DDoD: Dual Denial of Decision Attacks on Human-AI Teams

Tag, Benjamin; van Berkel, Niels; Verma, Sunny; Zhao, Benjamin Zi Hao; Berkovsky, Shlomo; Kaafar, Dali; Kostakos, Vassilis; Ohrimenko, Olga

Published in:
IEEE Pervasive Computing

DOI (link to publication from Publisher):
10.1109/MPRV.2022.3218773

Publication date:
2023

Document Version
Publisher's PDF, also known as Version of record

DDoD: Dual Denial of Decision Attacks on Human-AI Teams

Benjamin Tag, *University of Melbourne, Parkville, VIC, 3010, Australia*

Niels van Berkel, *Aalborg University, 9220, Aalborg, Denmark*

Sunny Verma, Benjamin Zi Hao Zhao, Shlomo Berkovsky, and Dali Kaafar, *Macquarie University, Macquarie Park, NSW, 2109, Australia*

Vassilis Kostakos and Olga Ohrimenko, *University of Melbourne, Parkville, VIC, 3010, Australia*

Artificial intelligence (AI) systems have been increasingly used to make decision-making processes faster, more accurate, and more efficient. However, such systems are also at constant risk of being attacked. While the majority of attacks targeting AI-based applications aim to manipulate classifiers or training data and alter the output of an AI model, recently proposed sponge attacks against AI models aim to impede the classifier's execution by consuming substantial resources. In this work, we propose dual denial of decision (DDoD) attacks against collaborative human-AI teams. We discuss how such attacks aim to deplete both computational and human resources, and significantly impair decision-making capabilities. We describe DDoD on human and computational resources and present potential risk scenarios in a series of exemplary domains.

Intelligent machines have increasingly been integrated into human teamwork settings. Besides substituting human workers, artificial intelligence (AI) systems are deployed at a large scale to support humans.[a] The efficiency and endurance of such machines effectively complement human capabilities. Therefore, rather than substituting workers with AI, an increasing number of companies are integrating it in collaborative human-AI teams,[1] to enable teams to work more effectively, improve decision-making, and increase productivity. By some accounts, the economic impact of this development is projected to add 13 Trillion USD to the global economy over the next 10 years.[b]

Unfortunately, such hybrid teams provide new attack vectors. While traditional attacks on humans often took the form of social engineering to steal credentials or gain secure access, we now have to consider attacks that target human-AI teams aiming to deplete resources and impair their combined performance.

In the machine learning (ML) domain, recent research has investigated the impact of the so-called sponge attacks on ML classifier performance.[1] These attacks aim to occupy computational resources and obstruct the model's behavior and availability. In contrast to adversarial attacks that aim to distort predictions, sponge attacks aggravate decision-making by wasting resources while appearing disguised as rightful requests.

In this article, we introduce dual denial of decision (DDoD) attacks by extending the notion of sponge attacks against human resources and advocate for an increased awareness of such attacks due to the risks they introduce to collaborative human-AI teams. Such risks include confusing humans and presenting them with unclear choices, as well as flooding them with inconclusive classification outputs produced by the machine. This consequently results in a high demand for cognitive resources, increased cognitive load (CL), and saturated attention, making humans more susceptible to other attacks, mistakes, and decreased performance.[2]

[a]2020 Deloitte Global Human Capital Trends, https://www2.deloitte.com/content/dam/Deloitte/cn/Documents/human-capital/deloitte-cn-hc-trend-2020-en-200519.pdf, last accessed Sep. 27, 2022.
[b][Online]. Available: https://hbr.org/2019/07/building-the-ai-powered-organization, last accessed Mar. 29, 2022.

Digital Object Identifier 10.1109/MPRV.2022.3218773
Date of publication 27 February 2023; date of current version 27 March 2023.

The human–computer interaction (HCI) community has intensely worked to understand effective human-AI collaborations better, focusing on explainability, transparency, and fairness. However, the robustness and efficacy of such human-AI collaborations is yet to be fully explored. To unpack the potential impact of DDoD attacks on human-AI teams, we analyze a series of traditional human-AI collaboration scenarios and discuss the potential implications and risks of these attacks on their performance. To this end, we discuss implications for HCI, AI, pervasive computing, and the increasingly popular human-AI research. We draw on the body of work in cybersecurity and human-AI interaction, paving the way for new research focusing on better protecting human-AI teams.

BACKGROUND

We distinguish between three types of attacks that influence the performance of an ML model: Adversarial, poisoning, and backdoor attacks. The main difference between these attacks is the attacker's capability, i.e., where in the ML pipeline the attacker interferes: during training, inference (i.e., when a trained ML model is deployed to process data and produce predictions), or in both phases.

Adversarial attacks harness benign samples that include small perturbations (e.g., noise on an image) to drastically alter a model's predictions.[3] An adversarial attack is executed at inference time without the need for involvement during training, instead exploiting inherent inconsistencies around the decision boundary from the training process. Such perturbations are typically imperceptible to humans, while they effectively deceive the model.

Poisoning attacks manipulate a small proportion of the data used to train a model to significantly reduce the model's prediction performance on any input data.[4] Attackers can compromise a training dataset by submitting poisoned data at crowdsourcing, planting poisoned samples for data crawlers to collect, or contributing to training data directly (e.g., sending malware or spam emails to a system that collects any inputs it receives to retrain the models and stay abreast of evolving threats).

Backdoor attacks, similar to poisoning attacks, have access to the training phase of the model, where an adversary can teach the model to behave in a predetermined manner when a certain trigger is presented at runtime.[5] These triggers have evolved from static shapes to invisible noise, while also taking real physical forms in the world to fool image-based applications.

All the abovemetioned attack vectors have been studied extensively. However, typically these attacks are considered against a standalone ML model without humans, unlike collaborative human-AI teams.

Humans as the Weak Link

When a human is part of the decision-making process, potential attacks on humans provide a new gateway for attackers to compromise collaborative human-AI teams. Social engineering is a well-known technique where the human is attacked directly to obtain control of a system (e.g., by enticing the user to follow fraudulent links in emails), leading to phishing, ransomware, or malware attacks. Social engineering describes the psychological manipulation of system users into unknowingly disclosing sensitive information and performing detrimental actions on behalf of the attacker, within the system.[6] While machines can warn humans of such attacks (e.g., spam filters, system permission requests), the increasing deployment of human-AI teams still presents underexplored avenues for attacks.

We note that while humans can leak private information (e.g., through phishing), complementary privacy attacks exist, whereby information about an ML model's training data is leaked.[7] However, in this article we focus on a new class of attacks against the integrity and robustness of the model, specifically attacks that change a model's behavior at inference or classification time, i.e., when the model is being deployed, and the impact such attacks can have in collaborative human-AI teams.

Sponge Attacks

Sponge attacks against ML models, introduced in Shumailov et al.'s work,[1] do not seek to compromise the predictive accuracy of the model, such as adversarial, poisoning, and backdoor attacks. Instead, a sponge attack provides inputs at inference time to drain the model's resources as much as possible, e.g., maximize the energy consumption and latency of inference by increasing the number of arithmetic operations or memory accesses required to process the input.

The effectiveness of sponge attacks depends on the model, with natural language processing (NLP) tasks shown to be particularly susceptible to such attacks. Shumailov et al. demonstrated an attack on Microsoft Azure's translator, a real-world NLP system, to show an increase in response time from 1 ms to 6 s.[1] Unlike denial of service attacks, sponge attacks increase the resource consumption of a system without increasing the number of requests to the system, thereby circumventing rate-limiting defenses. It is evident that sponge attacks

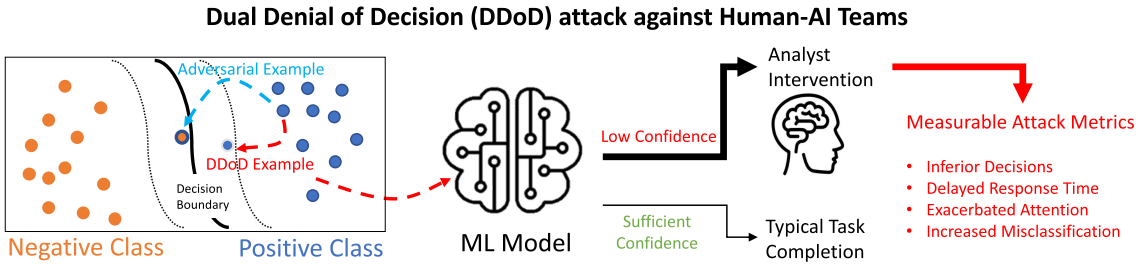

FIGURE 1. Schematic representation of a DDoD attack against human-AI teams. In contrast to an adversarial attack (light blue) a DDoD example does not change the class of a data point, but rather moves it into a space of low classification confidence.

have the potential for widespread negative effects on the availability and performance of ML models.

Along similar lines, Boucher et al. considered sponge attacks in the context of text-based ML models and identify "imperceptible perturbations" as an attack vector that is highly challenging to spot for humans.[8] They demonstrate how text can be crafted such that it appears legitimate to humans, but deceives computers. The authors outline four means to achieve this: invisible characters (e.g., zero-width spaces), homoglyphs (characters that look almost identical—Cyrillic and Latin "A"), reorderings (use of control characters to alter the rendering order), and deletions (use of control characters to conceal characters within strings). Boucher et al. presented how a sponge attack can be deployed against search engines, machine translation systems, and NLP models, with potential consequences including the ability to bypass content detection systems and negatively impact training data. While their examples feature NLP applications, such sponge attacks can be extended to other applications, such as image processing or malware analysis.

Sponge attacks have the potential to deceive machines while the rest hidden from humans. However, it remains to be verified whether such examples could be crafted not only to hide from the human but to concurrently influence both the machine and the human. As such, the ability and effectiveness of human-AI teams to sustain such attacks is a high-priority study area.

Human-AI Teams

The way we design systems to enable successful collaboration in human-AI teams has attracted increased interest in both the HCI and AI communities. Developing an understanding of end-user's expectations toward effective and efficient AI-powered collaborative agents is an active research challenge. For example, Zhang et al. studied the expectations of AI team members in a gaming context and describe that user expectations focus primarily on the AI's technical abilities, to provide the means to develop a shared understanding between the human user and the AI, and effective communication strategies.[9]

Integrating AI support in a real-world application requires a thorough understanding of, and adjustment to, the context in which it is deployed.[10] Examples of domains in which collaborative human-AI systems have been studied include data science,[11] clinical domains,[10] and creative tasks.[12] Despite diverse application domains, a common thread across these studies is the conflict between the desire to comprehend the AI's suggestions and choices (often under the label of "explainability") while simultaneously avoiding undesired interruptions to the user while completing their tasks.

DDOD ATTACKS ON HUMAN-AI TEAMS

To better differentiate the characteristics of this attack from traditional sponge attacks and capture the impact on human-AI teams, we introduce the term: DDoD attacks.

While sponge attacks were originally conceived to drain computational resources,[1] here we extend this notion to DDoD attacks, to be carried out against collaborative Human-AI teams to drain human resources. As presented in Figure 1, a DDoD does not change the class of a data point, but rather moves it into a space of low classification confidence. Consequently, the AI calls in the human collaborator to make the final decision, which may be biased by the low confidence provided by the classifier. Hence, it is important to articulate the range and variety of human-AI teams that can be affected. First, it is essential to consider who is the ultimate decision maker in a human-AI team: a human, or a computer? Second, it is relevant to identify the type of partnership within the team: is it monitoring and collaboration, or instruction?

These two aspects help us identify a variety of human-AI teams: 1) Human supervises computer, such as social media platforms moderation and filtering, airport security check, and passport control; 2) computer supervises human, such as antivirus software and driving safety systems; 3) human controls computer, such as driving and remote-controlled UAVs; and 4) computer controls human, such as warehouse employees picking up items, formfilling, and logins.

In this article, we focus on asymmetric human-AI teams, i.e., where the human is the ultimate decision maker aided by AI decision support. This is because such systems can easily scale up the computational resources, but not so much the human resources. Furthermore, depending on the type of Human-AI partnership, the attack can be either indirect (during which the human steps in to resolve issues related to the AI) or direct (whereby the attack aims specifically at deceiving and misleading the human). Due to their deceptive nature and ability to target systems as well as humans, we posit that DDoD attacks can substantially deteriorate the performance of human-AI teams.

Traditionally, social engineering attacks have long targeted humans as the weakest link in many computing systems, taking advantage of the scarcity of human cognitive resources. Short-term (i.e., working on timescales of minutes–hours) cognitive factors, such as vigilance (sometimes synonymously used with sustained attention), cognitive workload, and stress directly impact human susceptibility to fraud, deception, and distort their decision-making ability.[13] We consider those as the primary DDoD attack vectors on human-AI teams.

Vigilance describes the phenomenon of a fluctuating cognitive performance. This usually means that the longer a task demands cognitive effort, the more cognitive performance declines. Prior research has shown that performance significantly declines over tasks lasting 30–60 minutes,[13] making the human more prone to being less attentive to signs of fraud or deception.

Closely related to vigilance and often interrelated, CL describes the cognitive demands tasks put on the performers. These demands mainly depend on the task complexity (intrinsic CL), the format of the information presented (extrinsic CL), and the person's processing effort (germane CL).[14] The sum of all three CL components describes the total CL of a human. Rather than diverting attention, an attacker could exhaust the human with undue CL by interfering with one of the three types or a combination thereof.

Finally, changes in task load, declining attention, and unprecedented task demands (e.g., output from an attacked classifier) can lead to increased acute stress in the human. While stress in short bursts can be beneficial,[13] as it heightens attention, it often leads to an increased focus on the stress-inducing factor, and thus, to fewer attentional resources being available for other tasks and information.

We bring all these elements together by presenting a number of illustrative examples, reflecting a range of application areas to exemplify potential DDoD attacks on human-AI teams.

Scenarios

In this section we present a set of examples of established human-AI teams. Often mentioned in this regard is the control problem, which describes the failure of a human operator to detect malfunctioning machines due to complacency or overreliance. Until now, it has been recommended to install collaborative human-AI teams to counteract the control problem.[15] However, our examples of DDoD attacks show that human-AI teams present new vulnerabilities.

Medical Diagnosis

AI methods have been increasingly applied in clinical environments.[10] These are deployed for diagnostic, prognostic, and therapeutic purposes, primarily serving as a decision-support tool. That is, the AI does not have the authority to diagnose patients or determine treatment, but the output of the AI instead provides advice to a human clinician, who makes the decisions. This demonstrates an example of a collaborative human-AI team, where an AI can assist the diagnosis process (e.g., by proposing specific tests for an accelerated determination of a medical condition or minimizing unnecessary testing).

For example, as a decision-support tool for a human decision-maker a medical imaging AI can interpret images or videos and detect disease-specific symptoms. These may be highlighted in the images to streamline diagnostic decisions. A DDoD attack in this setting may entail increasing the uncertainty of the predicted diagnosis and highlighting wrong parts of the image. The latter will cause the clinician to waste precious time examining irrelevant parts of the image and potentially ordering unnecessary tests to reach a clear diagnosis. In the worst case, the increased uncertainty on the AI side may result in misleading the clinician in their diagnostic approach. Consequently, the clinician could go for the wrong tests losing crucial treatment time. Moreover, a lack of confidence expressed by a usually well-working AI, may lower the confidence and confuse the human. The clinician could call in for additional human support, which may be needed elsewhere.

Law Enforcement

Law enforcement often has to prioritize how to dedicate the limited human resources to enforce compliance. This has consequently led to the deployment of automated detection systems, supervised by humans, especially in the domain of traffic offenses.[16] Speeding, running red lights, and phone use while driving can be identified and captured by cameras. In most cases, an automatic plate recognition system would transcribe the plate number and issue an infringement to the registered vehicle owner. However, with varying environmental conditions, this recognition task may come with uncertainty requiring a human operator's intervention.

In this scenario, a DDoD attack would entail perturbing the license plate to create uncertainty forcing a review by a human. For example, a sticker could be affixed on the car to cause the classifier to erroneously detect multiple license plates, or be uncertain about the car's actual license plate. In cases where the perturbation of the license plate image is to a level that the human is not able to clearly detect the correct alphanumerical combination, offenses may go unpunished. In large jurisdictions this may occupy a large number of human eyes, which will then be missing at other ends.

Passport Control—Immigration

Another example is an immigration facility, e.g., at airports, where passengers are processed by an automated border control system, verifying the passport's authenticity, chip and biometric information. If any of these elements are in doubt, the individual is referred to a human border control officer.

A DDoD attack on this system would seek to overwhelm the human officers by diverting bulk amounts of individuals away from passing through the automated system. These systems capture facial information with a camera, and in the process will also capture background information. Hence, one potential attack vector is the hijacking, or purchasing of electronic ad spaces to display adversarial/sponge examples to influence the biometric operation, or creating a backlog of passengers that require human management, thereby diverting security resources away from other sensitive areas. This may also delay processing, which can lead to increased stress among passengers and security personnel.

Semiautomated Driving

Semiautomated driving has emerged as a desirable feature and has been becoming increasingly prevalent in commercial vehicles. Its ability to perform well in both regular and atypical settings, while ensuring the safety of passengers and the public, are critically important features allowing to distinguish a market-leading system from competitors. If the AI senses uncertainty about navigating safely, the control is typically relinquished to the human driver.

A DDoD attack would involve purposefully introducing uncertainty into the AI such that control frequently falls back to the human operator, even if not necessary. For instance, stickers or other visual decoys may be placed across a city, such that cars frequently handover control to the human driver, thereby degrading the perceived quality of such a system, and posing an additional burden to human drivers. This would not only impact comfort but potentially pose health risks to passengers as well as pedestrians, if the AI indicates a potential obstacle or problem that the driver has to immediately react to. In seemingly benign but high-speed situations this can have fatal consequences. Moreover, if the car frequently relinquishes control to the driver in such situations, drivers may lose trust in the system and refrain from using it.

Lending and Insurance

Lending and Insurance companies have an inherent tolerance to risk. With AI, additional details about a customer's spending and behavioral habits can be automatically analyzed to illustrate reliable risk profiles.[17] An output may indicate the size of payments, or if the customer should be accepted at all. However, such outputs may be provided to a supervising human expert who makes the final decision.

For example, consider an emergency loan program initiated to support businesses and individuals through an environmental disaster. Many such applications may be automatically processed to perform checks and balances to ensure only qualified parties are offered funding. If a human is integrated into the loop for the final decision, or to handle uncertain qualification status, this presents an opportunity for an attacker to waste the cognitive resources and time of these humans, thereby delaying the processing for legitimate applicants or rendering them illegitimate or questionable. For instance, certain keywords related to risk factors could be introduced into applications to increase the uncertainty of the classifier, which in turn flags more applications for human inspection. In an emergency situation, such as mentioned previously, this can lead to urgent cases not finding a human expert, thus, delaying potentially vital payments.

Recommender Systems

Recommender systems are a type of human-facing AI deployed in many online scenarios,[18] where human users may be faced with information overload. For

example, in scenarios, such as online shopping, booking accommodations, selecting a movie to watch, or deciding on a restaurant to dine, users may find tens to thousands of appropriate options and, thus, struggle to find the optimal one. Recommenders are capable of helping the users by filtering out the least appropriate options and presenting a list containing a small number of best-matching items.

One could conceive a DDoD attack that, instead of listing the best options, presents a list of highly similar items having a comparable probability to be selected by the user. In this attack, the compromised AI only increases the choice difficulty, as the burden of thoroughly examining the recommended options, identifying the subtle differences between them, and making the final selection is put on the user. This is likely to increase both the cognitive demand associated with the decision and the decision-making time.

IMPLICATIONS AND CONSIDERATIONS

While existing work on cybersecurity has primarily focused on technical flaws and attack surfaces, the introduction of human-AI teams broadens the vector for attacks. Human vulnerabilities, such as high CL, high degree of stress, low degree of attentional vigilance, poor domain knowledge, or lack of experience, make us more susceptible to attacks.[13]

Cognition: In the literature, various definitions of fatigue, alertness, and performance can be found and differences of opinion exist about their meaning. For our purposes, the term performance comprises cognitive functions ranging in complexity from simple psychomotor reaction time to logical reasoning, working memory, and complex executive functions. Fatigue refers to subjective reports of loss of desire or ability to continue performing. Alertness is a human resource that can be negatively influenced with carefully crafted sequences of inputs. One can see how an attack that leads to increased demand for human decision-making can quickly deplete alertness. Since alertness is seen as a combination of selective attention, vigilance, and attentional control,[19] it describes the readiness to respond to stimuli and plays a role in higher cognitive functions affecting productivity, decision-making , and memory. Cognitive performance significantly declines for tasks lasting 30–60 minutes.[13] Depleted levels of alertness, thus, lead to decreased productivity, slower or inhibited decision making, and an increased likelihood for mistakes. This can have particularly dangerous consequences, especially in situations where immediate reaction is required (e.g., traffic situations).

Trust: To safely integrate AI in organizations and society, it needs to be trusted by its users. A core component in building trust is the anticipation of intentional behavior in an entity, and the ability to predict (parts of) the decisions made by this entity and their impact.[20] Thus, trust guides the reliance on the output produced by increasingly complex AI that take on tasks that have a high cognitive demand on humans (e.g., predicting an illness within a split of a second based on thousands of images). This often happens without fully understanding the workings of the system. Consequently, when a system under a DDoD attack produces a large amount of results with low confidence, that can be either correct or wrong, the reliance on the system's performance deteriorates. The consequences of this are 1) trusting an unreliable system, which can result in false decisions, or 2) losing the advantages such a system has provided by not relying on it.

To increase end-user trust in autonomous decision-making, AI research community has made extensive efforts to improve the explainability of AI systems. While there is a great diversity in the intended use of such explanations (e.g., global versus local), explanations all face a similar challenge: they can interrupt a user's ongoing task and demand time and mental energy.[2,10] As such, merely including explanations for the recommended actions is insufficient to support a human-AI team in time-critical tasks, such as averting DDoD attacks. Therefore, identifying the most relevant recommendations and actions is a critical component of making explainability meaningful in human-AI collaboration.

Countermeasures: Monitoring DDoD attacks on human-AI teams, therefore, requires not only an assessment of system resource consumption, but also continuous and real-time monitoring of the human resource availability. The pervasive and ubiquitous computing community has provided a plethora of solutions in this domain. Through environmental and wearable sensors, CL, attention, and stress can be detected. Prior work has shown that these data can be obtained unobtrusively.[21]

Once extensive human resource consumption has been detected, the system can proactively act to support the user by looking for ways to reduce their CL. Such actions can range from suppressing interruptions to increasing the degree of AI support offered in the tasks presented. In addition, a third source of information is the set of decisions made by the user. Here, historical information can reveal unexpected decision-making, which in turn can flag potential attacks. Similarly, decision making can be used to

assess the human's current abilities—particularly in relation to the aforementioned detection of human resource consumption. Here we build on the concept of "golden questions" as frequently used in crowdsourcing scenarios—a small set of tasks or questions to which the correct answer is known and can therefore be used to assess the quality of the worker's output. Such verification steps can be incorporated into human-AI systems to monitor human performance.

FUTURE RESEARCH

Despite the existing knowledge of monitoring human biosignals, a wide range of considerations remains open. While the user typically takes the role of the final decision maker in many human-AI systems, the proposed monitoring of the human's performance and decisions implies the user to be monitored by the AI. This raises challenging questions, particularly associated with interruption and communication of system state. Therefore, future research should investigate:

- the extent to which human-AI teams are vulnerable to DDoD attacks;
- parameters, signs, and indicators that enable human-AI teams to detect DDoD attacks;
- potential defenses and ways to integrate DDoD defenses into the design of human-AI systems from the beginning and not as an afterthought;
- how to make the human collaborator cognitively less vulnerable to overload;
- how the output of a system under attack influences factors such as trust in the system;
- how decisions made by a human collaborator under attack influence the ML model;
- how to mitigate the control problem leading to human collaborators blindly trusting the AI output;
- format, design, and timing of explainability elements that help lowering the cognitive demand on the human.

CONCLUSION

The HCI community's extensive efforts toward establishing a better understanding of effective human-AI collaborations are far from complete. Among colossal challenges, such as explainability, transparency, and fairness, security and robustness have largely remained an underexposed element of collaborative human-AI systems. Preparing human-AI teams for sponge and DDoD attacks requires a thorough understanding of both technical and human limitations, as well as cross-disciplinary collaboration. Building on established HCI knowledge, future research could explore how to design and implement interactive systems that monitor both AI and human performance—alerting and supporting when either party faces atypical resource consumption. This article highlights how DDoD attacks can have severe consequences on the performance of collaborative human-AI teams and calls for active cross-disciplinary research on this emerging topic.

ACKNOWLEDGMENTS

This work was supported by the joint CATCH MURI-AUSMURI.

REFERENCES

1. I. Shumailov, Y. Zhao, D. Bates, N. Papernot, R. Mullins, and R. Anderson, "Sponge examples: Energy-latency attacks on neural networks," in *Proc. IEEE Eur. Symp. Secur. Privacy*, 2021, pp. 212–231.
2. J. Sweller, "Cognitive load theory, learning difficulty, and instructional design," *Learn. Instruct.*, vol. 4, no. 4, pp. 295–312, 1994.
3. I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2015, *arXiv:1412.6572*.
4. B. Biggio, B. Nelson, and P. Laskov, "Poisoning attacks against support vector machines," in *Proc. 29th Int. Conf. Int. Conf. Mach. Learn.*, 2012, pp. 1467–1474.
5. X. Chen, C. Liu, B. Li, K. Lu, and D. Song, "Targeted backdoor attacks on deep learning systems using data poisoning," 2017, *arXiv:1712.05526*.
6. R. J. Anderson, *Security Engineering: A. Guide to Building Dependable Distributed Systems*, 2nd ed. Hoboken, NJ, USA: Wiley Publishing, 2008.
7. R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *Proc. IEEE Symp. Secur. Privacy*, 2017, pp. 3–18.
8. N. Boucher, I. Shumailov, R. Anderson, and N. Papernot, "Bad characters: Imperceptible NLP attacks," in *Proc. 43rd IEEE Symp. Secur. Privacy*, 2022, pp. 1987–2004.
9. R. Zhang, N. J. McNeese, G. Freeman, and G. Musick, "'An ideal human': Expectations of AI teammates in Human-AI teaming," *Proc. ACM Hum.-Comput. Interact.*, vol. 4, no. CSCW3, pp. 1–25, 2021.
10. N. van Berkel, O. F. Ahmad, D. Stoyanov, L. Lovat, and A. Blandford, "Designing visual markers for continuous artificial intelligence support: A colonoscopy case study," *ACM Trans. Comput. Healthcare*, vol. 2, no. 1, pp. 1–24, 2021.
11. D. Wang et al., "Human-AI collaboration in data science: Exploring data scientists' perceptions of automated AI," in *Proc. ACM Hum.-Comput. Interact.*, vol. 3, no. CSCW, pp. 1–24, 2019.

12. C. Oh, J. Song, J. Choi, S. Kim, S. Lee, and B. Suh, "I lead, you help but only with enough details: Understanding user experience of co-creation with artificial intelligence," in *Proc. 2018 CHI Conf. Hum. Factors Comput. Syst.*, New York, NY, USA, 2018, pp. 1–13.
13. R. Montañez, E. Golob, and S. Xu, "Human cognition through the lens of social engineering cyberattacks," *Front. Psychol.*, vol. 11, 2020, Art. no. 1755.
14. E. Pollock, P. Chandler, and J. Sweller, "Assimilating complex information," *Learn. Instruct.*, vol. 12, no. 1, pp. 61–86, 2002.
15. J. Zerilli, A. Knott, J. Maclaurin, and C. Gavaghan, "Algorithmic decision-making and the control problem," *Minds Mach.*, vol. 29, no. 4, pp. 555–578, 2019, doi: 10.1007/s11023-019-09513-7.
16. T. Rademacher, "Artificial intelligence and law enforcement," in *Regulating Artificial Intelligence*. Berlin, Germany: Springer, 2019, pp. 225–254.
17. R. Balasubramanian, A. Libarikian, and D. McElhaney, "Insurance 2030–the impact of ai on the future of insurance," McKinsey & Company, 2018. Accessed: Feb. 8. [Online]. Available: https://www.mckinsey.com/industries/financial-services/our-insights/insurance-2030-the-impact-of-ai-on-the-future-of-insurance
18. Q. Zhang, J. Lu, and Y. Jin, "Artificial intelligence in recommender systems," *Complex Intell. Syst.*, vol. 7, no. 1, pp. 439–457, 2021.
19. H. P. A. Van Dongen and D. F. Dinges, "Circadian rhythms in fatigue, alertness and performance," *Princ. Pract. Sleep Med.*, vol. 20, no. 215, pp. 391–399, 2000.
20. A. Jacovi, A. Marasović, T. Miller, and Y. Goldberg, "Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI," in *Proc. 2021 ACM Conf. Fairness, Accountability, Transparency*, 2021, pp. 624–635.
21. S. H. Fairclough, "Fundamentals of physiological computing," *Interacting Comput.*, vol. 21, no. 1/2, pp. 133–145, 2009.

BENJAMIN TAG is a postdoctoral research fellow at the School of Computing and Information Systems, University of Melbourne, Melbourne, 2010, Australia. His research interests include digital human-AI interaction, emotion regulation, and human cognition, with a special focus on inferring mental state changes from biophysical signals in the wild. He is the corresponding author of this article. Contact him at benjamin.tag@unimelb.edu.au.

NIELS VAN BERKEL is an associate professor at Aalborg University, 9220, Aalborg, Denmark. His research interests include human–computer interaction, social computing, and human-AI interaction. Contact him at nielsvanberkel@cs.aau.dk.

SUNNY VERMA is a postdoctoral research fellow at Macquarie University, Sydney, 2109, Australia. Prior to that, he was a postdoctoral fellow at the Data Science Institute, UTS. His research interests include fairness and human cognition in AI systems and data mining. Contact him at sunny.verma@mq.edu.au.

BENJAMIN ZI HAO ZHAO is a postdoctoral research fellow at Macquarie University, Sydney, 2109, Australia. His research interests include authentication systems, security and privacy attacks against machine learning, and rapid malware triage. Contact him at ben_zi.zhao@mq.edu.au.

SHLOMO BERKOVSKY is a professor at Macquarie University, Sydney, 2109, Australia. He leads the Precision Health stream at the Centre for Health Informatics. The stream focuses on the use of AI methods to develop patient models and personalized predictions of diagnosis and care, and on the ways clinicians and patients interact with health technologies. Contact him at shlomo.berkovsky@mq.edu.au.

DALI KAAFAR is a professor of cybersecurity and privacy-preserving technologies at the Faculty of Science and Engineering. He is also the executive Director of the Macquarie University Cyber Security Hub. His research interests include digital privacy, distributed systems security, authentication systems, and security risks measurement and modeling. Contact him at dali.kaafar@mq.edu.au.

VASSILIS KOSTAKOS is a professor of computer science at the University of Melbourne, Melbourne, 3010, Australia, where he leads the Human-Computer Interaction Group. His research interests include ubiquitous computing, human–computer interaction, social computing, and the Internet of Things. Contact him at vassilis.kostakos@unimelb.edu.au.

OLGA OHRIMENKO is an associate professor at the University of Melbourne, Melbourne, 3010, Australia. Her research interests include privacy and security of machine learning and data analysis. Contact her at oohrimenko@unimelb.edu.au.