

Human Machine Interaction and Security in the Era of Modern Machine Learning

Anastasia-Maria Leventi-Peetz

Federal Office for Information Security (BSI), DE-53133 Bonn, Germany

ABSTRACT

It is realistic to describe Artificial Intelligence (AI) as the most important of emerging technologies because of its increasing dominance in almost every field of modern life and the crucial role it plays in boosting high-tech multidisciplinary developments integrated in steady innovations. The implementation of AI-based solutions for real-world problems helps to create new insights into old problems and to produce unique knowledge about intractable problems which are too complex to be efficiently solved with conventional methods. Biomedical data analysis, computer-assisted drug discovery, pandemic predictions and preparedness are only but a few examples of applied research areas that use machine learning as a pivotal data evaluation tool. Such tools process enormous amounts of data trying to discover causal relations and risk factors and predict outcomes that for example can change the course of diseases. The growing number of remarkable achievements delivered by modern machine learning algorithms in the last years raises enthusiasm for all those things that AI can do. The value of the global artificial intelligence market was calculated at USD 136.55 billion in 2022 and is estimated to expand at an annual growth rate of 37.3% from 2023 to 2030. Novel machine-learning applications in finance, national security, health, criminal justice, transportation, smart cities etc. justify the forecast that AI will have a disruptive impact on economies, societies and governance. The traditional rule-based or expert systems, known in computer science since decades implement factual, widely accepted knowledge and heuristic of human experts and they operate by practically imitating the decision-making process and reasoning functionalities of professionals. In contrast, modern statistical machine learning systems discover their own rules based on examples on the basis of vast amounts of training data introduced to them. Unfortunately, the predictions of these systems are generally not understandable by humans and quite often they are neither definite or unique. Raising the accuracy of the algorithms doesn't improve the situation. Various multi-state initiatives and business programs have been already launched and are in progress to develop technical and ethical criteria for reliable and trustworthy artificial intelligence. Considering the complexity of famous leading machine learning models (up to hundreds of billion parameters) and the influence they can exercise for example by creating text and news and also fake news, generate technical articles, identify human emotions, identify illness etc. it is necessary to expand the definition of HMI (Human Machine Interface) and invent new security concepts associated with it. The definition of HMI has to be extended to account for real-time procedural interactions of humans with algorithms and machines, for instance when faces, body movement patterns, thoughts, emotions and so on are considered to become available for classification both with or without the person's consent. The focus of this work will be set upon contemporary technical shortcomings of machine learning systems that render the security of a plethora of new kinds of human machine interactions as inadequate. Examples will be given with the purpose to raise awareness about underestimated risks.

Keywords: Human machine interaction, Deep learning, Spurious correlation, Multi-agent systems, Normal accident theory, Automatic speech recognition, Algorithmic instability, Image recognition, Cyber defence, Industry 4.0

INTRODUCTION

A human-machine interface (HMI) is classically defined to be that part of a machine or technical system that enables the human-machine interaction. It consists of hardware and software elements that allow users to interact with a machine or a system (Wikipedia, 2023). Undoubtedly HMIs have always been essential in modern life because they allow operators to control systems and processes with easiness and precision. At the same time HMIs are also considered to be the most vulnerable elements of systems which attackers can use in order to intrude in systems and take control. Artificial Intelligence (AI) is considered to be a game-changer as concerns the interaction between humans and machines because machine learning (ML) is a technology that broadens and intensifies the ways of interaction between humans and machines. Machines are able to perform tasks independently and learn continuously in the process. Also the interaction ways between humans and machines have developed and increased in number. What has started as physical buttons, keyboards and screen displays has evolved to web-based remote access control, tactile displays and haptic devices, voice and video recognition based communications etc. Computer technologies like Virtual Reality (VR), Augmented Reality (AR), eye tracking, gesture recognition etc. are also widely applied in this context. The new-generation of the so-called smart HMIs are intelligent, adaptive to various tasks interactive devices which should be easily deployable, intuitive to use, accurate and secure. HMIs are gateways to an increasingly connected society already employed in a broad range of industrial, automotive, infrastructure and even medical applications. To the key elements of smart HMIs there belong sophisticated software, especially machine learning (ML) algorithms which enable the fusing of multiple data streams to learn and recommend the optimal decision for the user after solving an optimization problem with many parameters involved. The focus of this work will be set on the complexity of the ML algorithms and their applications and the implicated risks, especially as far as security is concerned which in almost all of the cases is not a separable from safety concept.

MACHINE LEARNING: SUCCESS STORIES AND REAL WORLD PROBLEMS

There is a number of ML use cases which are widely-known because people come across them almost on a daily basis (Saitwal, 2022). Automatic speech recognition (ASR) which is also an important research area for human-machine communication, AI-powered chatbots for online customers service, and computer vision applications are from experience well known. The latter is a field of AI that trains computers and systems to derive meaningful information from digital images, videos and other visual inputs and take actions or make recommendations based on the extracted information.

Computer vision applications are used for various tasks, for instance picture recognition to photo tagging on social media, medical image classification, obstacle recognition and avoidance in automotive, home automation, robotics etc.

Well known are also the AI-based recommendation systems (RS) (or Recommendation Engines) which, as an alternative to search algorithms, learn to predict the users' choices and offer relevant suggestions. Such Engines are used for instance in e-commerce platforms to make product recommendations to potential clients, or in creating ride-sharing recommendations on the basis of customers characteristics and threshold times (taxi-recommender system with ridesharing service (Pamula, 2017)), or in e-learning context to provide users with personalized services by automatically identifying their preferences and needs and so on. Another rather well-known application of machine learning is the automated stock trading: it concerns AI-driven trading platforms with a higher frequency, designed to optimize stock portfolios and make hundreds or even millions of deals every day without human interaction.

Famous is the introduction of AI-enhanced cyber security and anomaly detection systems which are of capital importance for instance for the prediction or identification of new threats and the proper reaction in real-time. Currently the concept of human-AI interaction is implicitly added to the description of HMIs. And trust has been identified as a key feature that is fundamental for this interaction (Sapienza, 2022).

Critical infrastructures, smart grids, transportation networks etc. which are all relevant to cyber-defence are large-scale complex systems consisting of many interdependent sub-systems. Their ability to function under disturbances and threats strongly depends on distributed control systems composed of multiple interacting intelligent agents (MAS). An agent can be a human but in this context it is mostly a piece of decision-making software which interacts in a shared environment to learn and achieve certain goals. Subsystems are heterogeneous, diverse in their functionality and network structures, which demands different strategies and designs to achieve the individual resilience of each of them (Zhao, 2022). Their extensive interconnections with other subsystems create interdependencies which are not always known from the beginning or planned for. Unexpected errors or faults in one system can propagate over the network and cause fatal system failures. A holistic approach for protection of these systems becomes necessary. Distributed ML mechanisms and processes should help the system to develop cognitive capability and become able to demonstrate adaptive response to threats. This equals to the emergence of computer systems that can settle-up complex problems with little or without human intercession. Cyber-technologies should dramatically contribute to the increase of the cognitive capabilities of machines, changing machines from being reactive to self-aware. The actors in Industry 4.0 manufacturing systems can be human, organizational and technology-based agents. Artificial Intelligence is the glue for the synergistic combination of different technologies to produce a joint cognitive system. It is obvious that traditional system design methods do not properly address the autonomous capabilities of machine agents, because they observe humans as supervisors rather than teammates in a system. Several works study the development of trust models in the field of human-robot or human-AI interaction. These are mostly concerned with human-machine (H2M) interactions. A cognitive architecture for trustworthy human-robot collaboration has been also proposed

(Nikolaidis, 2017). H2M trust models mostly investigate human–machine relationships which are articulated but often limited to a one-to-one relationship. One-to-one models cannot account for collaborative scenarios though, where humans interact with many devices and devices have to manage multiple users. These users can be human and/or artificial agents. To anticipate reality, additional aspects have to get integrated into trust models such as modules representing the relationship of artificial agents toward humans (M2H) and machine-to-machine (M2M) ones.

The competence and availability of the system’s partners are decisive for the trustworthiness of the collaboration, whereby these properties are generally neither independent between participating partners nor linear in their contribution. It is very probable that cognitive machines will start operations already before their functionality has been understood or the risks connected with their operation estimated. At present, only the competence and reliability of ML algorithms can be generally and specifically get addressed. The next paragraph is dedicated to this issue.

Judging the Competence of Dynamic Complex Systems With AI-Driven Parts

Taking the automated stock trading as an example that was mentioned above, there already exist articles and books that analyze risks and key failures associated with algorithmic and especially high-frequency trading algorithms, including the notorious flash crashes (Min, 2021). In financial markets human traders have been replaced by automated computer algorithms. Markets have been actualized to allow interaction between human and non-human agents. However, new types of market failures and crashes are now linked to the technological risk of algorithmic trading.

Inspired by science and technology studies (Min, 2021) authors maintain that Perrow’s normal accident theory (NAT) with its perception of accident-prone technological systems offers the right background to analyze and understand automated markets and the individual organizations that are active on them (Dorner, 2014). Algorithmic interaction patterns are often nonlinear and demonstrate typical characteristics of those complex interactions that are associated with normal accident-prone systems. Because machine learning proliferates rapidly in the financial industry, it is been investigated how market participants assess, and manage these new complex techniques that gain an increasing significance to the sociology of financial markets. The matter is still a subject of ongoing research. Empirical studies trying to understand human–model interactions and entanglements in the financial markets have delivered some interesting first results (Hansen, 2020). Users of machine learning models are often concerned that models might learn the wrong things and thus, become deceitful rather than informative. Parallel to the algorithms, users utilize a so-called distributed cognition which is shared or unshared knowledge and beliefs in the context of financial markets. However, the increasing automation has as result that algorithms no longer inform human decision-making. Algorithms often also determine *what*, *when* and *how* to trade or invest and execute accordingly. Machine learning models

have the capacity to learn and optimize their strategies without interference from model developers and users. They continuously and dynamically adapt to new input data. In this kind of human-AI interface (the interface to the model) the user plays a very restricted role. He has the choice to trust a complex algorithm operating within a very complex environment that he hardly understands or use more classical methods of management of his financial assets. At this point it has to be underlined that also the developers of the algorithms themselves have a limited understanding of the models they develop. Machine learning expands the scope of data mining and data processing and thus, enhances the capacity to trawl markets in search of patterns and correlations to exploit. However, automatic fast algorithmic strategies involve the typical components of the tight coupling and complex interactions which are the basic conditions of (NAT). That means, that possibly even a minor local disturbance in the system, for instance an accidental network connection failure, can create new correlations to pervade the entire system, changing significantly patterns which the ML algorithm discovered some milliseconds ago and causing unforeseen results for the trader(s). Even if the ML model were explainable to the user, the effect of its application could still be unpredictable. One has to deal with complex highly dynamical systems with no global fixed-time synchronization constraints.

There exist parallels between the multiply interacting intelligent agents discussed in the previous paragraph, concerning the cyber-defence of large-complex systems, and the automatic algorithmic trading indicated here. The tight coupling, complex interactions and ML are the common elements. The main difference is that in the cyber-defence case agents cooperate to acquire a shared cognition about a common task that has to be fulfilled, namely the steady security of the whole system. In the algorithmic trading case, agents follow individual and also antagonistic tasks. The human interaction with such systems is the interaction with dynamically orchestrated stacks of AI-algorithms and services which makes any sensible human intervention appear hardly possible.

Algorithmic Reliability and Stability of ML Results

AI is a paradigm changing technology ultimately meant to replace humans in problem solving but it also replaces standard algorithms in computational science and engineering. Reliable numerical calculations are of fundamental importance. Accuracy and stability of algorithmic results are necessary. Unfortunately, statistical data-driven learning methods offer no guarantee that the trained ML model produces definite, stable or causal results and makes no mistakes when applied in real-world problems (Leventi-Peetz, 2021). In fact, different models can function equally well to fit the same data. This leads to the so-called *predictive multiplicity*, associated to the *Rashomon Effect* in ML, originally introduced by Leo Breiman in his famous work of 2001, by the title: *Statistical Modeling: The Two Cultures*. The ambiguity or instability of results which potentially leads to failing model reliability in the praxis has been extensively discussed in many scientific publications.


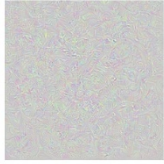

				<p>Article: Super Bowl 50</p> <p>Paragraph: " Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Cook had a jersey number 37 in Champ Bowl XXXIV."</p> <p>Question: "What is the name of the quarterback who was 38 in Super Bowl XXXIII?"</p> <p>Original Prediction: John Elway</p> <p>Prediction under adversary: Jeff Cook</p>
Task for DNN	Caption image	Recognise object	Recognise pneumonia	Answer question
Problem	Describes green hillside as grazing sheep	Hallucinates teapot if certain patterns are present	Fails on scans from new hospitals	Changes answer if irrelevant information is added
Shortcut	Uses background to recognise primary object	Uses features irrecognisable to humans	Looks at hospital token, not lung	Only looks at last sentence and ignores context

Figure 1: Deep neural networks often solve problems using shortcuts instead of learning the intended solution which leads to failures. This problem is observed in many real-world applications (Geirhos, 2020).

Unfortunately raising the algorithmic accuracy doesn't help against the failing definiteness of model predictions. Competing models of almost the same accuracy belonging to the so-named *Rashomon-set* can deliver different results. Deep learning, due to its unprecedented success in tasks such as image classification, has emerged as a standard tool in image reconstruction with application in many fields: medical image analysis and classification, clinical risk prediction based on electronic health records, drug response prediction, precision medicine, natural language processing (NLP) etc. (D'Amour, 2022). However, tiny and almost undetectable with the naked eye perturbations of an image can cause severe diagnostic errors (Antun, 2019).

In a recent paper, written by forty google engineers under the title *Underspecification Presents Challenges for Credibility in Modern Machine Learning* (D'Amour, 2022) the intrinsic problem of the underspecification of deep learning models has been identified as responsible for ambiguous and false results. Highly-parametrized models (NLP models possess hundreds of billions of parameters) tend to learn spurious correlations and shortcuts from their training data. Examples of shortcut learning applied by Deep Neural Networks (DNN) and leading to errors are depicted in Fig. 1. A spurious correlation example and an according false classification are depicted in Fig. 2.

CONCLUSION

Prior to the establishment of any conclusive discussion about the legal and ethical issues emerging out of the massive application of AI-algorithms in almost every field of modern life, it is an urgent necessity that a deeper understanding of the functionality but also the weaknesses and the problems of these algorithms will be gained and the consequences understood. Perhaps new, alternative ML-methods with smaller and therefore more transparent models which can be equally effective as the big models of deep learning have to be developed. Research on the way to prove the existence of simpler models has already produced promising results (Semenova, 2022). Training of smaller models is very appealing also because of their

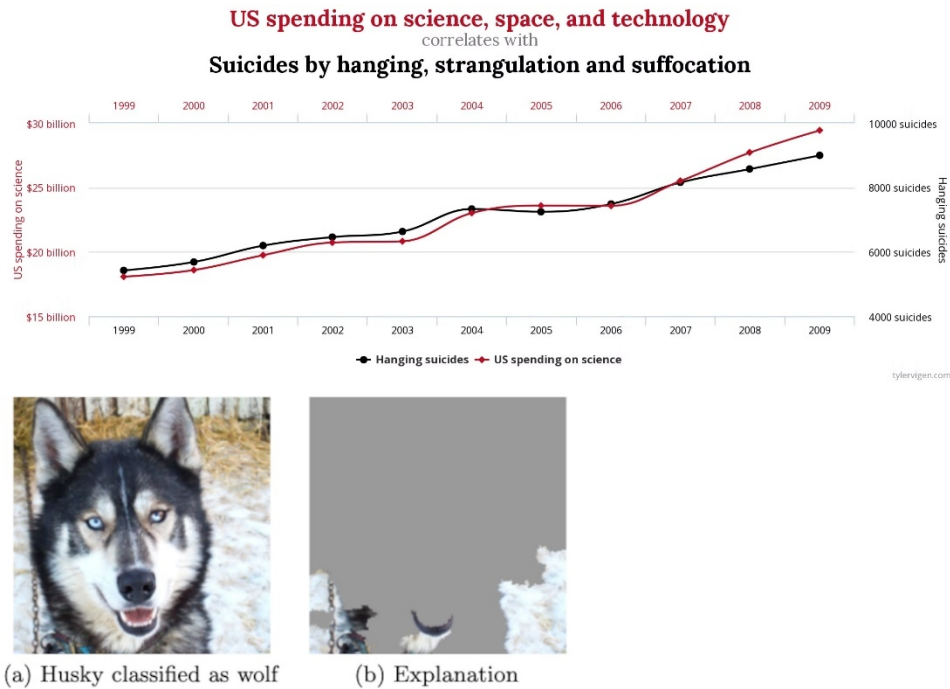


Figure 2: Top: Spurious correlation (curves sloping together with no existing causation (Berman, 2017)). Bottom: the picture of a husky wrongly classified as a wolf because of the snow in the background. The neural network (NN) was trained with images of wolves in snowy landscapes (Besse, 2019).

advantage to respect environmental sustainability. In April 2019, the European Commission released its *Ethics Guidelines for Trustworthy Artificial Intelligence*. The document encompasses seven guiding principles, among them transparency (the traceability of AI systems should be ensured), privacy and data governance (citizens should have control over their own data) and diversity, non-discrimination and fairness (which tackles the bias problems of AI systems) (Sahin, 2019). Now, almost four years later, the number of AI-algorithms and learning methods have increased while their transparency has hardly improved.

A statement of Jeanne Lim, Co-founder & CEO of beingAI, concerning the Next-Generation Human-Machine Interface, reads: *The borders between physical and virtual realms have started to disappear and smart devices and services pervade into every aspect of life. A new generation of intelligent, real-time interactive, and personalized agents are to be created to build engagement and trust between human and machine across devices and media platforms. These new agents should be a new category of intelligent, autonomous, evolving, and ubiquitous human-AI interface that traverse the physical and digital realms to engage with humans anywhere, anytime, as part of an interactive content and narrative experience. They will unite and elevate the core human experience in the upcoming age of AI and the advent of the immersive life in the metaverse.*

There are no security considerations in these words, no indication about the basis on which the trust between the human and the intelligent personalized agent can be built upon. A more human-centered definition is necessary, one which places humans higher than plain counterparts of intelligent agents, allowing them more control choices than the *blind* acceptance or denial to engage in a *dark* metaverse.

The new generation of Human-AI interface issue is still open.

REFERENCES

- Antun, Vegard et al. (2020). On instabilities of deep learning in image reconstruction and the potential costs of AI. In *Proceedings of the National Academy of Sciences (PNAS)*, 117 (48). <https://doi.org/10.1073/pnas.1907377117>.
- Berman, Robby (2017). 5 brilliant graphs that teach correlation vs. causality. In BigThink Health. Retrieved February 2023, <https://bigthink.com/health/data-connections-nicholas-cage-movies-to-drownings-wait-what>.
- Besse, Philippe et al. (2019). Can Everyday AI be Ethical? Machine Learning Algorithm Fairness. *Statistiques et Société*, 6(3) <https://ssrn.com/abstract=3391288>.
- D'Amour, Alexander et al. (2022). Underspecification Presents Challenges for Credibility in Modern Machine Learning. *Journal of Machine Learning Research* 23(226). <http://jmlr.org/papers/v23/20-1335.html>.
- Dorner, Michael (2014). Normal Accidents and Computer Systems. In *Proceedings of the Seminars Future Internet (FI) and Innovative Internet Technologies and Mobile Communications (IITM), Winter Semester 2013/2014*. https://doi.org/10.2313/NET-2014-03-1_04.
- Geirhos, Robert et al. (2020). Shortcut learning in deep neural networks. *Nat. Mach. Intell.* 2. <https://doi.org/10.1038/s42256-020-00257-z>.
- Hansen, Kristian Bondo (2020), The virtue of simplicity: On machine learning models in algorithmic trading. *Big Data & Society*, 7(1). <https://doi.org/10.1177/2053951720926558>.
- Huchler, Norbert et al., eds. (2020). Criteria for Human-Machine Interaction when using AI – Approaches to its humane design in the realm of work. White paper from Plattform Lernende Systeme. <https://www.plattform-lernende-systeme.de>.
- Leventi-Peetz, Anastasia-Maria et al. (2021). Scope and Sense of Explainability for AI-Systems. In *Intelligent Systems and Applications. IntelliSys 2021. Lecture Notes in Networks and Systems*, vol 294. https://doi.org/10.1007/978-3-030-82193-7_19.
- Min, Bo Hee and Borch, Christian (2021). Systemic failures and organizational risk management in algorithmic trading: Normal accidents and high reliability in financial markets. *Soc Stud Sci.*, 52(2) <https://doi.org/10.1177/03063127211048515>.
- Nikolaidis, Stefanos et al. (2017). Human-robot mutual adaptation in collaborative tasks: Models and experiments. In *The International Journal of Robotics Research*, vol 36(5-7) <https://doi.org/10.1177/0278364917690593>
- Pamula, Rajendra and Chakraborty, Rini (2017). Taxi recommender system using ridesharing service. *4th International Conference on Advanced Computing and Communication Systems (ICACCS)*, <https://doi.org/10.1109/ICACCS.2017.8014568>.
- Sahin, Kaan (2019). The Devil's in the Detail. *Berlin Policy Journal*. Retrieved February 2023, <https://dgap.org/en/research/publications/devils-detail>.
- Saitwal, Ashutosh (2022). Machine learning and its importance for Businesses. Retrieved February 2023, <http://www.klearstack.com/success-stories-in-fields-of-machine-learning>.

- Sapienza, Alessandro et al. (2022). Modeling Interaction in Human–Machine Systems: A Trust and Trustworthiness Approach. *Automation*, 3, <https://doi.org/10.3390/automation3020012>.
- Semenova, Lesia, Rudin, Cynthia and Parr, Ronald (2022). On the Existence of Simpler Machine Learning Models. *2022 ACM Conference on Fairness, Accountability, and Transparency*. <https://doi.org/10.1145/3531146.3533232>.
- UltraSense Systems Inc. (2022). The Next-Generation Human-Machine Interface. Retrieved February 2023, <https://ultrasensesys.com/the-next-generation-human-machine-interface>.
- Wikipedia contributors. User interface. In *Wikipedia, The Free Encyclopedia*. Retrieved February 2023, https://en.wikipedia.org/wiki/User_interface.
- Zhao, Yuhan et al. (2022). Multi-Agent Learning for Resilient Distributed Control Systems, preprint <https://arxiv.org/abs/2208.05060>.