generate explanations that will allow users to understand, appropriately trust and manage the AI software (Hoffman, et al., 2018; Mueller, et al., 2019).

This Panel brings together prominent researchers from both the Human Factors and AI communities to discuss the current state of the art, challenges and short-falls and ways forward in developing systems that engender appropriate trust. Panel members include Jessie Chen, Senior Research Scientist for Soldier Performance US Army Research Laboratory, William J. Clancey, Florida Institute of Human and Machine Cognition, Mica Endsley, SA Technologies, Robert Hoffman, Florida Institute of Human and Machine Cognition, and Marc Steinberg, Office of Naval Research.

Panelists will provide their perspective on the prospects and challenges for design of effective AI systems that promote appropriate trust. Among the questions to be addressed include, Does a computational system need to be explainable to be useful and usable and/or to promote appropriate trust? For example, there have been successful computational systems where the algorithm for generating the solution is opaque but enough context is provided surrounding the proposed solution for the user to be able to evaluate the quality of the solution for themselves (e.g., Roth, et al., 2017; Roth et al., 2018). Similarly does system 'transparency' imply 'explainability'? Related questions include, What kinds of explanations are needed for appropriate trust and at what points in time? For example, explanations of different forms may be more or less useful during system development and debugging by developers, vs. during the training of users on how the system works, vs. during real-time use when faced with a particular decisions as to whether to trust the automated software in a specific situation vs. during a 'post mortem' when trying to understand why the automated software made a wrong decision.

Panel members will also be asked to discuss ways forward to get to the desired end state of effective human machine teaming, and reflections back on 'lessons learned' from earlier waves of AI / automation – that eventually failed to live up to the hype.

JESSIE CHEN

Jessie Chen and her colleagues developed the Situation awareness-based Agent Transparency (SAT) framework, based on Endsley's Situation Awareness model, to identify the information requirements for effective human-agent teaming. Specifically, the SAT framework identifies the required communications from an intelligent agent to its human collaborator in order for the human to obtain effective situation awareness of the agent in its tasking environment. At the *first* SAT level, the agent provides the operator with the basic information about its current state and goals, intentions, and plans. At the *second* level, the agent reveals its reasoning process as well as the constraints/affordances that the agent considers when planning its actions. At the *third* SAT level, the agent provides the operator with information regarding its projection of future states, predicted consequences, likelihood of success/failure, and any uncertainty associated with the aforementioned projections. Chen and her colleagues are currently expanding the SAT framework into human-agent bidirectional transparency to support the agent's planning and performance. The challenge is to design the user interfaces that can support bidirectional transparency dynamically, in real time, while not overwhelming the human with too much information and burden.

WILLIAM J. CLANCEY

In the 1970s and 1980s, providing explanations was an integral design principle for developing medical expert systems, instructional programs, and assistant programs (broadly known as "symbolic AI"; Buchanan & Shortliffe, 1984). Research showed that explanation capability was enhanced by representing processes and subsystems in a modular, abstract way as *computational models* (e.g., in medicine, we distinguish and separately represent a disease taxonomy or causal model, the diagnostic reasoning strategy, and underlying physiological processes; Clancey, 1983; 1989). We thus developed domain-general modeling frameworks, facilitating not only explanation, but "software reuse" (adaptability to new applications) and maintainability. Thus, like alchemists, in the search to create "intelligent machines" we invented a computational method for creating and using scientific and engineering models by which a wide-variety of professional and everyday activities can be partly automated (Clancey, 1985; 1992).

Disappointingly, commercially prevalent software today, such as the common "map navigation" apps on a phone, mostly ignore these well-established and documented modeling methods and explanation principles. Furthermore, these programs increasingly incorporate data analysis algorithms (aka "neural networks") that derive associations people may find difficult to relate to familiar features, causal relations, and narratives by which they understand the world. Addressing the need for explanation, the DARPA Explainable AI (XAI) Program properly emphasizes reconfiguring data analysis algorithms so automated interpretations, plans, advice, etc. are understandable and trustworthy.

Meanwhile, since the late 1980s workplace studies of model-based automation (e.g., autopilot, office workflow; Luff et al., 2002) suggest that people benefit from interactive tools that fit how they think about and do their work. A well-established "work system design" methodology addresses this need by combining participatory design and participant observation to iteratively develop practical tools in the context of use (Greenbaum & Kyng, 1991)—a process that reveals and constrains what kind of explanation and hence what kind of computational methods are appropriate. Besides this contextual approach to tool building, we do well to consider other shortcomings of early "symbolic AI" systems that pose challenges and opportunities for the success of today's "neural networks."

MICA R. ENDSLEY

The ability of operators to successfully interact with automated and autonomous agents to carry out joint goals in complex systems is highly dependent on their ability to understand what the autonomy is doing, what it is projected to do in the near future, and its limitations for successful performance (Endsley, 2017). The need for both an accurate mental model of the automation in general and an accurate situation model of its behavior in real time affects not only operators' level of trust in the system, but also the level of shared situation awareness (SA) which is critical for allowing the automation and the human operator to operate successfully as a team.

Historically, explainability methods were employed to provide improved understandability of logic based AI. These methods sought to tell users how the system arrived at a particular conclusion or recommendation (usually by revealing the rules that had been executed), providing a limited understanding of the inner workings of the system. With the move towards learning algorithms as the favored method underlying today's AI, DARPA is rightly working to expand this approach to better derive logic from inherently opaque AI techniques to fulfill this same function. While there have been some successes at deriving rules from neural networks, for example, explainability approaches generally suffer from being incomplete, non-real-time, and often non-user-centric, with explanations being both vague and overly complex.

On the other hand, research shows that what people really need to interact effectively with automation is to understand in real time what the automation is doing currently and why (e.g. what is its current goal and tasking, what state is it in, what does it think is happening based on its sensors?), what will it do next (e.g. what is it planning to do?), and what are the limits of its performance (e.g. can it handle the present and upcoming operational conditions, or do I need to intervene?). This level of system understanding requires a significant amount of automation transparency.

While explainability and transparency are somewhat complimentary, transparency differs in that (1) it is provided in real-time to support dynamic decision making, (2) it is generally an inherent property of the automation interface displayed to the operator on an ongoing basis, and (3) it encompasses more of the needs of the human operator, providing the SA required to interact with automation to achieve successful oversight and interaction with the autonomy. Systems supporting high levels of SA provide understandability and predictability of the automation, understanding of key states and mode transitions, and understanding of system reliability (e.g. how well it is functioning, its confidence level in fused information, or system assessments), as well as its robustness (meaning its ability to handle current and upcoming situations).

While transparency is important with all automated systems, it will become even more important with learning based AI where the internal system model and capabilities may be constantly evolving and changing. This creates the need not just for system explainability and recurrent operator training, but also the development of system interfaces that provide ongoing, continuous reinforcement and automation understanding through automation transparency.

ROBERT HOFFMAN

This topic triggers a consideration of our terminology. First, machines can never be "team players" in the sense of engaging in genuine collaboration. Machines can in some respects and in some contexts act as if they are cooperating. But when it becomes apparent to "users" that their capacity for negotiating and engaging in common ground is limited by their lack of inferencing capabilities and lack of world knowledge, the teaminess might dissolve. Second, with regard to the interpretability of Deep Nets and Machine Learning systems, "transparency" is perhaps a misuse of a metaphor. (If something cannot be seen, it cannot be understood.) What is needed is machine "apparency." Third, calibration is perhaps an inappropriate metaphor for trust, if only because it attempts to reduce the human to the machine. (This is obviously ironic.) But even more important is that calibration feeds into the view that: (1) trust is a single state, (2) trust develops, and (3) as it develops it converges—or we want it to converge—on some metrical point on a scale of trust. In fact, trusting is a continuous dynamic of multiple relations. It does not develop, it morphs, and it is always manifest as some mixture of justified and unjustified trust and justified and unjustified mistrust in the various functionalities and affordances that a machine possesses (Hoffman, 2018).

The XAI field is characterized by terminology abuse. "Heat maps" are said to show what the machine "pays attention to." Deep Nets are said to "recognize" objects or actions that bear a human semantics. The "interpretability" of explanations means something in computational formalisms that is entirely different from its psychological meaning. Growing sensitivity to the potentially misleading nature of the jargon can only strengthen the XAI enterprise.

The core XAI concept is that an explanation capability could be added onto or into an AI system, enabling it to explain how it works and thereby engender appropriate trust and reliance. Initially, many XAI researchers held a tacit premise, that the property of "being an explanation" is a property of statements, and that statements are spoon fed to the user, who groks them, and then behaves appropriately. The initial concept was that the challenges of explanation could be solved without recourse to user models, knowledge bases, and symbolic inference.

The initial experimentation concept was to run large numbers of Mechanical Turkers in two conditions: Explanation versus No explanation, and show that the explanations helped. As the Phase 1 work proceeded, the approach to experimentation design, and explanation concepts was considerably enriched. AI researchers are now developing more reasonable and interesting experiments, including smaller-scale studies to target particular effects. Researchers are utilizing more refined experimental designs, including an awareness of the need for control conditions. Value is seen in the use of psychometrically-validated judgment scales (see

Hoffman, et al., 2018). On the other hand, there are issues regarding the utilization of parametric statistical testing, such as whether the data even conform to the assumptions of traditional tests. Additionally, the quest for statistical significance drives the desire to simply increase the sample size. Hence, some effects are being found that may be statistically significant but may not be practically significant.

The importance of the users' "mental models" has been acknowledged and some XAI researchers are attempting to reveal and study them. On the other hand, judgment and self-report data are sometimes devalued by the mere mention of the word "subjective, reflecting a belief in the mythical subjective-objective distinction. This engenders a hesitation to actually ask the research participants questions about how they are reasoning. This is especially the case for studies that rely on Mechanical Turk. Researchers can shy away from conducting post hoc interviews on the argument that it is difficult and time consuming to analyze the results.

Most of the explanations that are being machine-generated are local (*Why did it decide that?)* and not global (*How does it work?).* The means through which explanations are generated and provided are limited. However, some XAI researchers are coming to see explaining as a continuous, interactive dialog process. Additionally, some researchers are generating explanation systems that enable users to explore the boundary conditions of the competence envelope of the AI/XAI systems.

MARC STEINBERG

"Teaming" between people and systems containing AI is a popular concept among researchers today, but this term is often used loosely to cover a broad range of collaborative possibilities with different implications relative to the division of human/machine roles, responsibilities, relationships, and functions. This is an important distinction as achieving comparable capabilities to high functioning adult human teams with the addition of AI peer members will likely require solving some of the hardest problems in AI and may not be feasible for decades. However, many currently proposed ideas in teaming will require AI with much more readily achievable capabilities. Thus, it appears likely that we will see a broad spectrum of different types of hybrid human/AI "teams" that will open up increasingly as technology and cultural acceptance advances and that will require appropriate human factors methods, tools, and processes. This will be particularly the case if we imagine a future that sees the fielding of AI methods on large scales and in a great diversity of embedded, perceptive, and persistent devices.

Within the various intelligent systems communities, there has been a substantial increase of interest in incorporating the human element in a positive way. However, there is often a mismatch between what the more algorithm and device focused engineering and computer science communities want, and what can be provided from communities like human factors, psychology, and neuroscience. Some of the most popular things that these communities would like are "turn-key" models of humans that have good predictive power in particular contexts, models of humans "internal states," and actionable algorithm design guidance. However, there often is not a simple answer to these requests. There is a group of human factors researchers in high reliability applications that bring expertise across safety, health, physiology, organizational design, operations in off nominal and degraded conditions, and broad types of performance. However, due to the nature of these types of applications, the use of AI technology has often been conservative and domain specific. On the other hand, the Human Computer Interaction (HCI) community has explored many inventive computational methods and devices, but has focused more narrowly on a subset of human factors issues that is not sufficient for systems involved in more consequential decisions and actions. There is not yet an integrated community that can support tailorable processes and tools to fill the spectrum between the expensive, and difficult to apply ones of high assurance systems, and the more lightweight ones needed in many fast moving applications involving AI.

There is a need for more thinking about these issues in the context of end-to-end systems and with a deeper understanding of the human systems integration issues. First, there has been a long running argument that more human-like AI will inherently provide benefits in this regard. Secondly, there are a number of researchers that have argued for use of particular AI methods that may have some more natural translation into a form that supports understandability and transparency like natural language or visual display. Similarly, less well understood methods could potentially be converted to more easily understandable implementations. There has also been a recent growth of research on systems that attempt to implement more human-understandable plans and behaviors even if the underlying algorithm is somewhat more opaque and difficult to understand. Finally, there is a growing research area approaching these problems at a more meta-cognitive level. This includes both direct explanation mechanisms as well as techniques to assess things like in situ performance and broader proficiency across many similar environments or tasks. Finally, at a more practical level, field implementations have long developed effective decision aids to simulate and predict system behaviors and sensitivities and explore alternative scenarios and counterfactuals that particularly support highly skilled users. The sum total of these methods provides potential avenues for investigation, but there are many open human factors and human systems integration issues that have been explored only at a limited or fairly superficial level to date, and are in need of more multi-disciplinary research.

THE NATIONAL ACADEMIES BOARD ON HUMAN-SYSTEMS INTEGRATION

This panel is organized by the National Academies Board of Human Systems Integration (BOHSI). The National Academies of Sciences, Engineering and Medicine (NASEM) known as the "National Academies" in short, is a private, non-profit organization that is tasked with providing independent and non-partisan advice on all matters related to science, engineering and medicine to Congress and the federal

government since its inception in 1863. The National Academies currently comprises the National Academy of Sciences (NAS), the National Academy of Engineering (NAE) and the National Academy of Medicine (NAM). BOHSI, formerly known as the Committee on Human-Systems Integration and even earlier as the Committee on Human Factors, was established to better equip the National Academies in its efforts to assist the federal government on issues of national policy that involve human factors and human-systems integration. Housed within the Division of Behavioral and Social Sciences and Education, BOHSI is a standing board of the NAS and is sponsored by a coordinated consortium of several federal agencies and other organizations

REFERENCES

Atkinson, DJ, Clancey, WJ, & Clark, MH. 2014. Shared Awareness, Autonomy and Trust in Human-Robot Teamwork. *AAAI Fall Symposium on Artificial Intelligence for Human-Robot Interaction.* Arlington, VA.

Buchanan B.G. and Shortliffe, T.H. 1984. *Rule-based expert systems: The MYCIN experiments of the Heuristic Programming Project*. Reading, MA: Addison Wesley. Available: www.Shortliffe.net.

Chen, J. Y. C. (2018) Human-autonomy teaming in military settings. *Theoretical issues in ergonomics science*. 19, 255-258.

Chen, J. Y. C., Lakhmani, S. G., Stowers, K., Selkowitz, A. R., Wright, J. L., & Barnes, M. J. (2018) Situation awareness-based agent transparency and human-autonomy teaming effectiveness. *Theoretical Issues in Ergonomics Science*, 19, 259-282.

Clancey, W. J. 1983. The epistemology of a rule-based expert system: A framework for explanation. *Artificial Intelligence* 20(3), 215-252.

Clancey, W.J. 1985. Heuristic classification. *Artificial Intelligence* 27:289-350.

Clancey, W. J. 1989. Viewing knowledge bases as qualitative models. *IEEE Expert* 4(2), 9-23.

Clancey, W.J. 1992. Model construction operators. *Artificial Intelligence* 53:1–115.

Endsley, M. R. (2017). From here to Autonomy: Lessons learned from Human-Automation Research. Human Factors, 59, 5-27.

Greenbaum, J. and Kyng, M. (Eds.). 1991. *Design at work: Cooperative design of computer systems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Hoffman, R.R. (2017). A taxonomy of emergent trusting in the human-machine relationship. In P. Smith & R.R. Hoffman (Eds.) (2017). *Cognitive systems engineering: The future for a changing world* (137-164). Boca Raton, FL: Taylor & Francis.

Hoffman, R.R., Mueller, S.T., Klein,. G., & Litman, J. (2018). "Metrics for Explainable AI: Challenges and Prospects." Report on Award No. FA8650-17-2-7711, DARPA XAI Program.

Lee, J. D. and See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, 46, pp. 50–80

Luff, P., Hindmarsh, J., and Heath,C. (Eds.). 2002. *Workplace studies: Recovering work practice and informing system design*. Cambridge: Cambridge University Press.

Mueller, S.T., Hoffman, R.R., Clancey, W, Emrey, A., & Klein, G. (2019). "Explanation in Human-AI Systems: A Literature Meta-Review, Synopsis of Key Ideas and Publications, and Bibliography for Explainable AI." Report on Award No. FA8650-17-2-7711, DARPA XAI Program.

Roth, E. M., DePass, B., Harter, J., Scott, R. and Wampler, J. (2018). Beyond levels of automation: Developing more detailed guidance for Human Automation Interface Design. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 62, 150-154.

Roth, E. M., DePass, B., Scott, R., Truxler, R., Smith, S. and Wampler, J. (2017). Designing collaborative planning systems: Putting Joint Cognitive Systems Principles to Practice. In P. J. Smith and R. R. Hoffman (Eds). *Cognitive Systems Engineering: The Future for a Changing World*. Boca Raton: Taylor & Francis.

Schaefer, K. E., Chen, J. Y. C., Szalma, J. L. and Hancock, P. A. (2016) A meta-analysis of factors influencing the development of trust in automation: implications for understanding autonomy in future systems, Human Factors, 58, 377-400.

Woods, D. D. & Hollnagel, E. (2006). *Joint Cognitive Systems: Patterns in cognitive systems engineering*: Boca Raton, FL: Taylor & Francis.