Improving Automation Transparency: Addressing Some of Machine Learning's Unique Challenges

Corey K. Fallon(✉) and Leslie M. Blaha

Pacific Northwest National Laboratory, Richland, WA, USA
{corey.fallon, leslie.blaha}@pnnl.gov

Abstract. A variety of factors can affect one's reliance on an automated aid. Some of these factors include one's perception of the system's trustworthiness, such as perceived reliability of the system or one's ability to understand the system's underlying reasoning. A mismatch between the operator's perception and the true capabilities and characteristics of the system can lead to inappropriate reliance on the tool. This improper use of the system can manifest as either underutilization of the technology or complacency resulting from over-trusting the system. Increasing an automated tool's transparency is one approach that enables the operator to more appropriately rely on the technology. Transparent automated systems provide additional information that allows the user to see the system's intent and understand its underlying processes and capabilities. Several researchers have developed frameworks to support the design of more transparent automation. However, these frameworks may not fully consider the particular challenges to transparency design introduced by automation that leverages machine learning. Like all automation, these systems can benefit from transparency. However, artificial intelligence poses new challenges that must be considered when designing for transparency. Unique considerations must be made in terms of the type, and amount or level of transparency information conveyed to the user.

Keywords: Transparency · Machine learning · Automation

1 Appropriate Reliance

Through their interactions with automation, operators form perceptions of an automated system's technical competence, ability to function consistently (i.e., reliability) and understanding of the system's processing. These perceptions affect one's trust in the technology and ultimately one's reliance on the automation [1]. When a mismatch exists between the operator's trust and the actual capabilities of the system the operator may under-trust or over-trust the automation [1]. An operator under-trusts the automation when his or her trust is less than what is appropriate given the reliability and capabilities of the technology. For example, research on alarm systems has investigated the impact of under-trust on alarm system compliance. Several studies suggest that an unreliable alarm system which produces a high false alarm rate may result in slower and less frequent operator compliance with the system [2, 3]. Sorkin [4]

D. D. Schmorrow and C. M. Fidopiastis (Eds.): AC 2018, LNAI 10915, pp. 245–254, 2018.
https://doi.org/10.1007/978-3-319-91470-1_21

revealed that in some instances false alarms have led to a complete rejection of the technology, such as the deactivation of a warning system due to high false alarm rates.

Unreliable automation can also cause problems for the operator if the operator over-trusts the system. Over-trust can lead to complacency [1] which has been characterized by a reduction in system monitoring below what would be considered optimal, resulting in poor operator performance [5]. For example, a warfighter who over-trusts a sensor used for target detection may be less vigilant and fail to notice that the sensor is providing old or inaccurate information. In this example, the warfighter's lack of awareness would lead to inappropriate reliance on the automation (i.e., sensor) and poor performance. Complacency has been identified as one of the major factors contributing to accidents and incidents in aviation [6].

2 Automation Transparency

One possible way to calibrate trust may be to improve automation transparency [7]. In a human-system relationship transparency "is concerned with revealing information to the user and supplementing expected outputs, which reveal how a system works and/or what it is doing" [8, p. 2]. According to Lyons [9], transparency allows the operator to correctly perceive the ability, intent, and situational constraints of the automation or autonomous system. A review of the cognitive systems engineering and human factors literature identified several system design techniques to support operator cognitive performance. Techniques were identified to promote situation awareness such as providing access to historic information. Techniques for alleviating attentional demand, such as reducing visual clutter by highlighting critical information, were also identified. In addition to these techniques, the researchers included several guidelines for improving transparency such as providing operator access to unfiltered data and providing explanations for how raw data is filtered and processed. According to this review improving system transparency was identified as an important step toward supporting operator cognitive performance [10].

Empirical evidence suggests that providing operators with transparency can increase trust and operator reliance on an unreliable automated system. Fallon et al. [11] investigated the impact of Likelihood Alarm Displays (LADs) on trust. These displays generate an alarm signal coupled with additional probabilistic information regarding the signal's validity [12]. Fallon et al. [11] found that this additional transparency information significantly increased user trust in the system. In addition, Heldin et al. [13] manipulated the transparency of an automated target classification aid to support the target discrimination of fighter pilots. According to this research, providing fighter pilots with additional information about the uncertainty of the sensor increased user trust and the number of correct classifications. By improving the appropriateness of the operator's reliance on unreliable automation, the presence of transparency information may reduce errors. Also, in situations where operators have previously rejected an unreliable automated system, adding transparency to the automation may decrease operator workload by increasing reliance on the time-saving automation.

Although empirical evidence suggests increasing transparency may help calibrate operator trust and reliance on the automation, guidance for incorporating transparency information into design is lacking. Several researchers have made some initial progress. Lyons [9] proposed multiple models for informing the implementation of transparency in human-robot interaction. For example, this researcher draws a useful distinction between what he refers to as the Intention, Task, and Analytic Models. According to Lyon's [9] Intention Model, robot designers have a responsibility to ensure that the operator fully understands the machine's functionality and purpose. In contrast, the Task Model, emphasizes the communication of system goals and progress toward those goals. The machine should also share information about its ability to perform tasks and acknowledge its errors. Finally, Lyon's [9] Analytical Model highlights the importance of sharing the system's underlying analytical processes with the operator. Adherence to this model allows the operator to understand how the system is solving problems to accomplish its goals.

Chen et al. [14] also proposed a model for designing transparency. Their model was specifically developed to support situation awareness and is known as the Situation Awareness-based Agent Transparency Model. Chen et al.'s [14] model maps onto Endsley's [15] three levels of situation awareness: perception, comprehension, and projection. Similar to Lyon's [9] Task Model, Chen et al.'s [14] model stresses the importance of providing information about system's task performance and goals. According to these researchers this type of transparency information helps facilitate operator perception. Chen et al. [14] also suggest transparency information specific to the system's underlying reasoning allows the operator to comprehend how the system is working. This type of information is consistent with Lyon's [9] Analytical Model and according to Chen et al. [14] this information promotes Endsley's second level of situation awareness: comprehension. Chen's [14] model is somewhat unique in its emphasis on projection. According to their model, transparency information should support the operators' ability to make predictions about the system's future performance.

Both Lyons [9] and Chen et al. [14] provide researchers with useful organizational frameworks from which to build. However, advances in automation in the form of artificial intelligence (AI) may require researchers to expand on the existing guidance. Specifically, automation that leverages machine learning presents new challenges and requires additional design considerations to ensure that these systems are transparent. AI that leverages machine learning can improve its reliability and accuracy with experience. The advances in automation fueled by machine learning pose new challenges for designers to ensure system transparency and appropriate reliance.

3 Machine Learning

Broadly, machine learning refers to AI systems that train and learn from past activity without the specific improvements being explicitly programmed. The systems are given algorithms for learning together with training examples/data from which they determine what to learn. There are three broad classes of learning algorithms, each with different implications for the types of interactions or user inputs that will be required or informative to the process. Because of the different level of user engagement in the learning process, each type can have different implications for the types of transparency or information that must be conveyed to those users. Supervised learning requires a completely labeled training set from which the algorithm must draw its feedback. Traditional supervised machine learning requires all the training labels be provided up front. Advances in active and interactive machine learning are looking for ways to make this a more incremental process. In either case, the user may need a high degree of engagement with the system, implying the user must understand what the machine learner needs.

Semi-supervised and unsupervised learning algorithms need partial to no labeled input from users. This may simplify the process of constructing training exemplar sets, but it also changes the degree to which the user is engaged with the system. Unsupervised systems with minimal user interactions may also provide minimal information back to the user about the process, due to the lack of user involvement. In this case, it may be desirable to integrate corrective feedback for errors or other simple ways to engage the user, as well as training about the machine capabilities or explicit transparent information.

The increasing interest in deep learning systems has recently highlighted the impenetrable nature of black box machine learning for human observers. The hidden layers do not usually operate on human-recognizable patterns. This has resulted in a push for explainable systems, capable of translating those activities into something humans can understand. This push is based on the assumption that increasing human understanding through explanations, which are one form of transparency, will result in improved trust in the deep learning systems. What is unclear in this argument is if the explanations need to be about the internal reasoning processes or just about the classification outputs, or about some other aspect of the system entirely. Indeed, some recent work has shown that explaining the machine's reasoning can aid user in selecting the more effective classifier [16]. However, because machine learning can be used at multiple levels of automation and for multiple purposes in systems, *post hoc* explanations about machine reasoning may not always be necessary or enough to engender the appropriate trust and reliance. There are multiple types of transparency as well as degrees or levels of transparency that may be needed. Considerations of these will be informative to the system design process.

3.1 Type of Transparency

Automation equipped with machine learning adds an additional dimension of complexity and capability to the system that should be communicated to the user. At a basic level the tool should be transparent about its ability to learn and improve its own performance. This type of transparency is consistent with Lyons' [9] Intention Model. The tool should communicate to operators that it intends to learn from their input. This intention may be communicated during training or explicitly communicated to the operator during the first interaction.

In addition, the system should be clear what input is needed from the operator to improve system learning. If one of the operator's responsibilities is to teach the system, the system should provide guidance for accelerating its own learning. This is a key principle of active and interactive machine learning systems, particularly for robots, where the learner selects the data from which it will learn or requests the information or training feedback it needs [17, 18]. For example, if a recommender system based on machine learning learns only from specific user behaviors such as user product ratings, the automation should inform the user that only this information will be helpful for learning. If the operator is not aware of this behavior's importance, he or she may choose not to provide product ratings and ultimately stunt system growth. In general, failure to communicate the importance of certain behaviors for learning can lead to inefficient or unhelpful user interactions that slow the system's progress.

Designers should develop tools that clearly communicate what user actions are required for training. However, designers should also be considerate of the training burden placed on the user and take steps to reduce this burden. It has been found that simply treating the user as an oracle and repeatedly requiring the user to give right/wrong feedback is frustrating to users [19]. This interaction puts systems into a situation where users could stop giving any input at all, thus breaking the training cycle.

A principle of mixed-initiative systems is that user interactions should be used implicitly by the system to understand user goals and provide machine support toward those efforts [20]. While mixed-initiative systems are not necessarily all machine-learning based, the same principle applies to considering how observation of ongoing user interactions and implicit learning about the user can inform the machine learning. Semantic interactions were developed as one form of implicit learning about the patterns in data of interest to the user [21]. Jasper and Blaha [22] suggested that machines might learn implicitly from interactions with user interface metaphors. We note, importantly, that while implicit learning about the user may be helpful to the system and less intrusive to the user's analytic process, they may also require some degree of transparency so that the user provides interaction inputs that are valuable to the process.

Expanding on Lyon's [9] Task Model and Chen et al.'s [14] emphasis on projection, the tool should also provide information that will help the user gauge how quickly and smoothly the system will learn as well as the upper limits of the system's performance. This type of transparency will allow a new user to manage expectations about how the system will (or will not) improve with repeated use. For some systems, the relationship between operator input and system improvement may be linear. For other automated tools system improvement may come in fits and starts despite consistent attempts by the operator to train the tool.

Appropriately calibrating user expectations with the system's rate of learning may be particularly important for tools early in their learning (i.e., novice tools). For some systems, there may be an initial training ("burn-in") period where little to no performance improvement should be expected. In other instances, the system may overcorrect early in its learning resulting in performance errors. If these learning delays are not anticipated, the user may become discouraged and underutilize or even completely reject the technology.

It may also be useful for the tool to communicate the upper limits of its capability. If it is not reasonable for the operator to expect more than 80% reliability, this information should be communicated to the operator. Without this transparency, the operators may grow frustrated when they do not see performance improve above this threshold. This is a particularly salient aspect of working with machine learning systems, because the users can often see the mistakes. If the user is expecting 100% accuracy in classification or labeling, for example, then the errors can be surprising and unexpected or even result in a catastrophic loss of trust in the system from which the human-machine team may not recover.

For some system/environment combinations, it may be impossible for system developers to reasonably predict how quickly the system will learn or the upper limit of its performance. In these situations, it may be particularly useful for the system to provide users easy access to historical data. As Chen et al. [14] noted in their model of transparency, examining past system performance can help the operator predict future performance. AI that tracks its own performance on a task will help the operator understand how quickly the system is learning the task and the limits to the system's performance. In addition, allowing the operator to examine how much and how quickly the system learned in previous situations may provide insight into how it will learn in a new but similar environment.

3.2 Level of Transparency

The guidance above is simply an expansion of existing transparency models to support human-automation interaction [9, 14]. These models provide frameworks that organize transparency information by type such as distinguishing between task-focused and analysis-focused information. Selecting the correct type(s) of transparency information to display will help facilitate appropriate reliance and acceptance of the tool. However, designers must also consider the amount of transparency information that is appropriate for display and/or access at any given time. In this paper, we refer to the amount of transparency information as the level of transparency. We see parallels between levels of transparency and the levels of automation proposed by Parasuramen et al. [23].

The level of transparency information can range from a complete lack of transparency at the lowest level to a salient display detailing the system's performance and/or underlying reasoning process at the highest level. In addition, one might consider allowing access to detailed information within a menu structure a lower level of transparency than presenting this information on a display. Such access is consistent with Shneiderman's Visual Information Seeking Mantra broadly applicable to interactive interfaces: overview first, zoom and filter, details on demand [24, p. 365]. In practice, this means that when there is a danger of information overload, or more information than may be immediately useful to the user, the information should be made available in an easily findable way, on the demand of the user according to his or her needs. This same principle may be very useful to offering flexible degrees of transparency information, such that users desiring more details can access them, but they are not immediately cluttering the information displayed to users who do not desire the information. A variety of factors should be considered when choosing the appropriate level of information. Some of these factors include the operator's workload and experience with the tool, the reliability of the automation, and consequences of an error. Choosing the appropriate level of transparency is an important design decision for any automated system. However, in this paper we will focus specifically on the unique challenges posed by automated tools that improve with user interaction.

In static automated systems, the required level of transparency for appropriate reliance may drop as the user learns the capabilities of the tool and how it processes information. In the case of machine learning, increased familiarity with the tool coupled with improved performance may require the need for very little transparency. However, it is important to consider additional factors when deciding the level of transparency. One important factor to consider is the consequence of an error in the operational environment.

In an environment with relatively low stakes, a user who is familiar with the highly reliable automated tool may have little need for transparency. A high level of transparency information under these conditions may at best be ignored and at worst be a distraction that places unwanted attentional demands on the operator. Perhaps the ideal design approach for a low consequence environment is an adaptive display that reduces its level of transparency as both system reliability and operator familiarity increase. At peak human-machine teaming performance, the operator interacts with the automation seamlessly with little explicit communication. This relationship is similar to a high functioning human-human team that relies on implicit coordination [25, see Fig. 1].

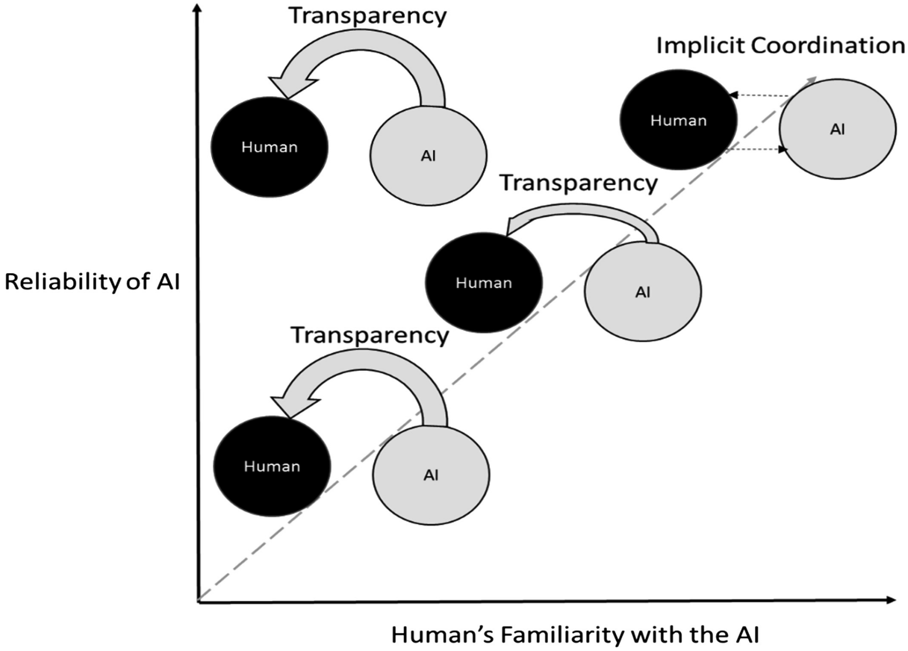

Fig. 1. The impact of operator familiarity and AI reliability on the level of AI transparency.

Figure 1 also depicts a scenario where the automation's reliability is high despite the user's lack of familiarity with the tool. Such a situation might exist if the AI has been trained by other users and is now available for a new operator to use. We propose that the level of transparency should be high in this situation despite the system's high reliability. A high level of transparency may be necessary because the user lacks the hands-on experience needed to become aware of the system's superior performance. The high level of transparency can be used to accelerate trust calibration and appropriate reliance in the absence of familiarity.

It is important to note in certain situations it may never be appropriate to design for very low levels of transparency. In high stakes environments where the consequences of an error are severe, it may always be appropriate to provide a high level of transparency. For example, despite familiarity with a highly reliable decision support tool, a high consequence environment may still require the need to understand the reasoning behind the system's recommendation [see Fig. 2].

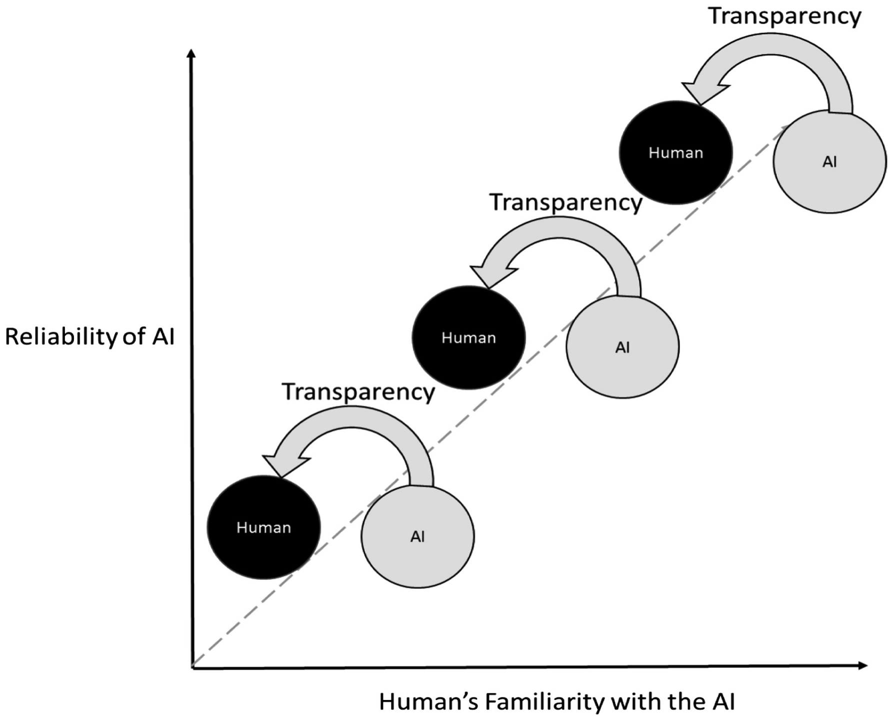

Fig. 2. The impact of operator familiarity and AI reliability on the level of AI transparency in a high consequence environment.

4 Conclusion

The success of a human-machine team is dependent on an operator's ability to appropriately rely on the automation. Over-reliance can lead to complacency, lapses in operator attention and error. Underutilization of the tool can result in increased workload and inefficiencies. Through trial and error the operator may be able to calibrate his or her trust in the tool and learn to rely on it appropriately. However, the trial and error technique can be time consuming and prone to errors as the operator attempts to understand the system's capabilities and limits. In addition, trial and error may never fully reveal the underlying analysis that governs the system's behavior. Increasing system transparency may be one technique to facilitate appropriate reliance more efficiently and with fewer errors.

As the capabilities of automated tools grow, the added complexity of these tools poses new challenges for transparency design. Machine learning in particular adds an additional dimension that must be considered when building transparency into these systems. This advancement in AI requires designers to consider a new type of transparency. In addition to communicating information about the systems' capabilities, goals and underlying reasoning, automation that leverages machine learning should communicate information about how the system learns.

Machine learning also creates additional challenges for designing the appropriate level of transparency. Systems governed by machine learning will likely improve with operator interaction. In low consequence environments, this increase in system accuracy and reliability may benefit from a decrease in transparency level. A level that adjusts automatically or is adjustable by the operator may be ideal given the potential for shifts in system performance over time.

Acknowledgments. This effort was sponsored by the Analysis in Motion Initiative at the Pacific Northwest National Laboratory. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government.

References

1. Lee, J.D., See, K.A.: Trust in automation: designing for appropriate reliance. Hum. Factors **46**, 50–80 (2004)
2. Bliss, J.P., Fallon, C.K.: Active warnings II: false warnings. In: Wogalter, M.S. (ed.) The Handbook of Warnings, pp. 231–242. Lawrence Erlbaum Associates, Mahwah (2006)
3. Getty, D.J., Swets, J.A., Pickett, R.M., Gonthier, D.: System operator response to warnings of danger: a laboratory investigation of the effects of the predictive value of a warning on human response time. J. Exp. Psychol. Appl. **1**, 19–33 (1995)
4. Sorkin, R.D.: Why are people turning off our alarms? J. Acoust. Soc. Am. **84**(3), 1107–1108 (1988)
5. Parasuraman, R., Manzey, D.H.: Complacency and bias in human us of automation: an attentional integration. Hum. Factors **52**, 381–410 (2010)
6. Funk, K., Lyall, B., Wilson, J., Vint, R., Niemczyk, M., Suroteguh, C., Owen, G.: Flight deck automation issues. Int. J. Aviat. Psychol. **9**(2), 109–123 (1999)
7. Fallon, C.K., Murphy, A.K.G., Zimmerman, L., Mueller, S.T.: The calibration of trust in an automated system: a sensemaking process. Paper published in The 2010 International Symposium on Collaborative Technologies and Systems, Chicago, IL, pp. 390–395 (2010)
8. Ososky, S., Sanders, T., Jentsch, F., Hancock, P., Chen, J.Y.C.: Determinants of system transparency and its influence on trust in and reliance on unmanned robotic systems. In: SPIE Defense + Security, pp. 90840E–90840E-12 (2014)
9. Lyons, J.B.: Being transparent about transparency: a model for human-robot interaction. In: Sofge, D., Kruijff, G.J., Lawless, W.F. (eds.) Trust and Autonomous Systems: Papers from the AAAI Spring Symposium (Technical report SS-13-07). AAAI, Menlo Park (2013)
10. Long, W., Cox, D.A.: Indicators for identifying systems that hinder cognitive performance. In: Proceedings of the Eighth International Conference on Naturalistic Decision Making, Asilomar, CA, pp. 171–175 (2007)

11. Fallon, C.K., Bustamante, E.A., Ely, K.M., Bliss, J.P.: Improving user trust with a likelihood alarm display. In: Proceedings of the 11th International Conference on Human-Computer Interaction, Las Vegas, NV (2005)
12. Sorkin, R.D., Kantowitz, B.H., Kantowitz, S.C.: Likelihood alarm displays. Hum. Factors **30**, 445–459 (1988)
13. Helldin, T., Ohlander, U., Falkman, G., Riveiro, M.: Transparency of automated combat classification. In: Engineering Psychology and Cognitive Ergonomics, pp. 22–33 (2014)
14. Chen, J.Y.C., Procci, K., Boyce, M., Wright, J., Garcia, A., Barnes, M.: Situation Awareness-Based Agent Transparency. (Final Report, Army Research Laboratory 6905) Aberdeen Proving Ground, MD 21005-5425 (2014)
15. Endsley, M.R.: Toward a theory of situation awareness. Hum. Factors **37**(1), 32–64 (1995)
16. Ribeiro, M.T., Singh, S., Guestrin, C.: Why should I trust you?: Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144. ACM (2016)
17. Amershi, S., Cakmak, M., Knox, W.B., Kulesza, T.: Power to the people: the role of humans in interactive machine learning. AI Mag. **35**(4), 105–120 (2014)
18. Guillory, A., Bilmes, J.A.: Simultaneous learning and covering with adversarial noise. In: Proceedings of the 28th International Conference on Machine Learning (ICML 2011), pp. 369–376. International Machine Learning Society, Inc., Princeton (2011)
19. Cakmak, M., Chao, C., Thomaz, A.L.: Designing interactions for robot active learners. Auton. Ment. Dev. **2**(2), 108–118 (2010). https://doi.org/10.1109/TAMD.2010.2051030
20. Horvitz, E.: Principles of mixed-initiative user interfaces. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 159–166. ACM (1999)
21. Endert, A., Fiaux, P., North, C.: Semantic interaction for sensemaking: inferring analytical reasoning for model steering. IEEE Trans. Vis. Comput. Graphics **18**(12), 2879–2888 (2012)
22. Jasper, R.J., Blaha, L.M.: Interface metaphors for interactive machine learning. In: Schmorrow, D.D., Fidopiastis, C.M. (eds.) AC 2017. LNCS (LNAI), vol. 10284, pp. 521–534. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-58628-1_39
23. Parasuraman, R., Sheridan, T.B., Wickens, C.D.: A model for types and levels of human interaction with automation. IEEE Trans. Syst. Man Cybern. Part A Syst. Hum. **30**(3), 286–297 (2000)
24. Shneiderman, B.: The eyes have it: a task by data type taxonomy for information visualizations. In: The Craft of Information Visualization, pp. 364–371. Morgan Kaufmann, Amsterdam (2003). https://doi.org/10.1016/B978-155860915-0/50046-9
25. Espinosa, A., Lerch, J., Kraut, R.: Explicit vs. implicit coordination mechanisms and task dependencies: one size does not fit all. In: Salas, E., Fiore, S.M., Cannon-Bowers, J.A. (eds.) Team Cognition: Process and Performance at the Inter-and Intra-individual Level (2002). https://doi.org/10.1037/10690-006