Olivia Y. Lee
OSPOXFORD 29: AI & Society – Final Paper

The Future of Human-Machine Interaction: Keeping Humans in the Loop

As Artificial Intelligence (AI) capabilities such as machine learning, natural language processing, and deep learning have rapidly evolved in the last decade, so has the idea that they will advance from learning from humans to rendering humans obsolete. In *Superintelligence*, Nick Bostrom warns of superhuman AI systems that pose an existential threat to humans, as do other academics warning of catastrophic risks of powerful AI systems in the imminent future. While such AI takeover scenarios are theoretically possible, they often disregard much of the recent developments in AI as well as the active steps developers and researchers can take to keep humans in the loop of AI development and deployment. The doomsday ending that humans will be demolished in the fierce intelligence competition with AI systems is remarkably enduring. However, it is an incredibly narrow view that fundamentally distracts us from AI's true potential as well as active measures that can be taken in the present day. This paper asserts that a key tenet of AI development going forward should be keeping humans in the loop. It identifies two broad classes of problems where AI will foreseeably be applied: non-immediate decision making (e.g., data analytics and robotics) and time-sensitive, safety-critical decision making (e.g., autonomous vehicles and aircraft). This distinction – that the evaluation of, for instance, AI-assisted determination of whether a tumor is cancerous versus whether to engage the emergency brake in an autonomous vehicle to save a life, is very different – is key to understanding different tools that can be developed to facilitate human-AI collaboration in each case. This paper then evaluates three concrete technical developments that can facilitate active human-machine collaboration and make tangible the notion of keeping humans in the loop of AI development and deployment.

I. Non-Immediate Decision Making: Human-AI Communication and Collaboration

AI technologies are already becoming increasingly integrated into areas such as healthcare, finance, supply chain management, insurance, and other areas that involve complex, high-volume data processing to inform data-driven decision-making. In many cases, AI systems do not make the final call; rather, they are used as tools for pattern recognition and outcome prediction, which are ultimately subject to review by human experts. There are many forms of both automation and augmentation in the present day, most notably in the widespread adoption of AI assistants from dashboards and smart homes to law firms, medical offices, and research labs, and many potential uses of AI have yet to materialize and scale, such as intelligent robotics. In these cases, AI is being incorporated into decision making processes that involves multiple steps over a long horizon, and often where human lives are not immediately at risk. In other words, the AI system has time to consult a human expert, if need be, and human-AI interaction and collaboration is an ongoing process, where AI systems must be adept at interpreting human feedback, just as humans must be able to interpret the outputs of AI systems quickly and clearly.

A concrete example of AI augmenting human capabilities is the application of assistive AI in the healthcare industry for patient supervision. For instance, AI systems can help for patient supervision and monitoring, as it is not feasible for nurses to constantly monitor patients. Smart-sensor technologies which are continually capturing data that are then processed by AI systems can alert healthcare providers when additional attention to a certain patient may be warranted, thus augmenting the work that of human nurses do by freeing up their capacities to handle tasks that technology may not be able to automate or help as much with. In addition to augmentative capabilities, AI can also automate necessary but tedious tasks, like applying machine learning models to processing medical literature or paperwork, thus freeing up time for medical professionals to spend connecting with patients. In this sense, automation can be instead viewed as an indirect augmentation of human capabilities, as it enhances the ability for human workers to engage in parts of their role where human connection and interaction is especially valuable, thus enabling humans to spend their time more meaningfully.

This is obviously an ideal scenario where humans and AI can collaborate seamlessly and AI can be said to truly enhance the capabilities of humans. There is still a lot of research and development needed to overcome the friction of introducing AI systems into workplaces and homes before AI can be said to effectively augment human capabilities and improve efficiency. Crucially, AI technologies need to get to the point where they are *transparent* or *interpretable*, as well as *intuitive*, if such technologies are to have any chance of being successfully integrated into society. With regards to transparency and interpretability, the outputs of the processing done by AI systems and how those outputs came to be generated cannot be opaque to humans, which is a limitation of current "black box" systems. Just as we would expect human collaborators to explain their reasoning motivating certain actions, AI collaborators should be held to a similar standard. This is crucial to building trust between humans and AI systems, which will facilitate their adoption. With regards to intuitiveness, AI systems should behave predictably and be responsive to environmental changes, without becoming an additional obstacle or hindrance to account for. This point may seem trivial but is especially relevant when introducing robots into environments with high human contact like workplaces, schools, and homes. An active area of research in robotics involves interactive robot learning from human feedback, investigating when it is appropriate for robots to stop and request human guidance or intervention. In the ideal case, a robot will operate autonomously and only ask for help when it predicts it will reach an irreversible or unsafe state, but this prediction task is highly nontrivial. At the same time, the robot would become an obstacle or hindrance rather than an asset if it constantly requests human intervention and supervision. Therefore, with more advanced AI technologies, especially embodied AI or AI agents capable of generalizing to diverse tasks, much development is needed to ensure the technology is intuitive and handles tedious tasks without requiring excessive supervisory effort from humans.

Overall, for non-immediate decision making, keeping humans in the loop is a generally uncontroversial view; doing so enables humans to ensure that AI systems function properly and fairly and allows humans to provide insights into human factors that AI systems may not understand or properly account for, while simultaneously reaping the benefits of AI capabilities in complex, high-volume data processing and predictive analyses. The more crucial issue is facilitating clear and effective communication between humans and AI systems, such that the augmentation of AI systems will result in a genuine increase in productivity and efficiency in the allocation of human resources. Much of this will depend on research and development into improving the transparency or interpretability of AI systems and their outputs, as well as ensuring these systems are intuitive and strike a balance between autonomous and interactive operation.

II. Time-Sensitive, Safety-Critical Decision Making: Human Involvement Pre-Deployment

Allowing AI systems to request guidance from human experts or supervisors is all well and good, but there are several situations with significant time constraints that make human intervention difficult or impossible. In such cases of decision making under time constraints, split-second decisions with major consequences, including risking human lives, are to be made. Therefore, AI systems have no time to consult human experts and allowing humans to intervene may actually increase the chances of catastrophic outcomes. Hence, solutions to time-sensitive, safety-critical decision-making problems trend towards fully automated, self-monitoring systems. Clear examples in this class of decision-making problems are autonomous aircraft and vehicles. Much of the following discussion will reference Langwiesche's article "The Human Factor" (2014), which breaks down the catastrophe of Air France Flight 447, as a springboard for broader insights that can be gleaned about keeping humans in the loop in the development of AI systems.

An unfortunate consequence of increased reliance on automation is a decline in human pilot capabilities, which further facilitates increasing automation to reduce the negative impact of human error. In Langewiesche's article, he describes a "paradox" in which "the incoherence of the pilots seems to have been rooted in the very advances in piloting and aircraft design that have improved airline safety". Upon further analysis, this negative reinforcing cycle is hardly paradoxical; it is, in fact, quite natural that increasing automation results in increasing reliance on automated processes and overall, less human intervention. Crucially, increased reliance on automation leads to a decline in the ability of human pilots to handle crises when they do arise. Because of advanced automation, the probability of human airline pilots being faced with crisis has become very low, but it also becomes increasingly unlikely that they can manually handle a crisis if one arises. Langewiesche describes the approach pilots take today as "to keep their hands off the controls, and to intervene only in the rare event of a failure." However, most pilots are incapable of intervening in such failure events, which is unsurprising since even seasoned pilots hardly encounter these situations and often lack sufficient experience to learn from and apply to emergency situations. Furthermore, Langwiesche highlights how automation shifts the pilot's role from active flying to passive supervising: "Once you put pilots on automation, … flying becomes a monitoring task, an abstraction on a screen, a mind-numbing wait for the next hotel". As Boeing's Delmar Fadden stated, "First they have to recognize that it's time to intervene, when 98 percent of the time they're not intervening. Then they're expected to handle the 2 percent we couldn't predict." Based on the series of events that transpired in the cockpit of Air France Flight 447, not only were the pilots' manual abilities lacking, their decreased flight awareness due to high reliance on automation also contributed to the flight's fatal end. Indeed, decreased awareness is a natural consequence of increasingly powerful automation systems.

In the field of AI today, we observe similar issues for handover procedures in autonomous vehicles, where a significant area of concern pertains to how humans should take over when the autonomous vehicle encounters a crisis that it cannot resolve. In Level 4 (high driving automation) or Level 5 (full driving automation) vehicles, it is quite natural that human occupants will not be paying full attention to the vehicle's surroundings. This is often touted as a significant benefit from autonomous vehicles, allowing them to free up time spent driving which can be spent on other more productive tasks. However, in the event of an emergency, one can easily envision a state of utter confusion about what is even happening in the environment, not to mention confusion about what to do to avert the crisis. On the road, passengers will likely have a matter of seconds to handle the emergency. However, those precious seconds will likely be spent merely recognizing that they need to intervene, with no time left for action. Ultimately, complications arising from breakdowns in interactions between humans and autopilot systems demonstrate that human involvement during deployment is risky, due to general human deskilling, naturally lowered operational awareness, and the inherent difficulty of emergency handover from machines to humans.

With these considerations in mind, in systems where split-second decisions with major safety consequences need to be made, it seems to be better to leave such decision-making processes up to self-monitoring automated systems, which can be extensively trained on the ground to account for low probability emergency situations that would take hours of accumulated experience for humans to learn to handle (if they even encounter them at all). Humans can then be kept in the loop in engineering and data collection processes, as well as pre-deployment testing, all of which will be crucial for the successful deployment of automated systems. On the ground, humans can consult the appropriate engineering, manufacturing, and legal experts to develop these systems, and engage in slower, more careful deliberations about how to handle emergency situations and edge cases. In this case, the transparency and interpretability tools mentioned in the prior section – specifically, tools that shed light on the deeper calculations and internal processing of the automated systems – will be less relevant since human operators on-board will likely not have time to fully analyze these detailed reports. During deployment, high-level alerts to communicate or signal to humans when certain actions or events are occurring, for example if a potential collision is near and steps taken to avoid it, will be more helpful than detailed reports to increase transparency. There are certainly difficulties with potential edge cases with limited time for communication, for instance if an animal jumps in front of a vehicle, as well as situations where preprogrammed controls (e.g., an emergency brake) may not always be the best course of action given circumstantial factors (e.g., if there is a truck driving behind the vehicle, if the vehicle is driving on ice, at a bend, or in the middle of the freeway, etc.). Therefore, how the system should respond to such cases will require human discussion and deliberation on the ground as well as thorough edge case testing before system deployment.

III. Achieving Active Human-Machine Collaboration

Having discussed approaches to keeping humans in the loop in non-immediate decision making scenarios (where the emphasis lies in transparency and interpretability of the outputs of AI systems) as well as decision making under time constraints (where the emphasis lies pre-deployment engineering and testing practices), the following section outlines three technological developments in the field of AI safety that can generate tangible progress towards realizing engineering systems that facilitate genuine human-AI collaboration. First, transparency and interpretability tools, which as mentioned prior more pertinent to non-immediate decision-making scenarios. Second, safe and robust exploration systems, which are more relevant to time-sensitive decision-making processes. Finally, monitoring systems, which are applicable to both scenarios, but to different degrees depending on the particular use case.

(a) transparency and interpretability tools

Advances in AI and machine learning have resulted in systems that are incredibly complex; they are now capable of processing numerous streams of high-frequency information and making high-level decisions through complex modeling. As a result of this complexity, AI systems often consist of numerous subcomponents within an integrated system that communicate between each other seamlessly, or more commonly referred to as end-to-end systems, with internal processing that is opaque to human operators. This masks a good amount of the reasoning that a human operator would be expected to provide when making important decisions. As the capabilities of AI continue to increase, humans will likely come to adopt a more supervisory role over AI systems. However, the lack of transparency in such design choices, combined with the rise in complexity, results in humans not knowing how to predict or intervene when AI systems fail. This is relevant to, for instance, dealing with biases that are incorporated into AI systems, but in the future can also be extended to situations like reward hacking and unexpected exploitation of loopholes when AI systems are operating in more complex environments. Interpretability tools, which serve to tease apart what each module or subcomponent of the integrated system is doing, would therefore be immensely helpful in facilitating human-machine interactions that engender trust.

Prior work in the field of transparency and interpretability of AI systems has primarily focused on feature visualization and channel attribution. Feature visualization involves converting abstract vectors of neuron activations into visualizations of neurons weighted by their activations, expressing a neuron's learned activation in terms of human-understandable input (Erhan et al., 2009). This allows us to understand what the network detects and attributes to and from hidden layers in the neural network, which is crucial to increasing the interpretability of the overall system (Simonyan, 2013). Attribution enables us to better understand the relationships between neurons in the neural network, specifically how the network assembles the individual neurons for future decision-making (Zeiler & Fergus, 2014), which is essential for explainability and interpretability of the network. In particular, channel attribution allows us to understand the extent of contribution of each detector to the final output (Kim, 2017). Applying a combination of these tools will improve predictions of the impact of AI systems after deployment, allowing engineers to make the necessary modifications to the system before deployment, for example through interpretability interfaces described in Olah et al. (2018). Such approaches combine building blocks of feature visualization and attribution to allow humans to interpret the input that the network recognizes and how the system's understanding and decision-making process develops with training.

(b) safe and robust exploration

Systems enabling robust and safe exploration will be critical for AI model training, especially for time-sensitive, safety-critical decision-making processes where exposure to edge cases in the real-world is not possible without compromising human safety. The process of exploration in, for instance, agent-like systems that are trained via reinforcement learning, is inherently risky as agents may attempt dangerous behaviors that lead to unacceptable errors in the real world. Simulation techniques can thus refer to either online virtual simulations or physical simulations in a safe testing environment, or a combination of both. The advantage of virtual simulations is that it is easy to reset the simulation to an initialization state, tweak or modify variables in the simulation, and deliberately put the agent in unusual or "edge case" situations to test its response. It is also easy to investigate situations in which the agent responds in unexpected manners. Once the safety of the system has been ascertained to a certain level, a physical simulation that mimics the real application environment could be implemented, so long as the necessary safety and precautionary measures are taken. The advantage of physical simulations is that "randomness" or obstacles that are inherent in real-world situations can be introduced that may not be accounted for in virtual simulations. For this technique to be successfully implemented, one must ensure the proxy simulation environment (both virtual and physical) is similar to what the system will encounter in reality. Techniques such as domain randomization (Peng, 2017) can also ensure less distributional shift between simulation and the real world and prevent the model from overfitting to situations in the training data and focus on the important aspects of the simulation. Such techniques are critical for AI system development to ensure that there is a wide range of coverage of potential states and robustness to low probability or unpredictable scenarios.

(c) monitoring systems

The development of systems that monitor the performance of AI systems post-deployment can help ensure its behaviors are in line with various safety constraints. They also serve to alert human engineers if the system demonstrates any unsafe or unpredictable behavior, at which point the system should be temporarily stopped and retrained. This is especially important for specific edge cases that may not have been captured during training and ensures the system does not respond in an unexpected manner. If any specific pattern is observed among instances when the system fails to respond, the system can be retrained to account for that specific instance. Such systems are much more developed in practical settings, for example in developing autonomous vehicles or trading, thus the safety constraints encoded within the monitoring system are often application specific. For example, for agents developed for algorithmic finance, a monitoring system could encode hard cutoffs that enable algorithms to be stopped immediately when out of distribution. This makes the problem of developing robust monitoring systems less of a theoretical problem and more so a recommended standard implementation for developers of AI systems that are intended to operate autonomously.

There are several broad approaches to the development of monitoring systems. The first approach involves techniques from human-robot interaction research, specifically using human interactions with robotic agents to detect when agents are not behaving as expected or in a safe manner (Najmaei & Kermani, 2011). This approach involves the agent making inferences about the safety of its own actions and behaviors based on the responses of humans that the agent co-exists with. This field remains an active area of research with its own suite of challenges (Alami et al., 2006), thus advances in this field will bring about significant improvements in the development of monitoring systems for robotic agents which physically interact with humans. The second approach is a more algorithmic approach, which involves developing a software that tracks information about an agent's state-action pairs and detects trends in actions where the agent fails or produces undesirable behavior. Human engineers can then identify states or groups of closely related states where the agent generates undesirable behavior and freeze and retrain the model to perform as expected on these (sets of) states. Sufficient progress in the development of monitoring systems will enable them to be incorporated into safe exploration systems discussed in the prior section. The development of an effective oversight agent that detects when the agent in the virtual simulation violates the safety constraints of the environment can be used to identify policies that operate within the constraints, or policies that deviate from the safety constraints during training.

A great deal of attention, arguably too much attention, has been focused on the substitution of human labor with AI algorithms or robots. Using AI to automate human intelligence and labor is an incredibly powerful vision, but also a very narrow one. Instead of automation, shifting our focus to augmentation will enable AI to complement humans to collaboratively tackle difficult tasks more quickly. The future of human-AI augmentation and interaction should therefore focus on keeping humans in the loop. In non-immediate decision-making scenarios, keeping humans in the loop during the deployment of AI systems is a natural outcome, but facilitating integrated human-AI collaboration is highly contingent on the transparency, interpretability, and intuitiveness of these systems, allowing AI systems to become an asset rather than obstacle. In time-sensitive, safety-critical decision-making processes, the inherent difficulty of machine to human handover in emergency situations suggests that keeping humans in the loop during deployment is infeasible. Humans can thus be more actively involved in pre-deployment engineering, development, and testing procedures. Three concrete technical developments that contribute to tangible progress towards keeping humans in the loop of AI development and deployment by facilitating active human-AI collaboration: transparency and interpretability tools (more applicable to non-immediate decision making), safe and robust exploration systems (more applicable to time-sensitive, safety-critical decision making), and monitoring systems (applicable to both and is application-dependent). Overall, envisioning ways that AI systems can operate alongside humans will spur innovation and create opportunities for humans to, with the assistance of AI systems, apply their unique skills and insights to tackle an expanded range of problems.

Bibliography

Alami, R., Albu-Schaeffer, A., Bicchi, A., Bischoff, R., Chatila, R. et al. (2006). *Safe and dependable physical human-robot interaction in anthropic domains: State of the art and challenges*. IROS 2006 Workshop on Physical Human-Robot Interaction in Anthropic Domains. Beijing, China. ff10.1109/IROS.2006.6936985ff. ffhal-01295366

Erhan, D., Bengio, Y., Courville, A. and Vincent, P. (2009). *Visualizing higher-layer features of a deep network*. University of Montreal, Vol 1341, pp. 3

Kim, B., Gilmer, J., Viegas, F., Erlingsson, U., Wattenberg, M. (2017). *TCAV: Relative concept importance testing with Linear Concept Activation Vectors*. arXiv:1711.11279.

Langwiesche, W. (2014). "The Human Factor". Vanity Fair, October 2014. https://archive.vanityfair.com/article/2014/10/the-human-factor

Najmaei, N. & Kermani, M. R. (2011). *Applications of Artificial Intelligence in Safe Human-Robot Interactions*. IEEE transactions on systems, man, and cybernetics. Part B, Cybernetics : a publication of the IEEE Systems, Man, and Cybernetics Society. 41. 448-59. 10.1109/TSMCB.2010.2058103

Olah, C., Satyanarayan, A., Johnson, I., Carter, S., Schubert, L., Ye, K., Mordvintsev, A. (2018). *The Building Blocks of Interpretability*. Distill. https://distill.pub/2018/building-blocks/

Peng, X. B., Andrychowicz, M., Zaremba, W., Abbeel, P. (2017). *Sim-to-Real Transfer of Robotic Control with Dynamics Randomization*. arXiv:1710.06537.

Simonyan, K., Vedaldi, A. and Zisserman, A. (2013). *Deep inside convolutional networks: Visualising image classification models and saliency maps*. arXiv:1312.6034.

Zeiler, M.D., Fergus, R. (2014). *Visualizing and understanding convolutional networks*. European conference on computer vision, pp. 818--833.