Transparency over Time: The Importance of Pre- and Post-Mission Interactions in Human Autonomy Teaming

Christopher A. Miller
Smart Information Flow Technologies, LLC
319 First Ave N., Suite 400
Minneapolis, MN 55401
UNITED STATES

cmiller@sift.net

ABSTRACT

In this paper, I briefly review recent work on the concept of "transparency" in Human Autonomy Teaming (HAT) and, while agreeing that with findings that it is a desirable property for human interaction with complex autonomous systems, nevertheless I point out a fundamental problem: that the presentation of transparent information at the time of its use in the system is at odds with the workload savings that is, frequently, the goal of including autonomy in the first place. A potential solution to this problem stems from the opportunity to "displace" the conveying of transparent information from the time of use to periods (frequently of lower workload) before and after execution time. I explore this opportunity and draw connections to the desirability of pre-mission planning and training and post-mission explanation and debriefing in human-human teaming, as well as the role of such transparency displacement in situation awareness and trust tuning. Mental Model Synchronization or Reconciliation is suggested as a mechanism by which benefits from displaced transparency might achieve these benefits. Finally, an experimental paradigm is suggested for exploring displaced transparency along with some specific predictions for its benefits.

1.0 INTRODUCTION

As Human-Autonomy Teaming (HAT) becomes more pervasive and complex, the nature and extent of that 'teaming' must be examined more closely. Two or more agents teaming to perform a function necessitate some communication and cooperation to minimize conflict and maximize joint performance. This is the "team work" which adds to "task work" (the work which directly accomplishes desired functions) and increases workload for the agents, though in effective teaming the task work accomplished exceeds the added workload from team work needs.

Trying to perform this added team work—making communication, behaviours, plans and their associated rationales "transparent"— in the moment of task execution seems counterproductive. After all, this "moment of execution" is the time frame of most critical performance constraints. Human performance limitations are, frequently, what drives the inclusion of additional team members (human or autonomous). Yet our focus in developing transparent automation systems has generally been on conveying information during execution, via user interface (e.g., a ground control station) which enables real or near real time awareness of the autonomous system and control inputs to it. While a worthy goal, transparency "in the moment" will inevitably run up against—and may exacerbate—the human workload and attentional constraints that motivated the development and inclusion of the autonomy in the first place. In short, the transparency required to enable HAT may not be possible in the same time frame where the team must perform its functions.

In this paper, I will examine this problem and potential solutions to it from Human-Human teaming approaches. These solutions centre around the concept of "displacing" the conveying of transparent information over time. Perhaps somewhat unexpectedly, this points to the link close linkage between a variety of concepts in Human Factors and Human-Autonomy Teaming: Trust, Mental Model sharing, Team Cohesion as well as the utility of training, after action reviews, explanation and overall Situation Awareness.

2.0 THE PROBLEM OF "TRANSPARENCY"

Autonomous systems, largely by definition, operate at times and in contexts in which the human supervisor/controller is not actively monitoring or issuing control inputs. This may be because inputs are impossible (such as in loss of communications due to jamming or weather or deep space communication time lags) or because humans are not able, willing or expected to intervene (e.g., due to workload constraints, inferior performance or simply to expectations of autonomous functioning). Yet humans still want such systems to behave for our benefit and within the instructions we provide. This puts the human operator in the role of a supervisor or, when further removed, a planner or even a designer

This displacement of control means that behaviour shaping inputs and interactions will, almost inevitably, be displaced relative to action execution. The displacement may well be geographical, but it will always be temporal. We necessarily task automation via commands that precede action—though the duration of the gap may range from micro-seconds to years. As automation becomes more complex and as humans are further temporally displaced from the immediate context of the automation's behaviour, however, such tasking will increasingly be via abstract guidelines, plans, and policies well before actions are taken and decisions are made, and we will be reviewing those actions and decisions (and, ideally, explanations for them) after they have been completed and their effects are known. That is, the conveying of information and the issuing of commands becomes increasingly "displaced" in time and perhaps geography from the locus of action.

This emerging situation with more complex and autonomous systems demands that we re-examined some of what we know about human machine teaming for "displaced interactions". We must ask questions such as the effect of varying degrees of temporal and geographic displacement on maintaining Situation Awareness (and what that means when communications are not possible or expected), what skills or personal traits are needed to accurately, effectively and safely interact with automated systems over increasing degrees of displacement, and what kinds of decision aids and training may make such interactions easier and more effective.

We have begun thinking about one such "displacement effect"—the role of transparency in displaced interactions. "Transparency", here, refers to the ability for the automation to be inspectable or viewable in the sense that its mechanisms and rationale can be readily known. Researchers since at least Charles Billings [1] and Sarter and Woods, [2, 3] have called for greater transparency in automation. Furthermore, recent research has

[1] Billings, C.: Aviation Automation: The Search for a Human-Centered Approach. Erlbaum; Mahwah, NJ (1997)

[2] Sarter, N. B., & Woods, D. D.: How in the world did we ever get into that mode? Mode er-ror and awareness in supervisory control. Human Factors, 37(1), pp. 5-19 (1995)

[3] Sarter, N. B., Woods, D. D., & Billings, C. E: Automation surprises. In: Handbook of hu-man factors and ergonomics, 2, pp. 1926-1943 (1997)

generally confirmed [4,5,6,7] that such transparency yields better human situation awareness, trust and, frequently, better overall human-machine performance than systems which include less or no transparency.

But there is a problem inherent in transparency, especially in displaced interactions, that begs for an answer to the displacement question. As noted previously [8] not all information about what an "autonomous" agent is doing and why can be shared if there is to be any workload savings in a multi-agent team. Indeed, for the human to attend to and process any information about what an "autonomous" agent is doing will come at the expense of the human's attention devoted to perceiving and understanding other aspects of their world and performing actions within it. We should expect this to produce additional workload and potential loss of situation awareness of other aspects of the work context if attention is oversubscribed. This will, of course, be particularly problematic in high tempo operations and critical response conditions. Finally, worse still, there are frequent situations in military contexts, especially those involving geographic and temporal displacement of team members, where the communication of information is simply not possible—perhaps due to jamming, distances, terrain and/or the need to remain undetectable.

So, we can conclude that transparency is valuable in human-automation interaction, but information to support it frequently cannot be communicated—at least in the moment when it is most needed--for a variety of reasons. Is there a way out of this dilemma?

3.0 THE BEGINNINGS OF A SOLUTION

This "problem of transparency" is hardly new in human experience; we have encountered it in human-human interactions since we began to collaborate and delegate for complex, organized activities, possibly beginning with pre-human hunting in groups. We may now be running into this problem with human-machine interactions precisely because machines are beginning to be complex and sophisticated enough that we are willing to use them in significant autonomous roles.

Humans interacting with other humans have encountered and wrestled with this problem throughout time. Human supervisors attempting to increase their capabilities through organizing and administering (human) subordinates have almost identical requirements: the need to maintain awareness of the performance of their subordinates so as to ensure that the supervisor's intent is accomplished as accurately as possible—even when that subordinate must necessarily act at a geographic or temporal distance, with lags or gaps in communication, from the supervisor.

---

[4] Mercado, J. E., Rupp, M. A., Chen, J. Y., Barnes, M. J., Barber, D., & Procci, K.: Intelli-gent agent transparency in human–agent teaming for Multi-UxV management. Human fac-tors, 58(3), pp. 401-415 (2016)

[5] Lyons, J. B., & Havig, P. R.: Transparency in a human-machine context: approaches for fostering shared awareness/intent. In: Proc. International Conference on Virtual, Aug-mented and Mixed Reality, Springer, pp. 181-190 (2014)

[6] Ososky, S., Sanders, T., Jentsch, F., Hancock, P., & Chen, J.: Y.: Determinants of system transparency and its influence on trust in and reliance on unmanned robotic systems. In: SPIE Defense+Security, ISOP, pp. 90840E-90840E (2014)

[7] de Visser, E. J., Cohen, M., Freedy, A., & Parasuraman, R.: A design methodology for trust cue calibration in cognitive agents. In: International Conference on Virtual, Aug-mented and Mixed Reality, Springer, pp. 251-262 (2014)

[8] Miller, C. A.: Delegation and transparency: Coordinating interactions so information ex-change is no surprise. In: International Conference on Virtual, Augmented and Mixed Re-ality, Springer, pp. 191-202. (2014)

As I have argued elsewhere [9], acting through a human subordinate is a process of delegation. Delegation inherently involves the communication of intent with oversight (including inspection and correction) of the subsequent performance of that intent by the subordinate. Furthermore, the communication of intent itself forms an "intentional frame" which can serve to increase situation awareness and decrease communication and cognitive processing demands in the future. Sheridan's original definition of supervisory control [10] included the provision that supervisors had to communicate their intent to subordinates or, in Sheridan's words, to "teach" subordinate automation what it should do.

Not surprisingly, humans have already developed approaches to this mitigating problem, at least in human-human tasking and delegation. These involve what might be termed "temporally shifted" or "displaced" transparency—providing transparent information at times other than the critical time of use. Transparency in human and human-automation work is, roughly, the provision of information about what the independent human or automated system is doing and why, potentially at multiple levels of abstraction. Thus, knowledge about what the system will or might do in various circumstances can be provided in pre-mission interactions which we might call planning or training. Similarly, knowledge about what the system has done and why can be provided post-execution in explanations or post-mission debriefings. In all cases, as with in-the-moment transparency, the goal is the synchronization of mental models about what is/will/should be done during missions.

Human to human communication in and near the moment of execution is an extraordinarily rich tool, especially when it is deployed against the backdrop of common cultural and professional understanding of the domain and its goals and methods, and even moreso when it is augmented by mutually-understood professional jargon. It serves to make the communication of intent from supervisor to subordinate, and status from subordinate to supervisor, extraordinarily efficient and rich.

But there is evidence that high-performing human-human teams, especially in high criticality domains, frequently exhibit *less* (and less explicit) communication than do less well-integrated teams. Entin and Serfaty [11] compared teams trained to adopt terse and implicit communication styles during periods of high operational tempo and stress with a control group that had not received such training and found that performance improved significantly with this training. They also found that multiple measures of team behaviour were significantly affected by the training, including particularly an increase in anticipations (and, presumably, a corresponding reduction in the need to explicitly request information) between teammates). Together, they took these findings as indicating that efficient teams were sharing and making more efficient use of mental models which had the effect of both reducing the need for communications and speeding those communications which were necessary.

This finding implies that part of the reason human natural language communication can be so effective, and a large part of the reason well-trained and experienced teams can be so sparse with their communications, is that such team members share an understanding of the domain and of work within it. This understanding is certainly acquired during training and experience, but on a day by day (or mission by mission) basis, it is also acquired through mechanisms such as training and planning before execution, and explanation and debriefing (or after-action reviews) after execution. Below, I will argue that these mechanisms provide "transparent" information to the human team-members—that is, the same kind of information that has been shown to improve performance

[9] Miller, C. & Parasuraman, R.: Designing for flexible interaction between humans and au-tomation. Human Factors, 49(1), pp. 57-75 (2007)

[10] Sheridan, T.: Supervisory Control. In: Salvendy, G., (ed.), Handbook of Human Factors pp.1244-1268. John Wiley & Sons, New York (1987)

[11] Entin, E., & Serfaty, D.: Adaptive team coordination. Human Factors, 41, pp. 312–325 (1999)

from “transparent” displays. They just do so at times other than the time of execution. That is, they provide temporally (and potentially, geo-graphically) displaced information about how an agent is, will, or should behave, or how and why he/she/it did behave, but do so at a time when workload and attentional demands are lower. In short, they provide ***Displaced Transparency***. Displaced Transparency may serve as a solution to the dilemma of transparent information in HAT contexts.

4.0 TRANSPARENT INFORMATION AND WHY DISPLACING IT WORKS

What is transparent information and why should we believe that it could be transferred before or after the temporal context of use? Similarly, why should we believe that there are frameworks that might mitigate the need for or improve the efficiency of real time transfer of transparent information?

Chen, et al., [12] have defined transparency as “… the descriptive quality of an interface pertaining to its abilities to afford an operator’s comprehension about an intelligent agent’s intent, performance, future plans, and reasoning process.” But the emphasis on “the interface” in the above definition puts a, perhaps undue, focus on information that is conveyed *during* execution—when an interface is typically used. I contend that much information which achieves the goals of transparency (i.e., affording “an operator’s comprehension about an intelligent agent’s intent, performance, future plans and reasoning process”) need not be provided only, or even primarily, at that workload-intensive time. This relationship is illustrated in Figure 1.

Indeed, Chen [12] also defines a scale or model for transparency, the Situation Awareness-based Transparency (SAT) scale, which in turn leverages Endsley’s [13] scale for Situation Awareness. Chen’s SAT levels are summarized below along with a summary of the transparency information intended to support them:

[12] Chen, J. Y., Procci, K., Boyce, M., Wright, J., Garcia, A., & Barnes, M.: Situation awareness-based agent transparency (No. ARL-TR-6905). ARL/HRED Aberdeen Proving Ground, MD (2014)

[13] Endsley, M. R.: Toward a theory of Situation Awareness in dynamic systems. Human Factors, 37(1), pp. 32–64 (1995)

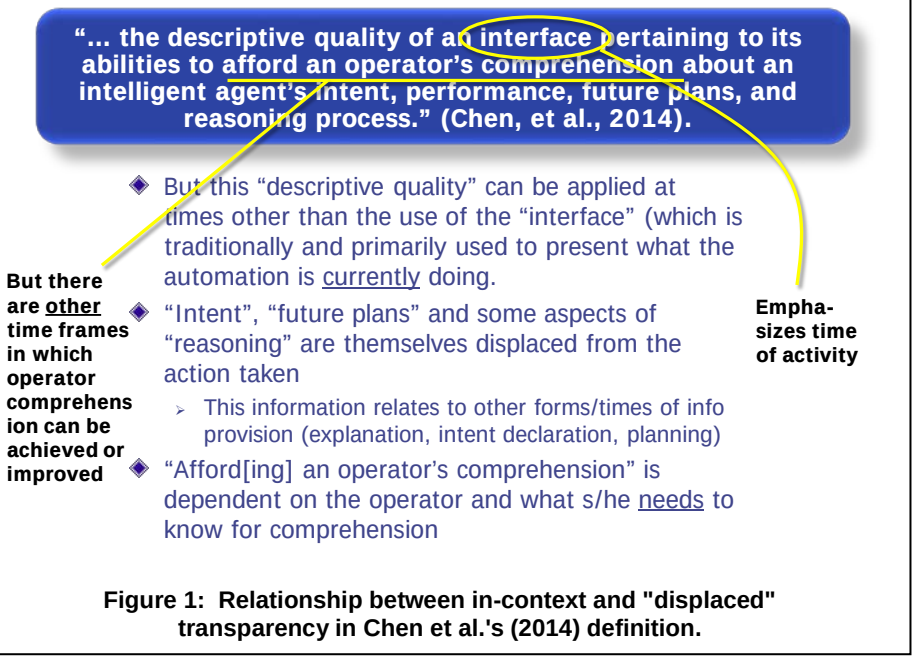

Figure 1: Relationship between in-context and "displaced" transparency in Chen et al.'s (2014) definition.

1. Level 1 SA (What's going on and what the agent is trying to achieve) which is satisfied by providing information about the agent's:
   - Purpose or Desire and Goal selection
   - Process and Intentions (including Planning and Execution) and Progress along that process
   - Performance of both the process and in general.
2. Level 2 SA (Why does the agent do what it does?) which is satisfied by providing information about the agent's:
   - Reasoning process for planning or decision making, including the agent's beliefs and broader purpose, including agent's current beliefs about the environmental and other factors which constrain it
3. Level 3 SA (What should the operator expect to happen?) which is satisfied by information about the agent's:
   - Projection to Future/End state
   - Potential Limitations including likelihood of error and history of performance.

If we take this information content as what is required for effective transparency, then it is worth noting that much of it could be—and probably is in much effective human-human teaming—provided either before or after the time of action execution. "What an agent is trying to achieve?" is something that can and generally should be worked out, at least at a high level, before the subordinate agent is deployed. Intent expressions (which may include goals, purposes, methods and priorities [14]), by their nature, occur before action, while an understanding of why an agent does what it does and what its belief were that might have influenced decisions and actions are

[14] Miller, C.: Delegation for Single Pilot Operation. In: 2014 HCI-Aero ACM, New York (2014)

precisely the focus of the explanations that occur in effective after-action reviews and debriefings [15]. Figure 2 provides a hypothesized annotation of Chen's SAT levels into items which can occur before or after the moment of action.

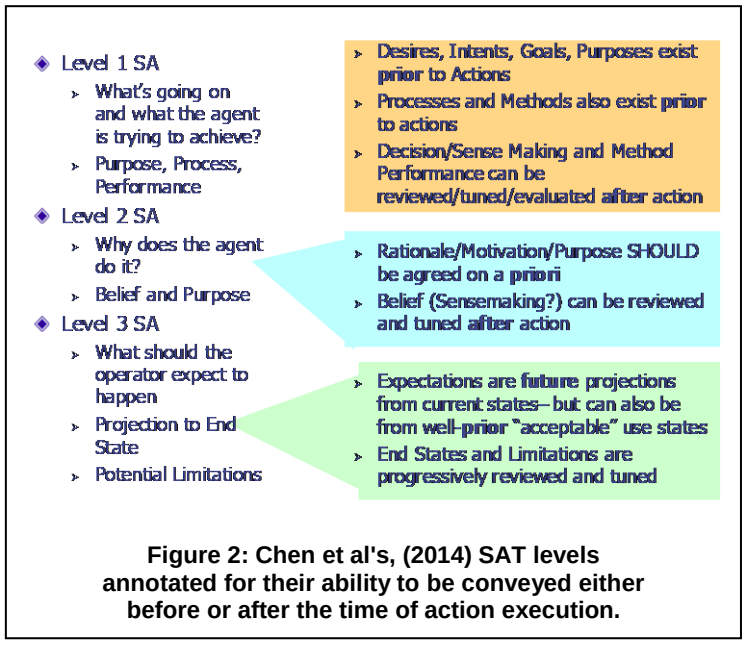

Figure 2: Chen et al's, (2014) SAT levels annotated for their ability to be conveyed either before or after the time of action execution.

We can illustrate the effects and power of pre- and post-mission transparent information transfer with a simple thought experiment. Figure 3 provides a simple timeline for a pre-planned workflow (e.g., for a military reconnaissance mission). Imagine that this timeline is worked out and agreed to by a supervisor and subordinates (e.g., a mission commander and his/her subordinate pilots) before a mission. The plan says that this mission will consist of a reconnaissance task consisting of an Ingress Phase to being at time T and end at time T+1:00 hour, to be followed by a Search Phase to run from the end of the Ingress Phase for another hour. These will then be followed by a decision point whose agreed-upon logic is that if a target has been detected it will be Monitored for another hour and if not, the recon aircraft will Traverse to an egress point.

The simple fact of having made this a prior plan affords substantial situation awareness to the supervisor *even if s/he has no further communication* with the subordinate. For example:

- Given that it is 45 minutes into the mission, the supervisor knows that the subordinate is (supposed to be) engaged in Ingress and even, approximately, where the subordinate is. This is "What's going on?" knowledge—Level 1 SAT.

[15] Tannenbaum, S. I., & Cerasoli, C. P.: Do team and individual debriefs enhance performance? A meta-analysis. Human factors, 55(1), pp. 231-245 (2013)

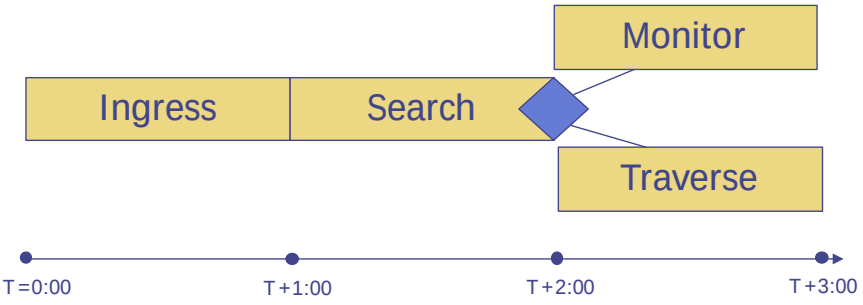

Figure 3: Simple timeline example to illustrate conveyance of "transparent" information before use via planning.

- Furthermore, the supervisor knows why the subordinate is ingressing: to get to the search area and begin search—Level 2 SAT.
- Finally, the supervisor knows that, at time 1:00, the subordinate will transition to Search. This is "What the operator should expect to happen?" knowledge—Level 3 SAT.

Granted, this awareness may be erroneous, if no further communication is available, because the plan has deviated from reality, but the deviations themselves will be probabilistic (i.e., they may not happen) and, if they do happen, they will represent partial deviations from the baseline plan. Note too that having an a priori plan in place makes communication more efficient. Instead of having to report all three levels of SA (what is going on, why, and what to expect next), the subordinate will generally have to report only current status or, perhaps, only deviations. The presence of pre-planned alternatives and acceptable error bounds (as represented partially by the pre-planned decision point in Figure 2) makes it possible to further reduce communication needs. If, anywhere from time 1:00 to 2:00 the subordinate reports finding a target, then the behaviour at the time 2:00 decision point becomes predictable. Even more, if at some time after 2:00 the subordinate reports that it is Monitoring, the superior is entitled to conclude that the reason is because a target was found during Search. Even in the event where the entire plan is made impracticable, having had a shared plan makes communication about future behaviours more efficient by giving all participants a shared ground to build or deviate from. If, for example, a fuel leak makes sustained Monitoring impossible, both supervisor and subordinate will know (at least approximately) where the agent is, how much range is required to get home, that a target has been detected, that the monitoring objective will have to be abandoned, etc. And all of this knowledge can be shared with little or no "in the moment" information exchange or processing.

5.0 TWO TYPES OF DISPLACED TRANSPARENCY

The above example emphasizes the role of pre-mission planning in establishing "displaced transparency", but that is not the only way transparent information can be displaced. After action reviews, debriefings and explanations also provide after-the-fact transparent information as well. Admittedly, this information is not provided in a timely fashion to enable a supervisor to override or correct behaviour which may not be desired for the current mission, but insofar as work with this particular subordinate continues in the future, it does play a mutual learning and trust building/tuning role. By learning how the subordinate thought and behaved in a specific situation, and potentially by offering advice or instruction about how future instances should be handled, the supervisor can shape future behaviour in much the same way that planning shapes behaviour (and information interpretation) before execution. As Tannenbaum and Cerasoli [15] have said, after-action debriefs

are an effective and efficient means of improving team performance and team cohesion. Their meta-analysis of debriefing studies (covering 46 samples and 2,136 individual participants) indicated that on average, debriefs improve effectiveness in future team collaborations and performance over a control group by approximately 25% (d = .67).

Figure 4 illustrates the "displaced transparency" relationship we posit. Most transparency research to date focuses on information which is conveyed in a narrow temporal window around an action or event of interest, perhaps because of the focus on the design of user interfaces to provide "transparent" information. Such research has generally shown improvements in performance, situation awareness and trust when transparency information is provided. But such research has rarely examined workload effects, especially in time critical and overloaded periods and/or has examined it with subjective and coarse-grained tools such as NASA TLX [16]. We posit that there are periods in human-automation interaction, just as there are in human-human interaction, where the human's attention, processing and comprehension capabilities are so sparse and/or over-subscribed, that the attention to transparent information will, at best, be incomplete, and at worst may provide a disastrous distraction.

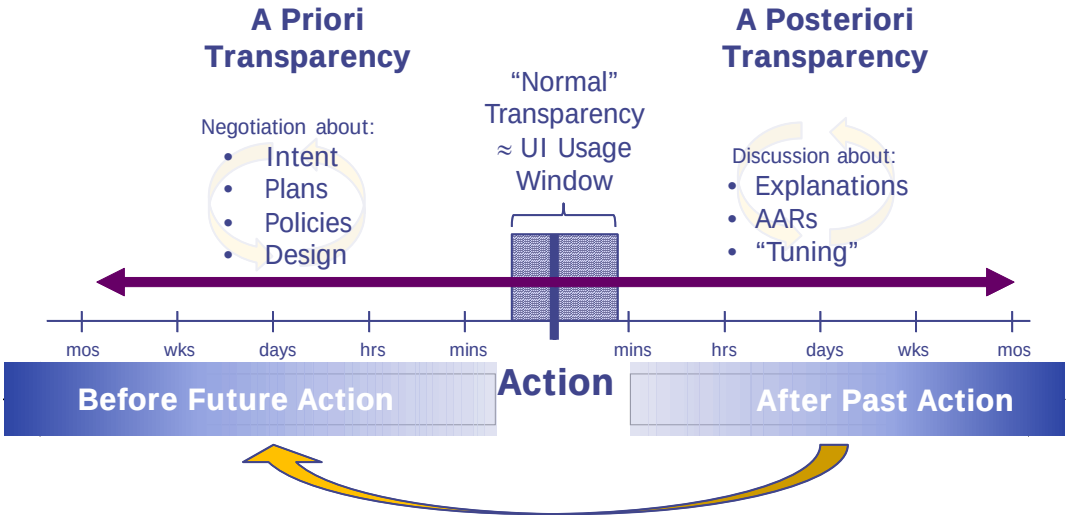

Figure 4. Proposed role of displaced transparency and its effects in team communication and interactions.

We suspect, however, that it will be possible to displace much of this transparent information transfer in time into periods which are substantially less overloaded and/or to provide frameworks and protocols for information transfer which greatly reduce the amount of information which needs to be transferred at critical times in order to achieve the same level of situation awareness. These periods may be anywhere from months or years before an action, or weeks to months after an action. This gives us, conceptually at least, two different periods in which displaced transparency information might be transferred: "***a priori transparency***" (transferring transparent information before the time and context of use) and "***a posteriori transparency***" (transferring it after the context

[16] Hart, S. G., & Staveland, L. E.: Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In: Advances in psychology, 52, pp. 139-183, North-Holland (1988)

of use). In a priori transparency, the processes by which this information transfer occurs are called training, discussion, planning, etc. and they occur before an action and the communication it. The focus of a priori transparency is intent, plans, policies, alternatives and how to decide between them. At extreme durations, the "planning" process becomes one of design of the team, its concept of operations or its equipment. For a posteriori transparency, where discussion occurs after the action, it is called explanation, debriefing, after-action review or various forms of "tuning" (including continuous quality improvement). Iterative cycles through either pre- or post-action discussion will likely only help the understanding of transparent information among team members. It is worth noting, in addition, that the process of pre-action planning and post-action review itself forms a virtuous cycle that can build team understanding for future actions. Although one might reasonably ask how and why providing transparent information *after* an action is helpful, the answer is that it will be helpful in providing team coordination and understanding for future actions, assuming that team must interact again in the future.

This may be why Entin and Serfaty [11] found that verbal communication of explicit information is actually lower among high-performing human-human teams than for less well-integrated teams: because such teams already know what each member will and should do, and trust each other to make good, coordinated decisions. It may also be because communication vocabulary has been streamlined a priori—a benefit we have claimed for playbooks [17]. This is also why Dwight Eisenhower [18] can say "Plans are worthless, but planning is everything"—because planning activities enabled his teams to develop a shared understanding of missions and each other. Explanations and post-mission debriefings work similarly to enable inspection, challenge, tuning and coordination of mental models, though such interactions only have benefit if the team members are able to work together (or work with similarly trained individuals) in the future.

6.0 A MECHANISM FOR DISPLACED TRANSPARENCY EFFECTS

The claim that displacing transparent information can serve much the same purpose as presenting it at the time of need begs the question of how such a process might work. The work of Kambhampati, Chakraborti, Talamadupula and others [19, 20] may provide at least the beginnings of an answer. Although working to provide explanations from machine or robot planners to humans, they nevertheless begin with a human social and cognitive model of the role of explanation itself. They take issue with many past approaches to providing machine explanations as rooted in presenting the machine's reasoning in its own terms—a process they call "soliloquy." [19] They say "Such soliloquy is wholly inadequate in most realistic scenarios where the humans have domain and task models that differ significantly from that used by the AI system." Instead, they say, effective explanation between actors must be a "Model Reconciliation Problem" [19]—that is, a process which is entered into only when a difference between "models" (roughly, mental models) is detected by one or both actors, and which proceeds until this difference is understood and repaired, at least to the level of affording translation between the two models.

---

[17] Miller, C. & Parasuraman, R.: Designing for flexible interaction between humans and au-tomation. Human Factors, 49(1), pp. 57-75 (2007)

[18] Eisenhower, D.D.: From a speech to the National Defense Executive Reserve Conference in Washington, D.C. (November 14, 1957); in Public Papers of the Presidents of the United States, Dwight D. Eisenhower, 1957, National Archives and Records Service, Government Printing Office, p. 818. Downloaded from en.wikiquote.org/wiki/Dwight_D._Eisenhower on March 3 (2018)

[19] Chakraborti, T., Sreedharan, S., Zhang, Y., & Kambhampati, S.: Plan explanations as model reconciliation: Moving beyond explanation as soliloquy. In Proceedings of IJCAI, pp. 156-163 (2017)

[20] Talamadupula, K., Briggs, G., Chakraborti, T., Scheutz, M., & Kambhampati, S.: Coordination in human-robot teams using mental modeling and plan recognition. In: IROS 2014, pp. 2957-2962, IEEE (2014)

Kambhampati has pointed our [21] that except perhaps in testing and evaluation circumstances, explanations are generally neither sought nor provided in circumstances where individuals believe their reasoning and behaviours are compatible or "reconciled". In other words, if I believe that I understand your mental model of the situation sufficiently that you and I would arrive at the same decision I have just arrived at about a course of action to pursue, then no explanation of my decision is necessary for you. It is only when our models diverge (or are believed to diverge) that explanations are invoked—either because you do something which is not compatible with my model of what should be done in the situation, or because you intend to do something that you believe will differ from my decisions or sensemaking in the situation.

Explanation is thus, in their view, largely a process of *Mental Model Reconciliation*. It is provided in order to synchronize the models of those who must work together. It is sought when models do not synch—and it may not be sought when models are assumed to synch, even if they do not.

While explanations can certainly deviate from our actual methods of decision making [22,23], they nevertheless represent how we are trained and acculturated to providing rationalizations for our decision making. Thus, it represents a form of team or group synchronization of thinking in its own right.

This viewpoint explains many phenomena in both explanation and transparency. For example, Entin and Serfaty's [11] finding that well-performing, efficient teams require less, not more, explicit communication can be seen as arising from the fact that such teams are likely to have trained and worked together extensively in the past or (as in their experiment) have been trained to attend to and to anticipate each other's information needs. Thus, their mental models of task and domain are likely to be well-synchronized—meaning that explanatory interactions are less likely to be required, commands can be abbreviated with contextual modifications presumed, and even task-based status information can be abbreviated and given with reduced contextual information because all parties are likely to understand what is needed when.

Similarly, team debriefing after a mission or project, with most of its associated benefits [15] can be attributed to the fact that effective debriefings involve team members interacting about their decisions and their behavioural processes. Simply knowing who knew what when and how they made decisions on the basis of that knowledge serves to educated other team members about their teammates' mental models, while group discussion about how things might be done more effectively in the future creates shared mental models going forward. Group discussion and shared rationalization of events and processes can also serve an educative function, effectively causing the group to converge on a series of broad thought processes and values that, within the "culture" formed by the group, count as valid and reasonable "ways of thinking" (or, at least, of explaining) [24]. Note, though, that this can cut both ways—resulting in adverse cases in "groupthink" [25] and "normalization of deviance" [26]. Finally, explanations offered within a group or to superiors play a social function as well [27],

---

[21] Kambhampati, S.: Personal communication. August 1, Arlington., VA (2017)

[22] Tversky, A., & Kahneman, D.: Judgment under uncertainty: Heuristics and biases. Science, 185(4157), pp. 1124-1131 (1974)

[23] Klein, G.: Naturalistic decision making. Human factors, 50(3), pp. 456-460 (2008)

[24] Miller, C.: Learning to Disagree: Argumentative Reasoning Skill in Development. Ph.D. Thesis. University of Chicago, August (1991).

[25] Turner, M. E., & Pratkanis, A. R.: Twenty-five years of groupthink theory and research: Lessons from the evaluation of a theory. Organizational behavior and human decision processes, 73(2-3), pp. 105-115 (1998)

[26] Vaughan, D.: The Challenger launch decision: Risky technology, culture, and deviance at NASA. University of Chicago Press; Chicago, IL (1997)

serving to reduce social distance and establish or reinforce power structures. While these functions may not directly contribute to a shared mental model about how decisions should be made within the group, they will serve to enhance team cohesion and affiliation.

One sense in which Eisenhower's "planning is everything" [18] may be true is that the activity of planning itself actively promotes model reconciliation. Extensively pouring over plans, rehearsing what could go wrong and what might be needed in contingent situations is a process which conveys and affords opportunities for model synchronization in much the same way that post-mission debriefing does. Participants are likely to emerge from such a process with a richer understanding of each other's models—and having had many opportunities to confront and refine those models in a process which tends toward synchronization. While this, too, can engender groupthink and the suppression of some voices, it also allows team members to "get inside each other's heads" and, thereby, increase their chances of knowing how each other will behave even in unanticipated situations.

Note that I am, in no way, claiming that measuring and computing model reconciliation needs and effects will be easy, especially between humans and machines. Chakraborti and Kambhampati and their colleagues have largely sidestepped this problem by using machine readable symbolic models for "simulated" humans to illustrate efficiency gains in symbolically characterized explanation content. By contrast, humans have evolved cultural, semantic and pragmatic markers rooted in natural language and "body language" for interpreting the need for model reconciliation and then effecting it—a process that, though complex and rich, is far from error proof (cf., [28] for examples of human-human and human-machine model mismatch where communication failed to avoid misinterpretations and, therefore, accidents).

It is worth noting that the process of building and "reconciling" mental models is clearly related to the process of building and tuning "analogic" and "analytic" trust as described by Lee and See [29]. Lee and See define three alternate routes to achieving and tuning trust: affective, analogic and analytic. Affective trust is based on emotion and intuitive pleasure in interacting with the system (or, presumably, the person): we tend to more quickly trust people and experiences that are pleasing to us. Analogic trust is trust based on similarity to individuals or situations we trust or accreditation by individuals or organizations we trust. Analytic trust involves determining how a person or system arrives at its decisions and/or behaviours. It is the slowest and most data-driven of the methods, but also the approach that results in the deepest awareness and most accurate predictions of future attitudes and behaviours. As such, mental model reconciliation is at least one process by which we can learn about (and revise as necessary) the thought processes of another agent, whether human or machine. Initial learning will help to establish whether or not the agent "thinks like it should"—that is, like one with expertise in the relevant domain—and is thus pertinent to accurate tuning of trust through analogic methods. Deeper and more extensive interaction and discussion (through planning, explanation, after action reviews, etc.) will provide the information necessary for analytic trust tuning.

---

[27] Brown, P., & Levinson, S. C.: Politeness: Some universals in language usage, 4. Cambridge University Press; Cambridge, UK (1987)

[28] Miller, C.: Social Relationships and Etiquette with Technical Systems. In B. Withworth and A. de Moor (Eds.): Handbook of Research on Socio-Technical Design and Social Networking Systems, Information Science Reference; Hershey, PA. pp. 472-486 (2009)

[29] Lee, J. and See, K.: Trust in computer technology: Designing for appropriate reliance. Human Factors, 46, pp. 50–80 (2004)

7.0 A SIMPLE ILLUSTRATION

It is reasonably straightforward to show how a process of model reconciliation which occurs either before or after a time frame in which the model is used to make a decision can lead to a reduction in the need to provide "transparency" information in that time frame. Consider again the simple mission sketched in Figure 3, along with an "Observing Teammate" (OT) who, let's assume, can observe and know everything that is happening in the region and to the vehicles involved in the mission, but who is completely unaware of the mission and the intentions of an "Enacting Teammate" (ET). Thus, OT has all and only the "What's going on?" portion of Level 1 SA in Figure 2. Note that this is not even all of the "transparent" information required for Level 1 SAT; the ""What the agent is trying to achieve" portion is missing, as well as essentially all of the information for Level 2 and 3 SAT.

In this scenario, any intent-related interpretation of the world and vehicle states and behaviours from ET will certainly increase OT's SA, especially at levels 2 and 3, but that is because OT's knowledge is essentially nil without such reports. Even observable events in the world (e.g., a headwind) will need to be interpreted for OT in order to provide him/her and understanding of the headwind on "What the agent is trying to achieve" as well as "Why the agent does what it does" and "What will happen next". Without such interpretation from ET, OT's ability to infer this information will be near zero.

OT has none of ET's mental model of the mission (though they may share a model of the world state in this example). All three levels of Chen's SAT [12] must be communicated for OT's full SA. On the other hand, if OT and ET shared the same mental model of the mission and can perceive the same world events, then ET might need to communicate nothing since events would be interpreted similarly and decision making would be identical between ET and OT.

Even when OT can't observe everything that ET can (e.g., in the case of a remote supervisor), the burden of communication is substantially reduced. When events are unfolding as planned, at most ET might need to communicate confirmations that the plan is proceeding as expected. Even when unexpected events occur (e.g., the headwind above), reporting them may be all that is necessary to synchronize ET and OT's models of the impact and revisions necessary and desirable to the mission (revisions in knowledge at Level 3 SA that will *not* have to be communicated since both sides will make them concurrently—though confirmation might still represent useful redundancy). As before, anticipated variations (e.g., the decision point about whether to remain and monitor or simply to traverse) can be communicated much more tersely since both parties know, a priori, what the significance of a detected target will be on this decision point, and/or what the valid reasons are for remaining to monitor vs. traversing.

Even mental model mismatches become easier to detect given this prior planning. Let's say that OT failed to notice that ET detected a target and thus, doesn't understand why ET is transitioning to Monitor rather than Traverse. Simply posing the question in the context of what was expected to be a shared model of the mission identifies the mismatch. "Why are you Monitoring?" conveys a violation of expectations for Monitoring (i.e., the prior detection of a target) and hones in on the piece of information which is needed to repair model mismatch.

This relationship is depicted in Figure 5, illustrating that much less information needs to be conveyed at execution time if full knowledge of plan, purpose and procedures are shared ahead of time. Note that the relative size of the blocks of information associated with Level 1, 2 and 3 SAT are purely speculative. Note too that while it may be true that the higher levels of transparency information may be omitted under conditions of shared a priori transparency information, I do not mean to claim that having this information available is useless.

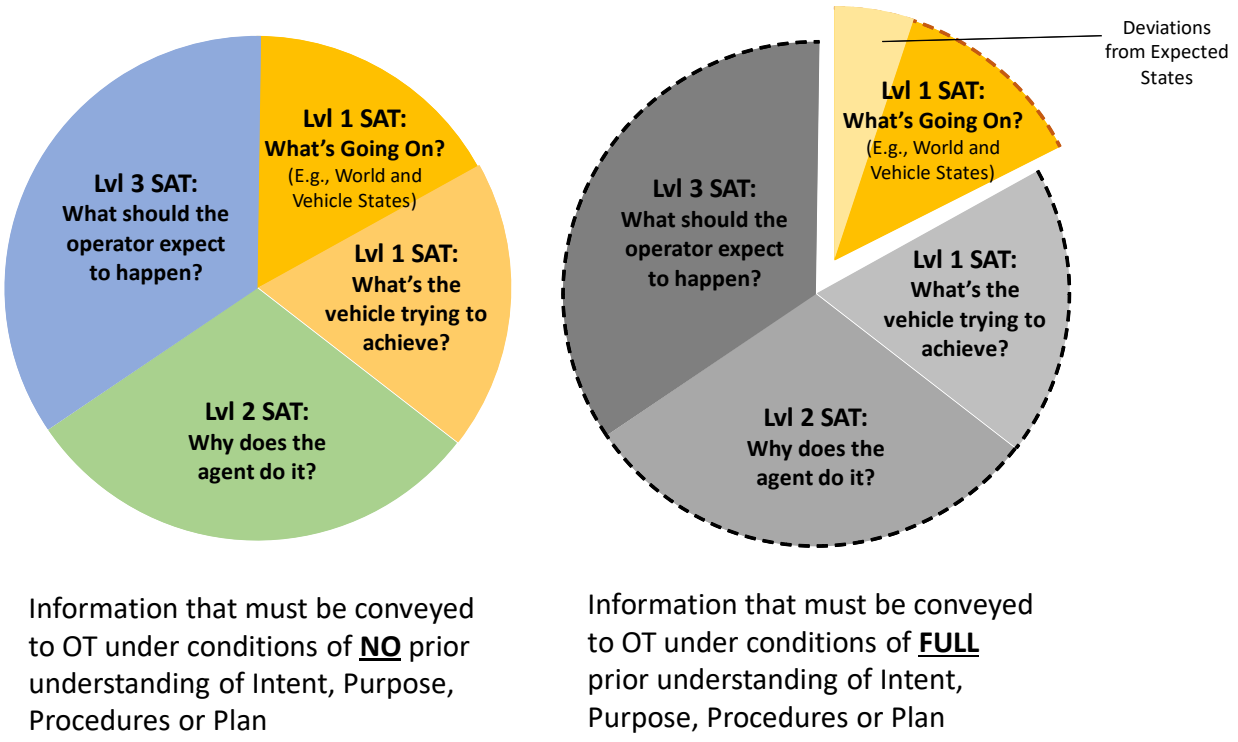

Figure 5. Types of information that need to be conveyed under conditions of no vs. full a priori sharing of Transparency-related information.

Since any sharing of mental models between supervisor and subordinate (even an automated subordinate) will necessarily be subject to uncertainty, continued sharing of higher level information types will help to trap potential mismatches, as well as helping to continue reminding the human members of the team what they *should* be aware of even without reminding.

Finally, although harder to quantify, post-mission debriefings which are later followed by subsequent missions can have similar effects. If, for example, OT learns that ET has a tendency toward speedy completion of missions, s/he might assume a bias or preference in ET for Traversing vs. Monitoring and, in the presence of an ambiguous target detection signal, make more nearly accurate predictions about what ET will do. This represents a variation in the mental model (specifically, in values or priorities) between ET and OT, but insofar as OT understands this about ET (that is, OT's model of ET contains it) it will be accurately factored in to OT's SA and result in accurate understanding of the situation.

8.0 PREDICTIONS AND NEXT STEPS FOR DISPLACED TRANSPARENCY

In this paper, I have presented a fundamental problem in providing "transparent" information in the context of high workload and high complexity. But I have also presented an argument for the desirability and effects of displacing the presentation of transparency information into a priori mission planning interactions and a posteriori explanations and debriefings, along with a hypothesized mechanism for why these effects might be

obtained. We know, from the sources cited above and others, that transparency information frequently provides detectable mission performance benefits. We also know (again from sources sited) that prior mission planning and explanation and debriefings also provide benefits for team cohesion, team satisfaction and team performance. It seems likely that these benefits obtain because they are making use of the same underlying mechanism: the communication of information which promotes situation awareness through mental model synchronization and reconciliation at the time of use. The fact that this information doesn't have to all be transmitted at the time of use, but instead can be spread into lower workload periods before and after usage, is a feature we should use more extensively in design—for human-human and also for human-machine interactions.

Some simple, testable predictions from this model are provided below. While we have not yet been able to conduct experiments to validate these predictions, a simple laboratory test seems eminently plausible. A relevant yet simple scenario is sketched in [20] where a human and a robot are located in a building with a long corridor and multiple side rooms. The human tasks the robot to fetch "a med kit"—one or more of which may be located in the side room(s). Which med kit is desired, expected, and provided is a function of elements of context (e.g., where robot, human, med kit(s) and other potential humans using med kits are located) as well as the robot's decision making algorithm. Of course, mismatches in elements of mental models (e.g., awareness of the physical context and of the robot's decision making process) are exactly what is required to provide SAT knowledge—and can be manipulated in an experimental design. The human may expect the robot to go to a different side room if s/he erroneously believes the med kit to be located there, or the robot may take longer and travel further than necessary if it has an erroneous model of where the human will be located. In this or a similar paradigm, we would predict:

With mental model synchronization between teammates, reduced time, workload effort and even communication bandwidth will be necessary to achieve a similar level of situation awareness compared to conditions without mental model synchronization.

Shifting the communication of transparent information into other time frames (before or after execution) will yield improved situation awareness (with reduced workload) even under conditions of communications restriction or constrained workload for the human recipient, given that model synchronization makes that information comprehensible.

A priori mental model reconciliation will produce more accurate inferences by team members even in unanticipated situations, even with little or no explicit communication of transparent information.

Particularly with regard to post-mission debriefing and explanations, effects of a posteriori model reconciliation will produce increased awareness and ability to predict teammates behaviour even in unanticipated situations in subsequent missions.

We note with interest that [20] reports that the inclusion of mental modeling capabilities in the reasoning of a robot agent, where the robot was modeling the expected reasoning of a human operator and reacting accordingly, produced a 44-75% improvement in robot decision making in terms of avoiding resulting resource conflicts in one analytic experiment they performed.

What is less well documented is the tradeoffs involved in shifting transparency information into time frames before and after it is needed. Somewhere between "no plan survives first contact with the enemy" (implying that "overplanning" is wasteful) and "Plans are worthless but planning is everything" (implying that planning activities are very valuable), there must lie a (probably context-dependent) happy medium. Where is that medium, and what parameters characterize it? It is likely that information theoretic models can provide us with

boundary conditions for this claim, but their relationship to actual human-human (or human-machine) interaction remains to be determined.

Finally, as automation becomes more capable, omnipresent and more "autonomous" in complex work domains, it is becoming clearer that it cannot provide all sufficient transparency information *in the moment* of action execution. Even if the human is capable of understanding it given time, it will all too frequently be the case that s/he will be engaged in other tasks and will be unable to devote sufficient attention and cognitive processing capability in a timely fashion. Instead, we need to strive to enable automation to participate in pre-mission planning and in post-mission debriefing and explanations in order to develop and accurately tune human trust and comprehension frameworks so that available capacity in the moment of use will be sufficient.

I make no claims that the recommendation and predictions provided above are particularly novel. As I said above, human-human interaction has wrestled with the problem of communicating intent and task status in communications-restricted domains throughout much of human history. Therefore, it is not particularly surprising that many of these methods of pre-mission planning and 1post-mission debriefing, along with other techniques for mental model reconciliation and improving the efficiency and predictability of intent communications have long been known and at least occasionally practiced in human-human teaming. It is also not particularly novel as applied to human-machine teaming. There have been calls for improved explanation capabilities for decades and my own work (e.g., [17,[30]]) has been specifically focused on improved human-automation intent communications for nearly 20 years now. I hope, however, that by showing the links between transparency, trust, situation awareness, explanation, debriefing, planning and mental model synchronization I may foster greater sensitivity to how these phenomena inter-relate and that that may, in turn, serve to make the development and use of human-automation systems more effective within the socio-organizational and macro-ergonomic systems where they are deployed.

Acknowledgments. This paper is entirely the work of the author and was not directly funded by any project or agency. I am indebted to Dr. Jesse Chen for providing a forum for initial thoughts on this topic, and to Rao Kambhampati for the insight that explanations generally need to focus only on mismatches in mental models. A prior version of this paper was presented at HCI International in Las Vegas on July 20, 2018 and appeared in the proceedings of that event.

[30] Miller, C. The FireFox fallacy: Why intent should be an explicit part of the external world in human automation interaction. In P. Smith and R. Hoffman, R. (Eds.): Cognitive Systems Engineering: A Future for a Changing World. CRC Press; Boca Raton, FL. Pp. 269-294 (2017).