Guest Editorial: Agent and System Transparency

WITH increasing capabilities of computers and algorithms to act automatically or autonomously, i.e., without the direct involvement of a human, and to act artificially intelligent, transparency of those systems and machine agents become one of the most critical system qualities. What might sound intuitively understandable at the first glance to most of us can become quite complex and non-intuitive with a more scientific and engineering perspective: What is agent and system transparency, scientifically? If machines become more autonomous and more intelligent, why is transparency needed at all; is this not a contradiction? Paradoxically, higher decision authority allocated to intelligent machines tends to result in greater human need/desire for transparency of the system. Yet, how much transparency do we need? Is maximum transparency really desirable? How is system transparency connected to other more established human factors concepts like situation awareness, workload, trust or cooperatively? How can transparency be defined, assessed and operationalized? How can system transparency be created and maintained? How much transparency needs/should be to be maintained over the course of a machine's lifecycle?

Before looking deeper into this special issue and human-machine transparency, let's consider the fundamental meaning of the word and the concept behind transparency. "Transparency" is generally used in the sense of "the characteristic of being easy to see through" (Cambridge Dictionary). If we do not restrict our understanding of transparency to machine agents or human-machine systems that we can literally see through (to moving parts), it immediately becomes clear that human-machine transparency is not a physical attribute that can be measured directly, but a metaphor (Greek, Meta = more highly organized, Phor = bear), which mentally transfers meaning from one thing (source) to another thing (target). It looks like the source is an experience in everyday life, which we are able to see through some material, while other, non-transparent materials would block our view. The targets of the transparency metaphor are human-machine systems. What is the "material" that can be transparent or non-transparent? Upon first consideration, machines are the "material" that should be transparent, such that our human mind could "see through" them. However, the situation could also be the other way around: the term "agent" refers to a thing in the environment (technical or biological). Consequently, humans can be transparent to machines as well as machines to humans. Agent transparency could work in both directions, and even if most of the articles, in this special issue, are mainly addressing the transparency of the machine to the human, we should keep this other direction of transparency in mind. These notions of bidirectional transparency have been noted in prior papers, yet a few studies have examined this space.

Continuing with examination of the transparency metaphor: What does "seeing through" mean? Is it the details of the machines, which we would like to be able to see with better human-machine transparency? If we look into other domains like politics or law, transparency is also used as a metaphor for an ability to access information, if necessary (e.g., Wikipedia), but also to understand the information in a way that actions can be understood or even forecasted. It looks like also for agent and human-machine system transparency, accessing information, understanding the meaning of it, and projecting it into the future, is at the very core of the concept of transparency. This connects transparency directly to situation awareness. The situation awareness-based agent transparency model (SAT) was created with this intent in mind.

If we follow the metaphoric translation of transparency from "see through" to "perceive and understand," it also becomes clear that perception and understanding has some purpose in agent behavior, specifically pre-stages to changing a situation or environment through action! If transparency is necessary for understandability and situation awareness, it becomes transparent that sufficient transparency is mandatory for proper action to influence or control a situation. Here, it becomes obvious that transparency is an important factor for controllability or meaningful human control, which is heavily debated right now in the realm of so called autonomous systems, e.g., in the car or military domain.

Before we consider human-machine related aspects of transparency, let's use the metaphoric analysis of the concept of transparency to directly address another question: Is maximum transparency really beneficial? Consider, for example, that biological agents (creatures in the biosphere) use transparency to their advantage to deceive or to hide their shape by revealing distracting details of the environment or internal structures (see Fig. 1). Other creatures of the biosphere are able to adjust translucency under different lighting conditions; that is, the desirable amount of transparency might be variable, and depend heavily on the situation.

Transferred to human-machine transparency, this could mean that agent and system transparency might be situation dependent, adaptable and adaptive, to allow the right amount of SA depending on user and situation. Unlike the biological agents discussed earlier, intelligent machines lack a natural (or even an intuitive) affordance to enable transparency. So, how do we "see through" machines to perceive, understand, and anticipate their actions and algorithms? This requires thoughtful deliberation and design based on what we "should see"—what is "most beneficial to see" rather than the gamut of "everything that is

Digital Object Identifier 10.1109/THMS.2020.2988835

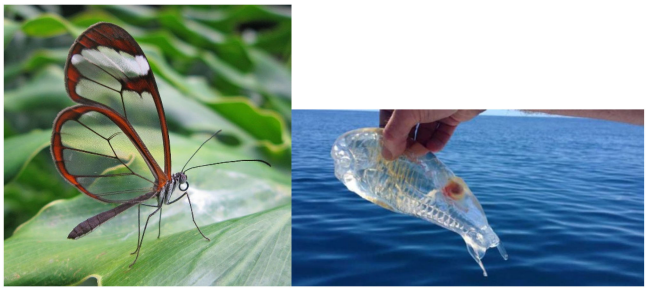


Fig. 1. Is maximum transparency always desirable? Example glass wing butterfly and sea salp (https://pulptastic.com/22-species-see-wildlife-really-letting-hang/).

possible to see," lest we equip operators with an overwhelming barrage of machine cues and information that detracts from the overall human-machine system purpose. So, how do we know what we "should" design for and what is most "beneficial"; this is overarching premise of this special issue and current/future research it may motivate.

This special issue contains eight papers and the concept of transparency is investigated in a variety of contexts of human-agent interaction—from a single robot to multiple heterogeneous agents and swarms. Two studies examine the effects of individual and cultural differences and, based on the compelling results, provide design recommendations related to transparent interfaces. One of the papers provides a thorough review on theoretical aspects of agent transparency and empirical findings on operator performance, situation awareness, trust and workload, among other outcomes. These measures are also among some of the most common metrics reported in studies in this special issue.

In item 1) of the Appendix], Nam *et al.* examine trust in the context of human-swarm interaction with operator control characterized by disparate levels of automation (LOA): manual control, mixed-initiative, or fully-autonomous modes. Trust in this domain is challenging for humans as swarm performance is often opaque. The authors created a series of models to better understand and represent trust behaviors of operators through Markov decision processes and inverse reinforcement learning across the range of LOAs. Results from empirical studies demonstrated that operator trust was often driven by the shape/appearance of the swarm rather than performance during a foraging task. The modeling methods were effective in capturing trust behaviors and support prediction of operator trust to facilitate calibrated trust of swarms.

In item 2) of the Appendix], Chien *et al.* investigate the transparency issue from a cross-cultural comparative perspective, with research participants recruited from three countries to represent distinct cultures based on the Cultural Syndromes Theory: United States (Dignity), Taiwan (Face), and Turkey (Honor). The participants' task was to manage a team of unmanned aerial vehicles in collaboration with a planning agent with different degrees of automation. The experimental results reveal the impact of agent transparency on system usage and a significant role of cultural differences in operator trust when interacting with transparent or opaque agents. These results suggest that when transitioning western-developed automation technologies to countries with different cultures (e.g., Face or Honor), cultural differences in automation reliance related to system transparency need to be incorporated into system designs.

In item 3) of the Appendix], Bhaskara *et al.* provide a review of the transparency literature. The authors present a deep dive into two theoretical transparency models: Lyons transparency model for human-robot interaction and Chen and colleagues SAT model. Then, the authors report on the empirical literature as it relates to transparency finding that:

1) most studies have used the SAT transparency model as a theoretical referent;
2) most use different scenarios and incorporate different testbeds;
3) most use different transparency manipulations in the studies.

The empirical studies reviewed suggest that transparency methods can have a positive impact on performance, trust, usability, and situation awareness; however, the authors note that the nuances of how transparency features are implemented within tasks can cause mixed effects for higher levels of transparency. The authors close with a call for future research to examine the tradeoffs between different transparency features and for research focused on identification of the mechanisms that drive transparency benefits in task contexts.

In item 4) of the Appendix], the transparency issue is examined in a military multi-agent manned-unmanned teaming context, which involves a helicopter crew working with a workload-adaptive cognitive agent to manage multiple unmanned aerial vehicles. The transparent interface design strategies were based on the SAT framework, and a simulation-based human factors experiment was conducted to test the effects of levels of transparency and automation. Experimental results were consistent with findings reported in item 3) that transparent interfaces can support operator situation awareness and performance; although, the subjective trust and workload results were less conclusive. However, there was some evidence that a transparent agent is perceived as more human-like than a more opaque agent, at least in certain tasking situations such as mission management.

In item 5) of the Appendix], Matthews *et al.* investigate the effects of individual differences in attitudes toward robots on human mental models of autonomy and their implications for transparent interface designs. Participants' attitudes toward robots were measured by the perfect automation schema and the negative attitudes toward robots scales; their mental models of robots were assessed in 20 scenario-based queries (in threat assessment contexts) in which they were asked the extent to which the robot's analysis was physics- or psychology-based and how confident they were in the robot's analysis and recommendation. Based on the results, the authors suggest that transparency information should be contextual and personalized in order to minimize human biases (e.g., unreasonable expectations of robot capability or negative attitudes toward humanlike

robots) and to support optimal operator trust calibration and situation awareness. Specific design suggestions to support interface transparency are also provided.

In item 6) of the Appendix], the transparency issue again is examined in a military human-autonomy teaming context, in which an intelligent agent assists the human operator in managing multiple heterogeneous unmanned vehicles in a base-defense scenario. Similar to item 4), transparent interface elements were designed based on the SAT framework, and the effects of transparency on operator performance were assessed in a simulation-based human factors experiment. The results show the benefits of transparent interfaces for supporting operator performance without increasing workload (although participants' response time did increase by a few seconds in the highly transparent condition). With transparency information, participants were able to calibrate their trust in the agent's recommendations more appropriately (e.g., rejecting incorrect recommendations) than in opaque conditions. Participants' subjective trust in the agent also increased as agent's transparency level increased except when uncertainty information was presented.

In item 7) of the Appendix], Wright *et al.* examined the role of transparency and reliability (in particular, low reliability) on user task performance (in a secondary task), workload, SA, SA confidence, and trust of an autonomous robotic squad member (ASM). The study leveraged the SAT model in a simulated training environment inclusive of an ASM wherein participants were tasked with target identification. Neither the transparency nor the reliability of the ASM influenced performance, workload, or SA; however, reliability influenced trust perceptions with ASM errors having a lasting negative effect on trust. Errors also influenced SA confidence, with errors being associated with less SA confidence. This study showed that agent transparency may not have a strong influence on task performance, workload, and SA when the human and agent's tasks are not interdependent, rather task interdependency may be an important moderator for future studies on transparency.

In item 8) of the Appendix], Vered *et al.* propose a transparency framework—similar to the SAT model and based on Endsley's situation awareness model—that consists of four levels of transparency but only focuses on the reasoning process of the agent. A simulation-based study is conducted to examine two delivery mechanisms of transparency information: sequential versus demand driven. The experimental paradigm and scenarios, similar to item 6), involved a human interacting with an intelligent planning agent to manage multiple heterogeneous unmanned vehicles for military surveillance purposes. The experimental results indicate that the demand-driven interface, compared with the sequential interface, is more effective for supporting operator performance (particularly speed) and trust in the agent. Based on the results, the authors suggest that interactive transparent interfaces appear to be a promising design strategy for human-autonomy teaming.

In summary, this issue provides some empirical insights toward addressing the questions we posed at the beginning of this editorial, particularly in the context of human-robot interaction. Furthermore, some of the studies have attempted to make theoretical advances in terms of modeling the concept of transparency and its impact on human performance outcomes.

Acknowledgment

The guest editors would like to thank all the authors and reviewers for their efforts in the review process, and also the editor-in-chief, Prof. D. Kaber, and Ms. M. Lau for their guidance and support.

Jessie Y. C. Chen, *Guest Editor*
U.S. Army Combat Capabilities
Development Command
Army Research Laboratory
Aberdeen Proving Ground,
MD 21005 USA

Frank Ole Flemisch, *Guest Editor*
Fraunhofer FKIE Wachtberg/Bonn
RWTH Aachen University
52062 Aachen, Germany

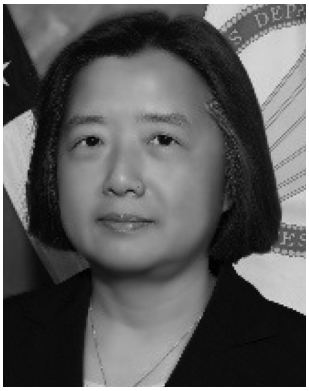
Joseph B. Lyons, *Guest Editor*
U.S. Air Force Research Laboratory
Wright-Patterson AFB, OH, 45433 USA

Mark A. Neerincx, *Guest Editor*
Delft University of Technology
TNO Netherlands Organization of
Applied Research
3769 DE Soesterberg, The Netherlands

Appendix
Related Work

1) C. Nam, P. Walker, H. Li, M. Lewis, and K. Sycara, "Models of trust in human control of swarms with varied levels of autonomy," *IEEE Trans. Human-Mach. Syst.,* to be published.
2) S.-Y. Chien, M. Lewis, K. Sycara, A. Kumru, and J.-S. Liu, "Influence of culture, transparency, trust, and degree of automation on automation use," *IEEE Trans. Human-Mach. Syst.,* to be published.
3) A. Bhaskara, M. Skinner, and S. Loft, "Agent transparency: A review of current theory and evidence," *IEEE Trans. Human-Mach. Syst.,* to be published.
4) G. Roth, A. Schulte, F. Schmitt, and Y. Brand, "Transparency for a workload-adaptive cognitive agent in a manned-unmanned-teaming application," *IEEE Trans. Human-Mach. Syst.,* to be published.
5) G. Matthews, J. Lin, A. R. Panganiban, and M. D. Long, "Individual differences in trust in autonomous robots: Implications for transparency," *IEEE Trans. Human-Mach. Syst.,* to be published.

6) K. Stowers, N. Kasdaglis, M. A. Rupp, O. B. Newton, J. Y. C. Chen, and M. J. Barnes, "The impact of agent transparency on human performance," *IEEE Trans. Human-Mach. Syst.,* to be published.
7) J. L. Wright, J. Y. C. Chen, and S. G. Lakhmani, "Agent transparency and reliability in human-robot interaction: The influence on user confidence and perceived reliability," *IEEE Trans. Human-Mach. Syst.,* to be published.
8) M. Vered, P. Howe, T. Miller, L. Sonenberg, and E. Velloso, "Demand-driven transparency for monitoring intelligent agents," *IEEE Trans. Human-Mach. Syst.,* to be published.



Jessie Y. C. Chen received the B.A. degree in linguistics from National Tsing-Hua University, Hsinchu, Taiwan, in 1987, the M.A. degree in communication studies from the University of Michigan, Ann Arbor, MI, USA, in 1989, and the Ph.D. degree in applied experimental and human factors psychology from the University of Central Florida, Orlando, FL, USA, in 2000.

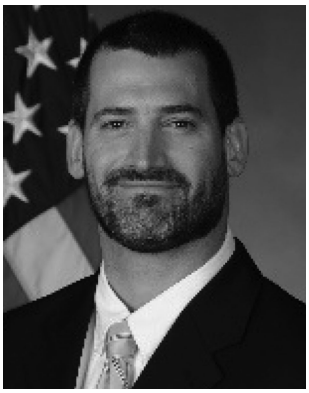
She is a Senior Research Psychologist (ST) for Soldier Performance in Socio-Technical Systems with U.S. Army Research Laboratory. Her research interests include human–agent teaming, agent transparency, human-robot interaction, and human supervisory control.

Dr. Chen is an Associate Editor for the IEEE TRANSACTIONS ON HUMAN-MACHINE SYSTEMS and IEEE ROBOTICS AND AUTOMATION—LETTERS, and she Guest Edited a special issue on "Human-Autonomy Teaming" for *Theoretical Issues in Ergonomics Science,* published in 2018.

Frank Ole Flemisch received the Dipl.-Ing degree in aerospace engineering with a specialization in system dynamics and the Dr.-Ing. degree in human factors from the University of Armed Forces, Munich, Germany, in 1989 and 2000, respectively.

He started as an Aerospace Engineer with a specialization in systems engineering and system dynamics. He spent many years in research on assistant systems and automation of aircraft, cars, trucks, helicopters, industrial sites and weapons systems with the University of Armed Forces Munich, NASA Langley and DLR, and served as the lead of a national standardization group and a technical expert in ISO TC204. He and his team, together with partners from academia and industry, coined the terms highly automated driving, cooperative automation and cooperative control. Since 2011, he has been leading the Department of Human System Integration, Fraunhofer FKIE Institute, Wachtberg/Bonn, Germany. He is a Professor for Human Systems Integration with the RWTH Aachen University—one of two Germany's Engineering Universities of Excellence, and he is member of the NATO-STO Human Factors and Medicine Panel.

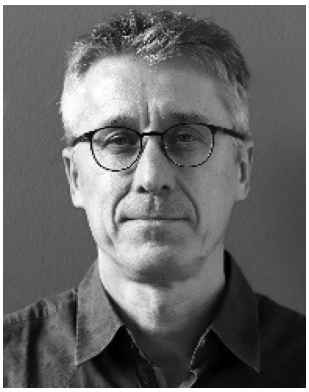
Prof. Flemisch is an Associate Editor for the IEEE TRANSACTIONS ON HUMAN-MACHINE SYSTEMS and he Guest Edited a special issue on "Shared and cooperative control of safety critical systems" for *Cognition, Technology and Work*, published in 2019.



Joseph B. Lyons received the B.A. degree in psychology from Bowling Green State University, Bowling Green, OH, USA, in 2000, the M.S. degree in industrial/organizational psychology from Wright State University, Fairborn, OH, USA, in 2003, and the Ph.D. degree in industrial/organizational psychology from Wright State University, Fairborn, OH, USA, in 2005.

He is currently a Principal Research Psychologist for the Air Force Research Laboratory, and between 2011 and 2013, he chartered the Trust and Influence Portfolio with the Air Force Office of Scientific Research. He is a fellow of the American Psychological Association and the Society for Military Psychologists. His research interests include trust in autonomy, human-autonomy teaming, and interpersonal trust.

Dr. Lyons serves as an Associate Editor for the *Military Psychology*, a Section Editor for *The Military Psychologist*, and has served as a Guest Editor for the IEEE TRANSACTIONS ON HUMAN-MACHINE SYSTEMS and *Frontiers in Psychology*.



Mark A. Neerincx received the M.Sc. degree in cognitive psychology from the University of Leiden, The Netherlands, in 1987, and the Ph.D. degree in cognitive psychology from the University of Groningen, The Netherlands, in 1995.

He is a Full Professor of human-centered computing with the Delft University of Technology and a Principal Scientist of Perceptual and Cognitive Systems with TNO (The Netherlands organization of applied research). His recent projects focus on the socio-cognitive engineering of artificial, virtual or physical, agents (ePartners) that show social, cognitive and affective behaviors to enhance performance, resilience, health and/or wellbeing. His ePartner prototypes are being developed for sharing situation awareness, harmonizing workload distributions and supporting stress-coping in high risk domains (e.g., robot-assisted disaster response teams), and for continual learning and behavior support (e.g., robotic partners for diabetes self-management or elderly care).