2

Human psychology and intelligent machines

Aaron Celaya and Nick Yeung[1]

The cheaper artificial intelligence (AI) and its related technologies become, the more they will become pervasive: they will be employed outside of the domains for which they were initially conceived. As the previous chapters have illustrated, automation will take over more mundane and repetitive tasks, thus making the role of human beings, paradoxically, more important in the decision process. As a result, the benefits of artificial intelligence will depend, critically, on the interaction between humans and machines, first, and on their acceptance. These two topics have little to do with the technology and much to do with the psychology of human beings. Machines can be, in theory, perfectly designed, but they will be of little utility if users have pratical biases in using or trusting them. Similarly, it will be difficult to benefit from intelligent machines if they are not accepted – not an unlikely outcome if we look at the historical record.[2] This chapter provides an introduction to the literature on the psychology of the interaction between human beings and machines, illustrating what aspects practitioners will have to consider – at the tactical, operational and strategic levels.

Artificial intelligence and human psychology

Research, development, and deployment of AI systems are increasing across the whole of society, in contexts as diverse as banking and finance, education, cyber defense, marketing, gaming, medical diagnosis, and security. AI systems are no less on the rise in warfare environments, where we are seeing development in military contexts ranging

[1] The views expressed in this NDC Research Paper are the responsibility of the authors and do not necessarily reflect the opinions of the NATO Defense College, the North Atlantic Treaty Organization, or any other institution represented by the contributors.

[2] C. B. Frey, *The technology trap: capital, labor, and power in the age of automation*, Princeton, Princeton University Press, 2019.

from targeting to intelligence exploitation, battlefield medical response, bomb diffusion, subterranean mapping, and automated Observe-Orient-Decide-Act (OODA) Loop inputs. The increasing influence of AI is widely expected to have a transformational, rather than incremental, impact on military operations. Former NATO Deputy Secretary General Rose Gottemoeller underscored the criticality and urgency of AI when noting that "we cannot fight tomorrow's threats with today's tools […] in the age of artificial intelligence".[3]

However, notwithstanding the many striking successes of AI in recent years – from mastering the game of Go to AI-aided disaster relief response – many of these systems or concepts are still in their infancy. In particular, AI capabilities that have matured to date, even those designed to learn from their experiences rather than execute pre-specified computations, are nevertheless specialized to perform certain, well-defined functions and subsequently remain limited in scope and application.[4] The most immediate goal is thus to develop AI that can mimic the core cognitive capabilities of children, 0-18 months old. The fact that this goal is genuinely ambitious highlights the challenge of developing AI systems that are capable of usefully contributing in contexts requiring judgment, flexibility, and the ability to make informed decisions on the basis of sparse information – properties that characterize most, if not all, political and military contexts.

Two related implications follow from this analysis: first, for the foreseeable future, AI systems will often be deployed in conjunction with human operators and decision makers, rather than acting autonomously. Second, the effectiveness of AI systems will depend critically on their usability and acceptability to humans. For this reason, "realising this potential will depend on understanding the relative strengths of humans and machines, and how they best function in combination to outperform an opponent. Developing the right blend of human-machine teams – the effective integration of humans and machines into our war fighting systems – is the key".[5]

As such, an understanding of human psychology will play a critical role in maximizing the potential of intelligent machines. At a basic level, the foundation of effective human-machine teaming will lie in identifying the respective strengths and weaknesses of human and machine intelligence. The 2018 US Department of Defense's Artificial Intelligence

[3] Remarks by NATO Deputy Secretary General Rose Gottemoeller at the Xiangshan Forum in Beijing, China, 25 October 2018, NATO Newsroom, Speeches & Transcripts Section, https://www.nato.int/cps/en/natohq/opinions_160121.htm (acessed 8 November 2019).

[4] D. Silver *et al.*, "Mastering the game of Go with deep neural networks and tree search", *Nature*, No.529, January 28, 2016, pp.484-489.

[5] *Joint Concept Note 1/18: human-machine teaming*, UK Ministry of Defense, May 18, 2018, https://www.gov.uk/government/publications/human-machine-teaming-jcn-118 (accessed 8 November 2019).

strategy, for instance, directs the use of AI "in a human-centred manner" that recognizes the complementary strengths and weaknesses of human and machine intelligence, with the goal of simplifying workflows and improving the speed and accuracy of repetitive tasks in order "to shift human attention to higher-level reasoning and judgment, which remain areas in which the human role is critical".[6]

However, as AI systems continue to develop in sophistication and reach, they will likely – and perhaps inevitably – move beyond the status of tools to be deployed by people, and instead take on roles as peers (or even the superiors) of human operators: setting goals, developing strategies, drawing inferences, and making meaningful decisions in domains that are currently the exclusive province of human intelligence. As such, with AI systems acting as partners to humans – true human-machine teaming – we foresee a central role for understanding the psychology of trust, teamwork, and communication in effective AI system design.

Human-machine trust

Teamwork relies on trust: the willingness of individuals to work interdependently, such that they accept roles in which they are vulnerable to the actions of other team members with the expectation that those others will act accordingly.[7] Existing research on human-machine teaming highlights the complexity of optimizing human users' trust in artificial systems. This challenge is particularly evident when errors in the system have been observed by the human user, with evidence of these users veering from irrational aversion to irrational overtrust in artificial systems.

Consistent with a longstanding observation that, though statistical prediction outperforms clinical or human-only prediction in many situations, people still prefer to get advice or diagnoses from other people rather than expert systems,[8] recent experimental research found that human decision makers consistently downweight information provided

[6] *Summary of the 2018 Department of Defense Artificial Intelligence Strategy*, US Department of Defense, 2018, https://media.defense.gov/2019/Feb/12/2002088963/-1/-1/1/SUMMARY-OF-DOD-AI-STRATEGY.PDF (accessed 8 November 2019).

[7] R.C. Mayer *et al.*, "An integrative model of organizational trust", *Academy of Management Review*, Vol.20, No.3, 1995, pp.709-734.

[8] P. Meehl, *Clinical versus statistical prediction: a theoretical analysis and a review of the evidence*, Minneapolis, University of Minnesota Press, 1954; W. Grove *et al.*, "Clinical versus mechanical prediction: a meta-analysis", *Psychological Assessment*, Vol.12, No.1, 2000, pp.19-30.

by intelligent machines such as expert systems and algorithm-based prediction systems.[9] For example, in one study, participants were asked to forecast the success of applicants to an MBA program (based on the information available in their file), and were then given the option of using forecasts from another person (a previous participant in the experiment) or a computer algorithm (a statistical prediction model). Absent any experience with forecasts of either the model or the past participant, most people preferred to use the algorithm's forecasts. However, when given experience with the statistical prediction model, or with both the model and the other person's forecasts, participants were consistently less likely to choose the model, even when the algorithm clearly outperformed both the participant themselves and human advisor. Apparently, such "algorithm aversion" arises because people rapidly lose trust in algorithms after seeing them err, but seem much more forgiving of human error.

Other works, however, have documented another problem, namely dramatic cases of overtrust in artificial systems.[10] Overtrust occurs when too much risk is accepted by the human user with the belief that the autonomous or AI counterpart will make up the difference and mitigate the accepted risk. In some recent studies, participants were willing to trust an emergency guide robot, even when they had observed it malfunction previously. In contrast to the works discussed before, human users were apparently very quick to forgive a robot for previous mistakes and to trust them even to the point of failing at their overall objective. One explanation is that users did not possess an accurate assessment of the robot's capabilities and accuracy. Another explanation holds that people just followed the robot's advice/directions simply because the robot stated that it was designed for the task, thus suggesting that designation mattered more than performance.

These studies identify divergent patterns of aversion versus overtrust, but converge in showing that human agents can learn incorrect models of machine trust and therefore use highly sub-optimal machine reliance strategies. The implication is straightforward: developing AI systems that can be trusted appropriately by human operators is of utmost importance. Our working hypothesis is that this challenge can be addressed by applying established principles of trust in normal human-to-human interactions. The concepts in human relationships form the foundation for those applied in human-machine teams.[11] For

[9] B. Dietvorst *et al.*, "Algorithm aversion: people erroneously avoid algorithms after seeing them err", *Journal of Experimental Psychology: General*, Vol.144, No.1, 2014, pp.114-126.

[10] P. Robinette *et al.*, "Overtrust of robots in emergency evacuation scenarios", *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, Christchurch, 2016, pp.101-108.

[11] J. Jian *et al.*, "Foundations for an empirically determined scale of trust in automated systems", *International Journal of Cognitive Ergonomics*, Vol.4, No.1, 2000, pp.53-71.

example, human decision makers exhibit egocentric bias, discounting advice from other people relative to their own opinions. Further, human decision makers are quick to form opinions of their advisors which are updated asymmetrically: trust is built slowly, but lost rapidly when advice turns out to be incorrect.[12] Human decision makers are also likely to be influenced more by advice they request and to discount advice less when the task is complex. Alternatively, human decision makers discount advice more as the difference between their initial judgment and the advisor's opinion increases.[13]

In our research, we apply such findings and theories of human trust as our framework for exploring trust in human-machine teams. A cornerstone of our approach characterizes the relationship between *trustor* and *trustee* as depending jointly on properties of each individual and the interaction between them (see Figure 1). The trustee is characterized by three traits:

- ability – the skills and competence of the trustee in their assigned tasks;
- benevolence – the extent to which the trustee is believed to want to do good to the trustor; and
- integrity – whether the trustee adheres to a set of principles that the trustor finds acceptable.

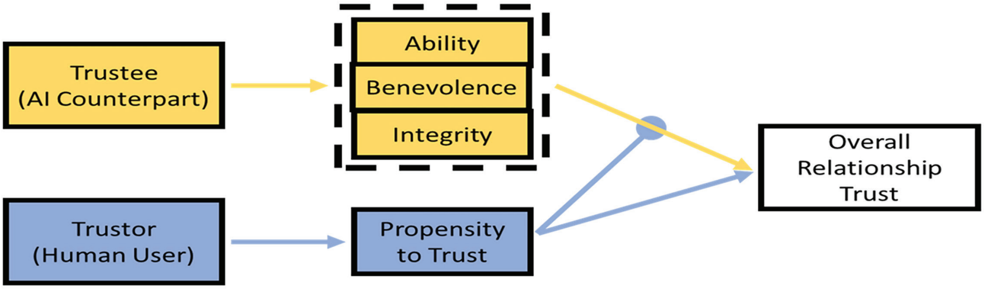


Figure 1: Trust model adapted from Mayer *et al.* (1995)

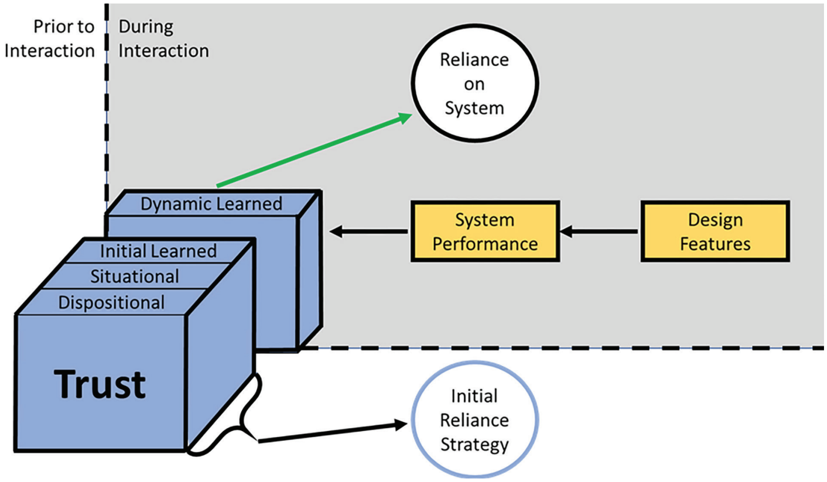
[12] I. Yaniv and E. Kleinberger, "Advice taking in decision-making: egocentric discounting and reputation formation", *Organizational Behavior and Human Decision Processes*, Vol.83, No.2, 2000, pp.260-281.

[13] S. Bonaccio and R. Dalal, "Advice taking and decision-making: an integrative literature review, and implications for the organizational sciences", *Organizational Behavior and Human Decision Processes*, Vol.101, No.2, 2006, pp.127-151.

The trustor, meanwhile, is characterized by the propensity to trust: individuals may vary in their baseline levels of trust in others, and also more subtly in the relative weight they place on ability, benevolence and integrity as critical to winning and retaining their trust.

Hoff and Bashir propose that trust is simpler in the context of human interactions with machines, being determined only by the ability of the machine and the human's propensity to trust (i.e., with benevolence and integrity of the machine being less important – a point we return to below).[14] Within this simplification, their model usefully elaborates how individuals vary in their propensity to trust: this propensity depends on the individual's general levels of trust in others (*dispositional trust*) that varies according to variables such as culture, age, gender or personality, as well as more context-specific factors (*situational trust*) relating to the current state of the person (e.g., their mood, their current workload), and to the situation in which they find themselves (e.g., the particular system they are working with, the complexity of the task being solved). Together, these dispositional and situational factors form an initial reliance strategy (see Figure 2). This strategy is then modified through interaction according to the performance and behavior of the machine (synonymous with ability in Roger C. Mayer *et al.*'s model), specifically how well it performs its given task. This dynamic learned trust continues to evolve over the course of system interaction and determines the trustor's reliance on the system.

Figure 2: Trust model adapted from Hoff & Bashir (2015)



[14] K. Hoff and M. Bashir, "Trust in automation: integrating empirical evidence on factors that influence trust", *Human Factors,* Vol.57, No.3, 2015, pp.407-434.

Current research efforts

We are conducting experimental work within these theoretical frameworks of trust, exploring human-machine teaming in tasks with requirements similar to those of operationally-relevant perceptual judgments (e.g., "Is there a surface-to-air missile site in this piece of satellite imagery?"; "Are there subtle changes over time in this region of interest?"). Our aim is to more fully understand trust in AI systems, as they depend on individual user differences (i.e., propensity to trust) and the properties of the AI systems (i.e., ability). In the experiments, human participants make a series of visual judgments, with the option of receiving and using advice from another person (a previous experimental participant) or a simple AI system (computer algorithm). We measure trust in terms of our participants' choice of advisors and the relative influence that advisors have on participants' ultimate decisions.

Figure 3: Correlation of advisor influence and choice

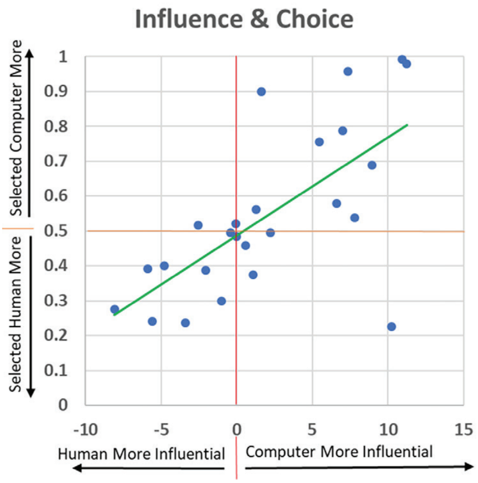


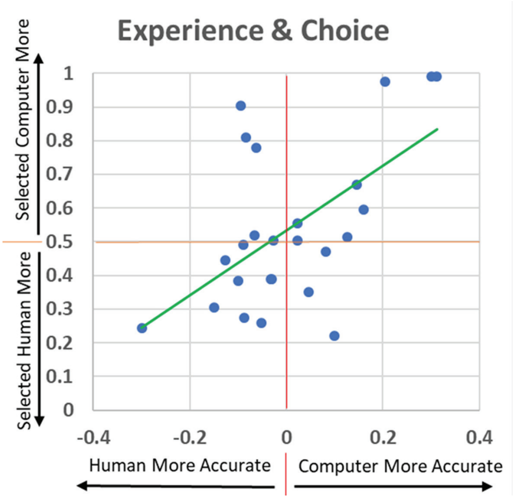
Figure 3 presents illustrative findings from one of our experiments, in which we controlled for system performance, design features, and initial learned information – all participants saw the same advice quality from the two (human and computer) advisors – in an effort to isolate and study individual differences in propensity to trust. In the figure, each point corresponds to one of our experimental participants, with the x-axis value indicating whether each participant was more influenced by human vs. computer advice, and the y-axis value indicating whether they were more likely to choose human vs. computer advice. The results reveal neither systematic aversion nor overtrust in AI, in contrast to the studies discussed before: the data points are evenly distributed either side of the red neutral midpoint, and above/below the orange neutral midpoint of each axis. However, large idiosyncratic individual differences are clearly apparent. Even though human and computer advice was objectively equally accurate, some participants were more influenced by the human advisor and chose him/her more (data points clustered to the lower left of the figure), whereas other people were more influenced by and chose the computer advisor more often (data points clustered in the upper right). Therefore, our participants are rational insofar as they choose the advisor that most influences them, but are characterized by substantial (irrational) variation in their preference for one advisor over the other. Leveraging these findings, we are now conducting experiments to understand what dispositional and situational factors drive these preferences with a view to developing profiles of individuals' trust in AI.

Other experiments have allowed us to characterize the dependence of human-machine trust on observed ability of the system. For example, in line with evidence that human-human trust is rapidly lost and only painstakingly rebuilt after bad advice, we observe that early experience with advisors is critical when determining appropriate reliance strategies (Figure 4).[15] In this experiment, the average accuracy of the human and computer advisors was perfectly matched across the whole experiment. However, from trial to trial there is natural variability in advisor ability as perceived by the participant, reflecting the uncertainty inherent in difficult judgments. Variability in the first interactions with advisors turned out to have long-lasting effect on participants' trust. If a participant's initial experience was of the human advisor being more accurate than the computer, they tended to choose that advisor for the remaining interactions (data points clustered to the lower left of the figure), and the same was true if initial interactions favored the computer advisor (data points towards the upper right). This over-reliance on one advisor resulted in better perceived performance by that advisor, as predicted by Kevin A. Hoff and Masooda Bashir.[16] Translated into practical implications, these findings indicate that training human operators on new systems might benefit from demonstrating system performance via simulation and classroom instruction. Without this component, the human operator could form an initial reliance strategy based on unrepresentative early experiences that are different from what will be developed later, after real-world system interaction.

[15] I. Yaniv and E. Kleinberger, *Advice taking in decision making*, pp.260-281.

[16] K. Hoff and M. Bashir, *Trust in automation*, pp.407-434.

Figure 4: Correlation of early experience and choice



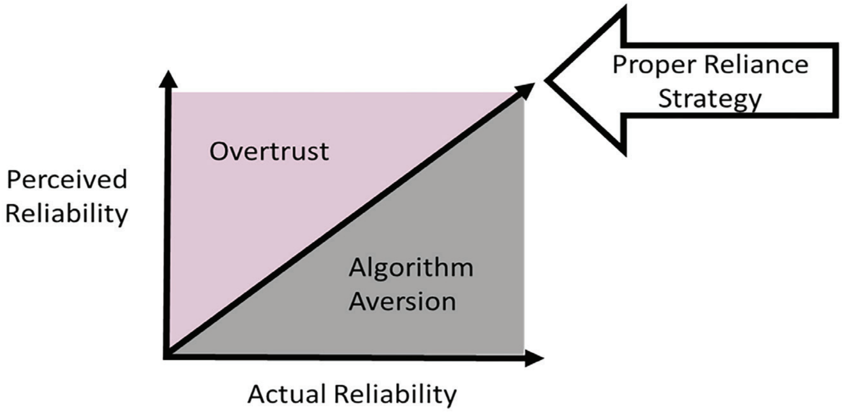
Outlook

To the extent that AI systems remain embedded in human-machine teams rather than acting fully autonomously, understanding human psychology will be critical to maximizing the potential of these emerging technologies. Here we have focused on one example: applying principles of human trust to ensure appropriate reliance on machines. Within Keith Gempler's[17] depiction of the relationship between perceived and actual system reliability (Figure 5), we have seen how human operators deviate from proper reliance strategies as a function of their disposition, pre-conceptions and experiences, such that they exhibit aversion (underestimation of machine reliability) or overtrust (overestimation of reliability) in machine partners. These findings indicate the importance of system designs that mitigate against these human biases, which can again be informed by known psychological principles of trust maintenance and repair. For example, in human teams, communication of confidence and uncertainty is crucial to both effective group decision making and trust calibration, suggesting that effective human-machine teaming may depend on the development of AI systems that know (and signal) when they might be wrong.[18]

[17] K. Gempler, *Display of predictor reliability on a cockpit display of traffic information*, Urbana, University of Illinois at Urbana-Champaign, 1999.

[18] B. Bahrami *et al.*, "Optimally interacting minds", *Science*, Vol.329, No.5995, 2010, pp.1081-1085; N. Pescetelli and N. Yeung, *The role of decision confidence in advice-taking and trust formation*, 2019.

Figure 5: Reliability diagram adapted from Gempler (1999)



Looking further into the future, we foresee the importance of looking beyond simplified characterization of human-machine trust as dependent on human propensity to trust and machine ability alone,[19] and instead embrace a broader conception of trust that includes benevolence and integrity as key characteristics.[20] Thus, as AI systems grow in flexibility and autonomy, by necessity they will come to set goals and develop strategies that are independent of the original intentions of the system designers (as we are already seeing in David Silver *et al.*'s algorithm for the game Go, which taught itself super-human strategies that comprehensively outmaneuvered the very best human players of the game). Moreover, the goals and strategies developed by AI systems may be imperfectly aligned with those of their human partners. As AI systems become intentional agents in this manner, the benevolence of these systems towards human team members, and their integrity in terms of operating by principles that can be understood and endorsed by humans in the loop (as subordinates, peers and controllers of these systems) will be critical factors in assuring appropriate levels of trust and control in human-machine teaming.

[19] K. Hoff and M. Bashir, "Trust in automation: integrating empirical evidence on factors that influence trust", *Human Factors,* Vol.57, No.3, 2014, pp.407-434.

[20] R.C. Mayer *et al.*, "An integrative model of organizational trust", *Academy of Management Review*, Vol.20, No.3, 1995, pp.709-734.