Being Transparent about Transparency: A Model for Human-Robot Interaction

Joseph B. Lyons, PhD
Air Force Office of Scientific Research
875 N. Randolph St., Suite 325
Arlington, VA 22203
Joseph.Lyons@AFOSR.AF.MIL

Abstract

The current paper discusses the concept of human-robot interaction through the lens of a model depicting the key elements of robot-to-human and robot-of-human transparency. Robot-to-human factors represent information that the system (which includes the robot but is broader than just the robot) needs to present to users before, during, or after interactions. Robot-of-human variables are factors relating the human (or the interactions with the human; i.e., teamwork) that the system needs to communicate an awareness of to the users. The paper closes with some potentials design implications for the various transparency domains to include: training and the human-robot interface (including social design, feedback, and display design).

Human-Robot Interaction

The world is on the cusp of a robotics revolution given advances in technology and software relating to robotic systems. Yet, despite cutting-edge technology, fundamental research is still needed to better understand the dynamics of human-robot interaction. Robotic systems in the future will likely possess greater autonomy than current systems. Further, robotic systems within the military may be operated in hostile, complex situations and may be given the authority to execute lethal decisions within the battlespace (Arkin, 2009). The interactions between humans and robots will also likely get more complex as systems increase in their autonomy. For instance, instead of the teleoperation of contemporary military robotic systems such as Uninhabited Aerial Vehicles (UAVs), future operators will likely execute supervisory control of multiple robots. This evolution of robotic capabilities coupled with increased supervisory control from humans adds additional layers of complexity in the human-robot interaction, thus making the humans' trust of robotic systems a key aspect of the overall human-robot system. Understanding the trust a human has of a robotic system is important because trust will influence whether or not the human relies on the technology (Lyons & Stokes, 2012), and history shows that our reliance behavior is often suboptimal. There are numerous examples of major accidents in the aviation community being attributed to mis-calibrated reliance on technology. Thus, the research community is called to improve the trust calibration process when humans interact with advanced technological systems. Yet, there is a surprisingly little systematic guidance in these regards. Having shared context between humans and robots will be a critical facet of the overall system performance of human-robot teams (Stubbs, Wettergreen, & Hinds, 2007) and this will likely facilitate "appropriate" reliance on the robotic system (Lee & See, 2004). The current paper suggests that transparency between the robot and the human is one mechanism to facilitate effective interactions between humans and their robotic teammates. The terms robot and autonomous system are used interchangeably in the current paper to represent a system that must operate autonomously in a dynamic environment under some conditions of uncertainty.

Robot-to-Human Transparency

With added capability comes added responsibility, at least in theory. This raises questions concerning accountability with systems that operate under some degree of autonomy.

It turns out that humans do hold robots accountable for their mistakes (Kahn et al., 2012), at least more so than they would an inanimate object such as a toaster. However, the accuracy of such perceptions would benefit from accurate perceptions of the robot's ability, intent, and situational constraints, transparency in other words. While the idea of transparency has intuitive appeal, few studies have examined the impact of transparency on human-robot performance. Furthermore, the construct of transparency has been constrained to explanations of the robot's behavior, which may limit the value of the construct in explaining variance in human-robot performance. In one prior study, transparency was operationalized as the user's understanding of why a machine (in this case a robot) behaved in an unexpected way (Kim & Hinds, 2006). Kim and Hinds demonstrated that transparency had a greater impact on user perceptions of the robot (greater blame of the robot and less blame on others) when the robots had greater autonomy (Kim & Hinds, 2006). This suggests that transparency will increase in importance as a system's autonomous capabilities increase. Other studies of automated systems have defined transparency in terms of understanding the reliability of the system. Giving users information about the reliability of a system helps the users to calibrate their trust of the systems during information uncertainty (Wang, Jamieson, & Hollands, 2009). This is logical in that the added reliability information may help users to understand when to rely on the system and when not to. Other studies have defined transparency in terms of communicating information to users of an automated tool relating to the system's tendency for errors in a given context (Dzindolet, Peterson, Pomranky, Pierce, & Beck, 2003). This study asked individuals to use an automated tool that would use pattern recognition methods to identify a target in wooded scenery. The automation was not 100% reliable, and thus trust of the system fluctuated following errors, however individuals' trust recovered more quickly when they were given extra information about why the system failed. This added information about the reliability of the automated tool did help users appropriately calibrate their trust of the system.

Each of the above conceptualizations of transparency has merit and incorporating these lessons into new robotic systems could add value. However, there has yet to be a comprehensive treatment of the transparency construct in the literature and with increasing autonomy/capabilities of robotic systems the need for transparency is amplified. To fully address the complexities of transparency, one must consider information that a robotic system needs to convey to a human, as well information that the system needs to convey awareness and understanding of about a human. The former are classified as robot-to-human factors and the latter are labeled robot-of-human characteristics. The robot-to-human transparency factors include: an intentional model, task model, analytical model, and environmental model.

Intentional Model

Robotic systems are designed for a purpose, typically to support humans during some physical or analytical processes that humans either cannot do (or not do well) or tasks that humans do not want to do. Researchers believe that the physical appearance of a robot can afford cues to the users pertaining to the robot's functionality (Fischer, 2011; Goetz, Kiesler, & Powers, 2003). For example, cleaning robots may be designed to look like maids, or administrative assistants may have the appearance of a secretary. However, the match of appearance and functionality may not always be ostensible for users. For instance, automated systems built within cars may not have the opportunity for a direct linkage between function and physical appearance. Furthermore, robotic systems where the human-robot interaction occurs virtually also do not confer the opportunity for function-appearance alignment (at least not on a continuous basis). Therefore, it is important for the user of a robotic system to fully understand the intent or purpose of the robotic system. This higher-level understanding of the function of the robot can be distinguished from the task model relating to specific actions of the robot. Such an understanding will help users of robotic systems to put the actions of the robot in the proper strategic context. This is important because future robotic systems may interact with multiple users who may have different levels of baseline knowledge about the particular robot.

The intentional model is much deeper than simply providing a reflection of the robot's intended functionality. The intentional model should represent the design, purpose, and intent of the system. In other words, users should clearly understand "why" the robot was created, whether to provide customer service, elderly care, emotional support, medical advice, or some other form of support. In addition to "why", users to should understand "how" the system seeks to perform these actions. In this sense, the "how" should be defined in terms of broad categories of behavior perhaps akin to the three laws of robotics coined by Isaac Asimov. Rather than a task-driven model of behavior (which is presented in the task model), the users should have an understanding of the robots moral, albeit, programmed, philosophy of interaction with humans. If for instance, the robot was to override a human directive it would be useful for the human to first understand that such behavior is possible but also to understand why and when such behavior might occur. Further, it would be effective for human-robot teaming for the human to understand the robot's priority for general taxonomies of behavior. Specific examples of such could be as varied as that of humans, yet understanding the moral structure of a person certainly facilitates a rapid heurist processing of that individual which enables teaming or lack thereof.

Task Model

Once users understand the purpose of the robot they can begin to analyze the actions of the robot within a particular cognitive frame. The task model will provide the details to inform that cognitive frame during human-robot interactions. The task model could include an understanding of a particular task, information relating to the robot's goals at a given time, information relating to the robot's progress in relation to those goals, information signifying an awareness of the robot's capabilities, and awareness of errors.

For starters, the robot must communicate an understanding of the task at hand to the user. This will promote a shared awareness between the user and robot in terms of what actions need to be accomplished for a given task. For example, you might depict the tasks associated with a search and rescue mission to be something like: identify emergency location, calculate optimal route to search location, travel to search location, search for victims, identify life signs of victims, notify emergency personnel, and return to base. While this is a simplification of a highly complex scenario, simple task analyses could be used for a variety of robotic tasks to further break down the actions/behaviors of the robots for a particular scenario. The robot must also communicate its intent in terms of what goals it is trying to accomplish for a given task. This will provide useful to the human regarding where the robot is in terms of its task sequence and why it is performing a certain action/behavior. Using the above scenario as one example, the robot could communicate to the user that is currently in the "identify life signs of the victim" phase of the task which could explain to the human why the robot is hovering in a specific spot for an extended period of time. If the human observed the same behavioral pattern but understood that the robot was trying to return to base then the human would quickly surmise that there might a problem. Information such as this will improve the human's situational awareness of the robots actions and will aid the human in a supervisory control capacity.

An important facet of the task model would be the robot's awareness of it capabilities in a given context. Understanding for example, that the reliability of the robotic system is questionable under specific conditions would do a great deal to promote appropriate trust from the human users of the systems. Context-specific reliability estimates can also enhance the performance the human-robot team as the robot may be redirected either internally if autonomous or by the human to engage in situations where its reliability is the strongest, if that option is possible. Again moving back to the search and rescue example above, if the robot could cue the human teammate that it the search location was in a mountainous region and that its sensors had lower reliability to detect life signals in that terrain then the human could make a decision to use a different platform or to send in a human search party. Alternatively, the human could decide to allow the robotic system to continue but with the understanding that its performance may be degraded. The juxtaposition of the awareness of environmental constraints and robotic capabilities will be a key aspect of robot-to-human interaction.

The final aspect of the task model could include an awareness of progress in relation to some goal state. This requires that the robot understand that it has some goal in relation to some task, the task at hand, and its progress in relation to that goal. This kind of self-regulatory function at the outset appears highly complex, and for non-physical tasks it may be highly complex, but not impossible if the robot understands it's higher-level goals, understands what it needs to do in relation to those goals, and has the capacity to self-monitor. If the robot could self-identify mistakes, that would have a significant impact of the user's ability to calibrate their reliance with a system.

Analytical Model

One of the key benefits of automated systems is that their capabilities for processing large amount of data often exceed that of a human. Yet, because of the complexity of the information these systems are asked to analyze it is often possible for human users (especially non-computer science types and non-robotics experts) to be confused about how the robot is doing the analysis. The analytical model needs to communicate the underlying analytical principles used by the robot to make decisions. This information will help the human understand how the robot makes decisions. Such awareness will be very useful during situations where there is a great deal of uncertainty regarding the appropriate course of action. For instance, knowing that a particular robot fuses information from satellite imagery and ground sensors in determining where potential emergency zones are located could be useful if the human knew that the ground sensor networks had been compromised. The human would then be able to use more manual control of the robot's navigation system, as one example. The analytical model is very much a knowledge-based component where the users need to understand the analytical structure of the robot's decision process.

Environment Model

Given the harsh conditions, potentially hostile environments, and time constraints that future military robots will be asked to operate within, it will be critical for robotic systems to have the capability of understanding the dynamics of its surrounding environment. Robots should be capable of communicating to humans an understanding of the geographic variance (i.e., terrain), weather conditions, potential for hostility, and temporal constraints within a particular environment. This provides a user with a glimpse of what the robot is experiencing and it further enhances the human's situation awareness (SA) during uncertainty. Knowing that the robot understands its environmental conditions will aid the human in calibrating their reliance on the robot especially if the robot is capable of communicating an awareness of its potential limitations in specific environmental conditions. For instance, a robot may be forced to increase its altitude because of hostilities in an area. If the robot communicated an awareness that its sensors were less effective at higher altitudes human teammates could appropriately recalibrate their trust of the robot's performance. In addition to environmental characteristics, the robotic system should be able to communicate awareness of temporal constraints. For instance, it is possible that robotic systems deployed in future military domains will need to adjust to extended periods of inactivity to rapidly evolving demands without disruption of the mission. Boredom, one unfortunate concomitant of long inactive periods, is often reported by soldiers in a combat zone because such missions tend to have long periods of inactivity coupled with short bursts of activity. Boredom is not a limitation of robotic systems, however such systems will need to easily transition from low to high or high to low periods of demand and adapt their behavior accordingly.

Robot-of-Human Transparency

The previous section focused on information that the robot needs to share with the human that represents, to a large degree, the robot's view of the world ranging from the task at hand, its analytical underpinnings, and awareness of its goals and limitations in particular environments. The robot also needs to communicate an awareness of factors relating to its human teammates, and these factors are herein termed robot-of-human factors. The precursors to such capabilities are already evident in modern automotive designs. Recent advances in automotive safety systems may allow a machine to intervene when it detects cues from drivers that they are operating their vehicle at unsafe levels (Inagaki, 2008). The notion of robotic systems monitoring human performance and intervening when necessary is of growing interest as automated systems acquire more dominant roles within domains such as driving and aviation. For such systems to work as intended they need to understand what human-centric metrics are related to performance, who is responsible for what tasks, and what the goals of the human user are. In other words, just as the human user needs to understand information about the robot, the overall human-robot system would benefit from the robot having awareness and understanding of several human-centric factors.

Teamwork Model

A crucial element of human-robot teaming is for both humans and robots to understand the division of labor for a given task or set of tasks. Parasuraman, Sheridan, and Wickens (2000) provide a useful framework for division of labor between humans and robots in their discussion of different stages of information processing and their discussion of levels of automation. They discuss information processing as consisting of information acquisition, information analysis, decision analysis and selection, and action implementation. Even such a high-level framework as this could be useful in terms of fostering a shared awareness between a human and a robot. Once the higher-level division of labor is shared and understood between both parties, it will be important for a set of norms to be defined to negotiate uncertainties and dynamic nature of teamwork, this will be especially true if the human is executing supervisory control of multiple robots at one time.

The levels of automation discussion by Parasuraman and colleagues (2000) can be a useful way to formulate "social" norms between the robot and the human. At the lowest end of the continuum the human executes complete control of the system with no inputs from the automation, whereas, at the highest level the automation operates completely autonomously. More realistic scenarios are those that exist somewhere between the extremes. In support of the teamwork model, the robot should convey an understanding of what tasks it is responsible for, what tasks the human is responsible for, and what level of autonomy it is currently operating under. This information will allow the human to forecast certain actions from the robot which ultimately leads to some sense of predictability (an important driver of trust in automated systems; Muir, 1994). In the presence of dynamic temporal constraints it would also be useful for humans to have the capability to toggle between different levels of automation for given platforms. Such capabilities will likely be a key aspect of successful supervisory control of multiple robotic systems.

Human State Model

After the robot-human team develops a shared awareness of the task, and the division of labor with regard to that task, it will be important for the robot to communicate an understanding of the humans' cognitive, emotional, and physical state. Like a typical effective human teammate, the robot should be able to sense when the human is under distress. The robot might monitor human states such as the cognitive workload of the human. When the robot senses that the human is overloaded it might recommend increasing its level of autonomy while the human recovers and addresses key tasks. When the robot senses that the human is experiencing certain emotions such as frustration, anger, or fear it may prompt the robot to ask the human if he/she needs assistance of additional information. Finally, when the robot senses physical vulnerabilities such as fatigue it might execute protocols to alert the human or even assume autonomous control if the human's limitation could cause a safety concern. Such systems are being developed within automotive industry to sense fatigue among drivers and if necessary assume control of the vehicle. Examples of three types of systems may include: arousing drivers' attention to motivate them to reallocate attention to another aspect of the driving task, warning systems that encourage to make the right decisions and avoid accidents, and systems that take action when a lack of action is detected from the human drivers (Inagaki, 2008). Inagaki (2008) discusses the importance of the concept of transparency in such systems, "To do this the machine will need to understand the human's psychological and or physiological state, the situation, the intent of the human, and whether the human's actions match the needs of the situation."

Designing for Transparency

Once the information facets relating to the robot-to-human and robot-of-human are identified, they will need to be incorporated into the human-robot system. There are two likely opportunities to inject transparency into the human-robot system: training and the human-robot interface.

Decades of scientific endeavor have solidified the notion that team training is a useful method to foster effective team performance (Salas, Cooke, & Rosen, 2008). Therefore, it is logical to believe that team-based training will also enhance teams comprised of humans and robots. Training on the robotics systems themselves can be crucial in understanding the intentional model and analytical model of the robot. This type of background information and a deep understanding of the robotic system can sometimes be missing from studies where novice users are asked to interact with robotic systems for a brief duration. Researchers should treat such studies with caution as a deep understanding of the robot could significantly change how the robot is perceived and interacted with (i.e., relied on). Training with the robot (either through staged exercises or on-job-training) can be useful in establishing the task model, environmental model, teamwork model, and in understanding how the robot incorporates the human state model in adapting its behavior. Specifically, training that cues users of systems to understand the limitations of a robotic system in a given context will be very useful in supporting appropriately calibrated trust among human users. In one study looking at trust in an automated system the researchers found that humans more readily adopted better calibration strategies when they interacted with unreliable automation relative to highly reliable automation, which ultimately led to higher complacency (Rovira, McGarry, & Parasuraman, 2007). Thus training offers an interesting method to foster calibrated trust of robotic systems.

In addition to training, the human-robot interface offers another opportunity to foster transparency between the human the robot. The interface between humans and robots may exist at a number of levels including: informational (e.g., information displays), communicative, and physical. Interface features at the informational level could use information displays to fuse information geospatially, temporally, and dynamically. This information fusion could provide useful information relating to task model, analytical model, and environmental model. Though such information fusion displays should be approached with some caution as too much information, or a non-intuitive display may confuse and frustrate the users of robotic systems. Aspects of the teamwork model could also displayed and potentially manipulated within an informational display. Cues to signify the division of labor overtime would be useful in this regard, provided that the display mirrors the dynamic nature of the task and the human-robot interaction. Indicators of reliability can be a useful piece of information to display to users as this may help the humans calibrate their trust of the robots (Wang et al., 2009). Even the mere presence of a face associated with a decision aid has been shown to impact trust and decision accuracy among users of a medical decision aid (Pak, Fink, Price, Bass, & Sturre, 2012), suggesting the potential for added social design features.

The communicative interface would naturally involve a voice or text exchange between the human and the robot. The style of communication between the robot and the human would of course matter. Research has shown that user trust and performance is influenced by the social etiquette of the feedback provided by an automated tool, in this case patient versus impatient (Parasuraman & Miller, 2004). Patient styles were associated with higher trust and better performance for users who were interacting with the automated system. Finally, the physical interface could include features such as robotic emotional expression and gestures, which in combination are effective in communicating the depiction a robots' emotional state (Zecca et al., 2009). This level of interface shows promise since recent research has reported that humans use similar cues to gauge the trustworthiness of robots as they do humans (Desteno et al., 2012).

Conclusion

The current paper discussed the premise that an extended conceptualization of transparency may benefit the human-robot interaction. To date, the transparency construct has been limited to explanations for anomalous behavior, reliability indices, and attempts to define the analytic underpinnings of a system. These aspects of the machine are certainly relevant and should be designed into novel systems, however they have been considered in isolation and additional information relating to the robot-to-human and robot-of-human factors could add considerable value in complex human machine interactions where robotic systems have high degrees of autonomy. Transparency in this sense is a more comprehensive treatment of the information that a human operator may need or want when dealing with autonomous systems under high stress, workload, and uncertainty.

Robotic systems represent the future. It is plausible that robots will revolutionize daily life as the internet and social media have done in recent years. However, to fully realize the potential of such systems and to recognize their inherent limitations researchers and engineers will need to consider the information that drives the human-robot interaction. Given the imperfect track record of automation use in recent years it is imperative that researchers consider the elements of human-robot interaction that allow individuals to properly calibrate their reliance on these systems, particularly as technology gets more complex technology is fielded in increasingly complex scenarios. A broader operationalization of transparency offers one mechanism to foster optimal calibration between humans and autonomous systems.

References

Arkin, R.C. 2009. Governing lethal behavior in autonomous robots. Boca Raton, FL: CRC Press.

Desteno, D.; Breazeal, C.; Frank, R.; Pizarro, D.; Baumann, J.; Dickens, L.; and Lee, J. 2012. Detecting the trustworthiness of novel partners in economic exchanges. *Psychological Science*. In press.

Dzindolet, M.T.; Peterson, S.A.; Pomranky, R.A.; Pierce, L.G.; and Beck, H.P. 2003. The role of trust in automation reliance. *International Journal of Human-Computer Studies*. 58: 697-718.

Fischer, K. 2011. How people talk with robots: Designing dialogue to reduce user uncertainty. *AI Magazine*, 31-38.

Goetz, J.; Kiesler, S.; & Powers, A. 2003. Matching robot appearance and behavior to tasks to improve human-robot cooperation. Proceedings of the 2003 IEEE International Workshop on Robot and Human Interactive Communication (55-60). Millbrae, CA.

Inagaki, T. 2008. Smart collaboration between humans and machines based on mutual understanding. *Annual Reviews in Control*. 253-261.

Kahn, P.H.; Kanda, T.; Ishiguro, H.; Gill, B.T.; Ruckert, J.H.; Shen, S.; Gary, H.E.; Reichert, A.L.; Freier, N.G.; and Severson, R.L. 2012. Do people hold a humanoid robot morally accountable for the harm it causes? *Proceedings of the 7th ACM/IEEE International Conference on Human-Robot Interaction*. 33-40.

Kim, T.; and Hinds, P. 2006. Who should I blame? Effects of autonomy and transparency on attributions in human-robot interactions. *Proceedings of the 15th International Symposium on Robot and Human Interactive Communication (RO-MAN06)*. 80-85. Hatfield, UK: IEEE.

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

Lyons, J.B.; and Stokes, C.K. 2012. Human-human reliance in the context of automation. *Human Factors*. 54(1): 112-121.

Muir, B.M. 1994. Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics*. 37(11): 1905-1922.

Pak, R.; Fink, N.; Price, M.; Bass, B.; and Sturre, L. 2012. Decision support aids with anthropomorphic characteristics influence trust and performance in younger and older adults. *Ergonomics*. 1-14.

Parasuraman, R.; and Miller, C. 2004. Trust and etiquette in high-criticality automated systems. *Communications of the ACM*. 47: 51-55.

Parasuraman, R.; Sheridan, T.B.; and Wickens, C.D. 2000. A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, man, and Cybernetics – Part A: Systems and Humans,* 30: 573-583.

Rovira, E.; McGarry, K.; and Parasuraman, R. 2007. Effects of imperfect automation on decision making in a simulated command and control task. *Human Factors, 49*(1): 76-87.

Salas, E.; Cooke, N.J.; and Rosen, M.A. 2008. On teams, teamwork, and team performance: Discoveries and developments. *Human Factors*. 50(3): 540-547.

Stubbs, K.; Wettergreen, D.; and Hinds, P.J. 2007. Autonomy and common ground in human-robot interaction: A field study. *IEEE Intelligent Systems*. 42-50.

Wang, L.; Jamieson, G.A.; and Hollands, J.G. 2009. Trust and reliance on an automated combat identification system. *Human Factors*. 51: 281-291.

Zecca, M.; Mizoguchi, Y.; Endo, I.F.; Kawabata, Y.; Endo, N.; Itoh, K.; and Takanishi, A. 2009. Whole body emotion expressions for KOBIAN Humanoid Robot: Preliminary experiments with different emotional expression patterns. Paper presented at the 18th IEEE International Symposium on Robot and Human Interactive Communication. Toyama, Japan.