Trustworthy Artificial Intelligence: Comprehensible, Transparent and Correctable

Ute Schmid

Abstract With the digital transformation, artificial intelligence (AI) applications are also finding their way into more and more areas of work and life. In particular, models learned from data are being used, which are mostly opaque black boxes. The fact that people can understand why an AI system behaves the way it does is necessary for various reasons: The model developers themselves must be able to assess properties of the learned models—in particular, possible biases due to overfitting to the data used for learning. For safety-critical applications, aspects of certification and testing are also becoming increasingly relevant. Domain experts—for example, in medical diagnostics or quality control in industrial production—must be able to comprehend, verify and, if necessary, correct system decisions. Consumers should understand why a system—a smart home control, a driving assistance—behaves in a certain way and why they are recommended certain products, offered certain tariffs or denied certain offers. After a brief introduction to the topic of AI, the chapter gives an overview of methods of the so-called third wave of AI. Central to this are approaches of *explainable* AI (XAI), which are intended to make the decisions of AI systems comprehensible. The main approaches are characterized and shown for which objectives and applications they are suitable in each case. It is shown that in addition to the highly regarded methods for visualization, methods that allow system decisions to be described in a differentiated manner are also particularly important. It is also argued that, in addition to comprehensibility, interactivity and correctability of AI systems are necessary so that AI systems do not restrict human competences but support them in partnership.

U. Schmid (✉)
Cognitive Systems, University of Bamberg & Bavarian Research Institute for Digital Transformation (bidt), Bamberg, Germany
e-mail: ute.schmid@uni-bamberg.de

H. Werthner et al. (eds.), *Introduction to Digital Humanism*,
https://doi.org/10.1007/978-3-031-45304-5_10

1 Introduction

Artificial intelligence (AI) is that field of research in computer science in which algorithms are developed to solve problems that humans are currently better at solving (definition according to Rich, 1983). AI is a research domain within computer science. In general, AI approaches should only be applied for problems which cannot be solved with standard algorithms. While standard algorithms—at least in principle—guarantee correctness (an input results in the intended output) and completeness (for all possible inputs, an output can be computed), this does not in general hold for AI algorithms. For many safety-critical domains, AI algorithms are usually not an option. For instance, the controller of an airbag should react in the intended way in all situations. AI systems become necessary for one of the following two reasons: (1) A problem is too complex that a solution can be computed efficiently. That is, it would take an unacceptably long time to generate an output. In this case, heuristic algorithms (one of the core approaches of AI) are used to compute approximate solutions without a guarantee how near the produced solution is to a desired or optimal solution for a problem. (2) It is not possible to give a full explicit description of the problem, and consequently, it is not possible to even define an algorithm. In this case, the algorithm for processing inputs into outputs is approximated from data, that is, by machine learning. Between input and output, there is now not an explicit, inspectable program but a machine-learned model which has generalized over data.

The field AI was given its name "artificial intelligence" in 1956 by computer science pioneer John McCarthy at Stanford University. The two main families of AI methods are knowledge-based methods and machine learning (see the most widely used textbook by Russell & Norvig, 2020). Both areas have been considered from the beginning. The first implementation of a machine learning program was a program to learn a strategy for the game of checkers and realized by Arthur Samuel in 1952. Early approaches also included the perceptron as a model of a single neuron and decision tree algorithms as an example for symbolic/interpretable machine learning (Rudin, 2019).

The 1980s was the peak period for knowledge-based methods in the context of applications for expert systems. It was hoped that AI systems could relieve or support human experts in many areas—from medical diagnostics to the planning of production processes to the use of intelligent tutoring systems in teaching. In the context of research on knowledge-based systems, efficient algorithms for drawing conclusions emerged. Special AI programming languages such as the logic programming language Prolog and specific hardware for more efficient processing, especially the Lisp machine, were developed. Research on machine learning still took place, but was dominated by the knowledge-based approaches. The heyday of expert systems accordingly has similarities to the current hype in machine learning. Again, one direction strongly dominates, and special program libraries are developed for deep neural networks as well as special hardware in the form of GPUs (*graphics*

processing units) allowing to multiply matrices particularly efficiently with multiplication of matrices of real numbers as core operation for neural networks.

The high hopes placed in expert systems could ultimately only be partially fulfilled, especially due to the so-called knowledge engineering bottleneck—the realization that human knowledge is only partly explicitly available and can be formally represented. Large areas of human knowledge, especially perceptual knowledge and highly automated action routines, are implicit and cannot be captured or can only be captured inadequately with knowledge acquisition methods. The phenomenon is also called Polanyi's paradox: *How can we humans know more than what we can talk about?*

Impressive successes in the application of deep neural networks have heralded a new peak phase in AI since around 2010—this time with a focus on machine learning. The main reason for the new great interest in AI is that for the first time, it was possible to learn almost directly from different types of data, such as images or texts, without complex pre-processing (*end-to-end learning*). Most machine learning approaches, including classical neural networks as developed since the late 1980s, expect data in the form of feature vectors as input. Many data are available in tabular form anyway—for example, customer data or patient data. However, if you want to learn from image data such as photos of objects or even X-ray images, for example, you first have to extract features such as textures or color distributions from the available image data for the classical machine learning approaches. Just as for the knowledge-based approaches of AI, perceptual tasks also posed a challenge for machine learning.

In 2012, a deep neural network—a *convolutional neural network* (CNN) called AlexNet—won the ImageNet Challenge for the first time (Krizhevsky et al., 2012). In the challenge, images from 1000 categories, for example, animal species, vehicle types and buildings, are to be classified. Several million images are available for this purpose, for which the objects depicted are annotated by hand. Unlike earlier machine learning approaches, AlexNet could learn directly from the images. Comparable developments exist for natural language processing, such as machine translation (DeepL) or text generation (GPT-3). Again, however, expectations of what these novel AI methods can do are overblown. Polanyi's revenge (Kambhampati, 2021) has swung the pendulum from a near-exclusive focus on AI methods for explicit knowledge to a sole focus on AI methods for tacit knowledge. For any given problem, learning from lots of data is seen as the only meaningful approach. Existing knowledge, including carefully acquired knowledge about causal relationships, is thrown overboard to learn things imperfectly from data for which explicit knowledge is available. At the same time, traceability and control are abandoned, since deep neural networks calculate inputs in a complex mathematical way and are thus black boxes.

2 Problems with Data-Intensive Machine Learning: Unfairness, Biases and Missing Robustness

Even though data-intensive machine learning with the new generation of deep neural networks opens up new possibilities for various application areas, it also brings new problems. The requirements for quantity and quality of data are extremely high. The ImageNet already mentioned consists of 14 million images and 20,000 categories. It is often overlooked that the effort of capturing knowledge and formalizing it for processing by AI methods does not disappear with machine learning, but is deferred to the correct annotation of training data. Clickworkers have to manually annotate each example with the correct category—or even mark objects in images. The more complex the architecture of a neural network, the more data is needed to train it. If too little data is available, it is duplicated (augmented). Images, for example, are changed in their color values. In complex application areas where it is unclear which complex combination of information is responsible for a certain category, this can lead to unwanted *biases*. For example, when diagnosing tumors from tissue sections, the tissue is often colored. A model that decides whether a tumor is present, and if so which category, could be misled by training data with different staining than the original.

Supervised machine learning approaches, and this includes many deep neural network approaches, require a sample of training data that is as representative as possible for the problem and that is annotated with the correct output—this is called *ground truth labelling*. Especially in medicine, but also in other application areas, it is often not clear what the correct decision is for a given datum. For example, it could be that one medical expert decides on tumor class pT3 for the same image of a tissue section, while another decides on pT4. If certain types of data are missing from the training set (*sampling bias*) and data are not correctly annotated, this has a direct impact on the quality of the learned model (see Bruckert et al., 2020). In addition, models generated from data can typically only generalize for similar data that lie within the distribution of the data in the training set, but not for data that lie outside the distribution. If one has trained a model that can distinguish car types and it later receives a washing machine as input, it will classify it in terms of similarity to the car types it has learned. A human being, on the other hand, would say, that's something completely different from what I've seen so far, I can't say anything about that. Learned models do not have this kind of meta-cognition by default. A knowledge-based AI system, on the other hand, would not process an input outside the domain under consideration. So the quality of learned models depends heavily on the selection and quality of the data it has been trained with.

But even if the data are collected representatively and annotated correctly, undesirable effects can occur. Unfairness in reality is represented in the data. If there are significantly fewer women working in IT than men in a company and one naively simply trains a model for application selection with the existing data, the result is that a female applicant is no longer considered for a position in IT at all, as happened with Amazon's recruiting tool in 2018 (Dastin, 2018). If one is aware of

such unfair distributions in the data in advance, this can be taken into account through appropriate methods in the learning process. In general, however, unfair models cannot be ruled out completely.

Both human and machine learning are inference from a sample of data or experience to a population. Such inductive inferences can never be completely correct. Human concept acquisition is generally very robust. For example, we have no trouble distinguishing cats from other animals, even with very different types of cats, lighting or backgrounds. In other areas, people tend to overgeneralize and form stereotypes and prejudices. Prejudices related to gender or ethnicity cannot be eliminated, but they can be recognized and also corrected. But with both human and machine learning, it is true that mistakes can be made. With machine-learned models, one estimates what the error rate will be for unseen data. A *predictive accuracy of* 99% does not sound bad, but it means that the model will make an error every hundredth case. If you use a search engine to look for pictures of cats, it doesn't matter if every 100th picture shows something different. Here, the advantages of automated image retrieval outweigh the disadvantages. You look at the pictures and choose a suitable one. In contrast, if, in a medical diagnosis, a disease were mistakenly diagnosed or—even worse—overlooked in every 100th case, that would be intolerable. Similarly, it is certainly undesirable that every 100th person is wrongly denied a loan or an insurance rate is set too high for no reason.

In order to be able to recognize and correct such undesirable model decisions, it can be very helpful to comprehend which information of the input data has been taken into account by which the model came to its decision. However, many machine learning approaches, especially deep neural networks, construct non-transparent models that are black boxes even for the model developers themselves.

3 Explainable Artificial Intelligence: Comprehensibility of Machine-Learned Models

The growing interest in the use of data-intensive AI methods impacted more and more application areas since around 2015. It quickly became clear that an exclusive focus on black-box machine learning approaches is often neither possible nor desirable. Possible applications are limited by the data quantity and quality requirements discussed above, but especially by the high effort required to annotate the training data. In addition, it has been realized that—especially in safety-critical areas such as medicine—systems where it is not possible to understand the basis on which they arrive at a decision or a recommendation for action are not acceptable. In areas that have a direct impact on consumers—from personalized advertising to lending—the right to transparency was also soon demanded (Goodman & Flaxman, 2017).

In spring 2017, DARPA (*Defense Advanced Research Projects Agency*, USA) launched the *Explainable Artificial Intelligence* (XAI) program. The aim of the

program is to develop methods that (a) lead to machine-learned models that are more comprehensible than black-box models but at the same time retain a high degree of predictive accuracy and (b) enable users to understand this emerging generation of partnered AI systems, to trust the decisions appropriately and to interact effectively with the systems (Gunning & Aha, 2019). Using the classification of a cat by a neural network as an example, it was shown that an explanation of the model decision can include both verbalizable features such as "has fur, whiskers and claws" and prototypical images of typical visual features such as the shape of the ears (see https://twitter.com/darpa/status/843067035366187008, 18.3.2017). However, the term *explainable* led to misunderstandings outside of the research community, as it rather suggests that the workings of AI systems are explained in a way that is understandable to laypersons. However, XAI means to provide methods which allow to make the decision-making process of an AI system, specifically a machine-learned model, more transparent. In parallel, terms such as "comprehensible machine learning" (Schmid, 2018) or interpretable machine learning (Doshi-Velez & Kim, 2017) were proposed. In the meantime, the term *explanatory* machine learning is also frequently used (Teso & Kersting, 2019; Ai et al., 2021). Furthermore, transparency is now usually understood more generally than explainability: it refers to the principle that it should be made clear when a recommendation or decision is based on the use of AI methods or if an interaction is not with a human but with an AI system such as a chatbot.

In the meantime, a standardization of terminology has developed: After the initial focus on explainability for deep neural networks, the relevance of methods for generating explanations, in short XAI methods, is now seen for all types of AI systems. On the one hand, explanation methods are being developed for various black-box approaches to machine learning (this includes methods such as *support vector machines* or k-nearest neighbor approaches; see, e.g. Kersting et al. (2019) for a general introduction). On the other hand, explanatory methods are also being developed for knowledge-based AI systems as well as for *white-box* machine learning approaches. For these systems, it is in principle comprehensible how a decision is reached. But—comparable to large software systems—the models are often too complex to see through the entire process of information processing. In addition, the models are stored in special representation formalisms that enable processing by computer programs and must be suitably translated into comprehensible explanations. Recently, it has been established to refer to white-box machine learning approaches, such as decision tree methods, as interpretable machine learning (Rudin, 2019).

In the meantime, a wide range of XAI methods exists that are suitable for different target groups and different information needs. There are numerous methods that show the relevance of specific information from the input for the current decision. This can be features, words or parts of images. For example, the LIME approach (Ribeiro et al., 2016) shows which groups of pixels must be present for a classification decision—for example, that eye and ear are relevant for whether the model recognizes a cat. LIME is a so-called model-agnostic explanation approach: to generate an explanation, the learned model is not interfered with; instead, the input

data is manipulated, and the resulting model decision is considered. An approach that was developed specifically for image classification with (deep) neural networks is LRP (*layer-wise relevance propagation*; Bach et al., 2015). Here, those image points are highlighted that had a particularly strong influence on the output of the network. In contrast to LIME, LRP is model-specific, which means that the method must be integrated directly into the learning algorithm. Highlighting the information that is particularly relevant to a learned model is especially useful for model developers to check whether the model has generalized meaningfully. During learning, it can happen that the model uses irrelevant information for prediction that correlates with the class to be predicted. In other words, the model adapts too much to the training data (*overfitting*), which can lead to problems with the prediction for data that has never been seen before. This is also referred to as "right for the wrong reasons" or "Kluge Hans" predictors. For example, it could be that by chance a part of the photos showing horses is given with a source reference (e.g. a website). The learning algorithm can then use this simpler information to correctly indicate when a horse is seen for the available data. However, highlighting the pixels used can show that this output is based on the source cue (Lapuschkin et al., 2019).

For domain experts and also for end users, highlighting only relevant information is usually not very helpful. For example, visual highlighting can show that a certain tumor is actually visible on a tissue section. However, in order to understand why the model has decided on tumor class pT3 and not on pT4, much more complex information is required that can be better expressed in language. This includes spatial relations, such as the position of the tumor relative to other tissue types, or the concrete expression of individual features, such as the diameter of the tumor (Bruckert et al. 2020; Schmid, 2021). Such explanations can be generated, for example, combining black-box machine learning approaches and interpretable approaches (Rabold et al., 2020a).

For consumers, simple explanations such as those familiar from recommendation systems are often relevant (Tintarev & Masthoff, 2012). For example, if a certain product is recommended in an online shop, one can ask on what data basis this recommendation was made. Typically, one is then shown previous purchases that have been compared with the purchase profiles of other people for a similarity comparison. When it comes to making transparent how algorithms (with and without AI components) at banks, insurance companies or other companies come to certain decisions, such as the rejection of a loan or the amount of an insurance premium, counterfactual explanations are particularly helpful (Wachter et al., 2017)—for example: "You did not get the loan because your annual income is €45,000. If your annual income was €55,000, you would have received the loan". Such explanations give the relevant information to customers while avoiding the need for companies to reveal their algorithms. In case a model decision has been based on erroneous assumptions about a customer, it should be possible for the customer to complain and ask for a correction (actionability).

Prototypical as well as contrastive examples provide another possibility for explanations. Such examples offer experts in particular the opportunity to better understand how the model is structured. The XAI methods considered so far explain

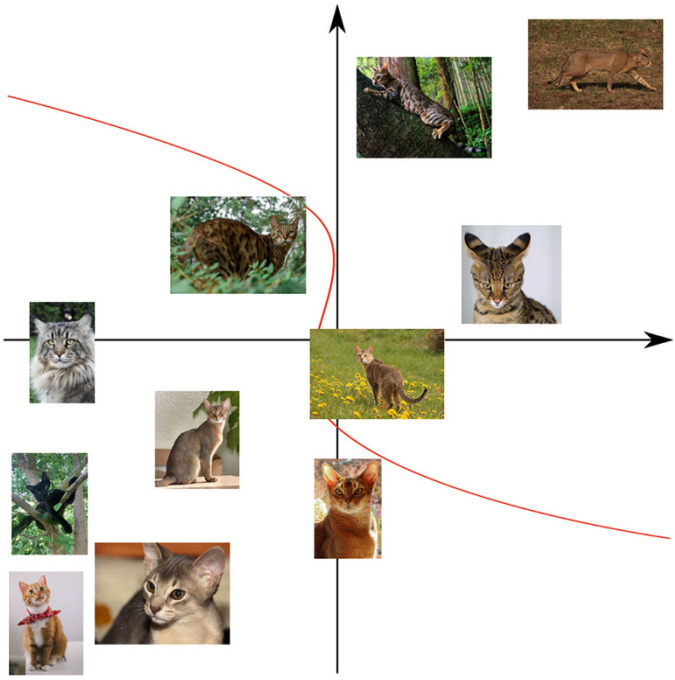

Fig. 1 Explaining image classifications by example. On the left side are house cats, and on the right side are small wild cats. For the house cat, the grey sitting cat might be the prototype for the class. The cat in the dandelion field is a near-miss example for the class house cat—a cat very similar to house cat examples but classified (correctly) as a small wild cat. Alternatively to the cat domain, one can think of images indicative for two tumor classes II and III or images for defect or acceptable parts in industrial quality control

how a specific decision was reached (local explanation). Specially selected examples can (1) show which data a model evaluates to be particularly typical for a certain class—a prototypical representant with respect to the decision region the model induced for this class; instead of identifying a prototypical example, a synthetic representant for a concept might be constructed, as it is usually proposed in psychology and philosophy; (2) examples which are situated near to the decision boundary for a class help to get insights in the discriminative features of the model. This can be a borderline case for the considered class or a near-miss example (Rabold et al., 2022), that is, an example similar to objects of the considered class but being classified as a member of a different class (see Fig. 1).

Another type of explanation tries to explain the entire model—so-called global explanations. While a local explanation supports understanding why a specific example is classified as belonging to a certain class (e.g. *Why do you classify this image as indicating tumor class II?*), a global explanation supports understanding of what constitutes a class given a specific model (e.g. *What features are in general relevant to decide that an image is indicating tumor class II?*). An explanation by

prototype can be seen as a special instance of a global explanation. Another possibility is to learn symbolic rules as a surrogate model. Such rules can be based on identifying concepts and their relations. For instance, a model classifying faces should take into account the presence of eyes, nose and mouth as concepts together with their spatial relations (e.g. that the nose is above the mouth; Rabold et al., 2020b).

Explanatory AI thus consists of a growing set of different methods, each fitting different information goals. Theoretical and empirical analyses of the properties and effects of explanations from psychology are increasingly being incorporated into research on XAI (Miller, 2019). XAI methods are an important contribution to the comprehensibility of AI systems, especially machine learning. However, in the respective application context, especially in the professional environment, it must be carefully checked that explanations are actually used to control system decisions and thus not blind but justified trust in an AI system can develop (Thaler & Schmid, 2021). The danger is that the mere existence of the possibility of an explanation leads to system decisions being adopted without reflection (Lee & See, 2004).

4 Third-Wave AI Methods: Hybrid, Comprehensible and Correctable

Explainable AI methods are also referred to as the third wave of AI—after the first wave of knowledge-based approaches (*describe*), followed by data-intensive machine learning (*categorize*), which is to be replaced by approaches that adapt to the interests of the users depending on the context (*explain*). It is increasingly argued that the methods required for the third wave must not only address the generation of explanations, but that machine learning should allow interaction, especially corrections of the model (Teso & Kersting, 2019; Müller et al., 2022). Furthermore, it is seen that a combination of knowledge-based approaches and machine learning can lead to more data-efficient and robust models (see Fig. 2). This direction of research is referred to as hybrid AI or neurosymbolic AI (De Raedt et al., 2020).

The combination of explanatory and interactive (human-in-the-loop) machine learning is a useful approach to counteract the problems with the quantity and quality of data discussed above. For example, experts can simply accept a system decision that they can directly understand, question a system decision more closely by requesting one or more explanations from the system as to how the decision was reached and, in the third step, also correct this decision. While most work on interactive machine learning only allows the correction of the output, there are now first approaches that additionally allow the correction of the explanations. This allows the adaptation of the model to be controlled in a targeted manner (Schmid, 2021). Interaction thus allows targeted human knowledge to be introduced into the learning process (see Fig. 3). Corrections are also possible when knowledge cannot be made completely explicit. For example, an expert can often recognize

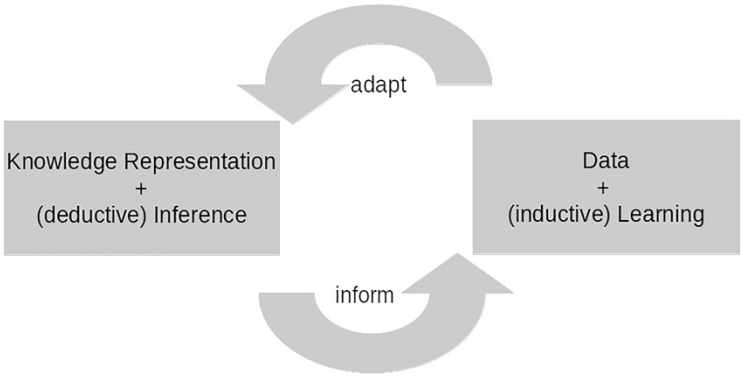

Fig. 2 Combining knowledge-based and data-driven AI: What we already know we do not need to learn (over and over) again

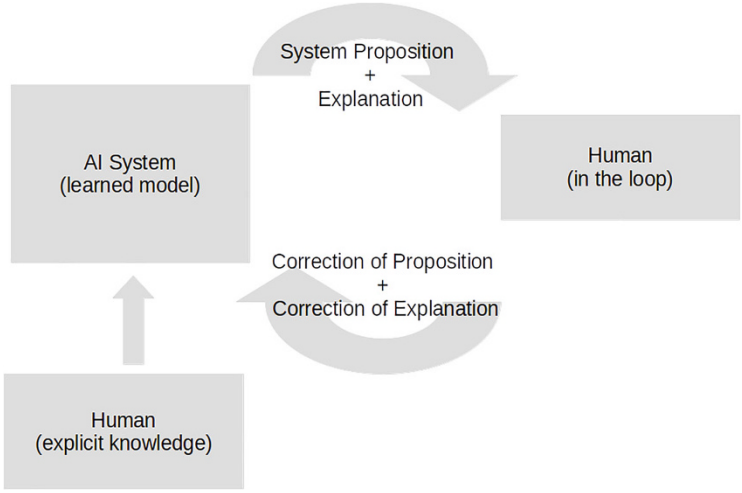

Fig. 3 Human-in-the-loop machine learning: Making use of explicit knowledge and of corrections to guide model generation and adaptation

whether a diagnosis is acceptable or not and possibly also identify faulty assumptions in its justification. At the same time, it can be assumed that the possibility of correction leads to a stronger sense of control and self-efficacy and thus there is less danger of blindly adopting system decisions.

Finally, there is a growing realization that purely data-driven machine learning is often not very efficient. While people use knowledge and skills they have already acquired and can thus learn increasingly complex things, machine learning involves learning everything from scratch over and over again. If prior knowledge could be incorporated into the learning process, data could be saved, which in turn could lead to less effort for annotation as well as savings in energy for storage and processing.

In addition, existing knowledge can be used specifically to guide the learning process. Models that take existing knowledge into account are less prone to unwanted biases and more robust with respect to data that lie outside the data distribution in the training.

Deep neural networks have brought the research field of artificial intelligence back into the public eye after many years. Increasing digitalization and global networking make it possible to learn from large amounts of data. For a responsible use of AI methods, the new research topics of explainable, interactive and hybrid AI provide the opportunity for AI systems to emerge in partnership, which do not curtail human competences but expand and promote them.

5 Conclusions

Explainability, the combination of data-driven and knowledge-based AI, and interactive approaches to machine learning have been introduced as relevant ingredients for trustworthy AI systems. However, one has to be careful that the presentation of an explanation does not result in unjustified trust. It is not guaranteed that an explanation is faithful to the model. That is, an explanation might be not correlated to the way in which a model did process the data. Similar effects can be observed in human explanations—one might give a reason to justify one's behavior which is not the true one. If a person is giving a wrong reason by design, the person is not truthful. However, often we have no full access to the motives underlying a specific behavior and come up with an explanation we find plausible. Furthermore, explanations as an additional source of information might result in cognitive overload (Ai et al., 2021). Therefore, explanations should only be given when a specific information need exists.

In the ethics guideline for trustworthy AI of the European Commission a non-exhaustive list of requirements for trustworthy AI is given, among them data quality (governance), inclusiveness (design for all), human oversight, fairness (non-discrimination), human autonomy, privacy, robustness, safety and transparency. The topics discussed in this chapter can contribute to realize these requirements: Explainability contributes to human oversight and transparency. Hybrid systems contribute to robustness and safety. Interactive machine learning contributes to human oversight and human autonomy.

The recent developments in the domain of generative AI systems such as large language models or dialog systems such as ChatGPT bring new challenges with respect to trustworthiness. When an output, for instance, an answer to a question, is generated, one might be interested in several aspects to evaluate the trustworthiness of the output: (1) Are statements concerning factual knowledge correct or is the system hallucinating? (2) Has the presented output originally been obtained from sources with copyright? (3) What are the original sources for the information given? (4) Why was a specific information included in the output and another omitted? Current XAI methods are not suited to explain the output of generative AI models.

First, ideas on how to provide explainability are currently being explored, and hopefully, we will see results in the near future.

Discussion Questions for Students and Their Teachers

1. If a domain expert, for instance, in medical diagnosis, wants to understand why a machine-learned model classified an image as indicative for a specific tumor class, which type of explanation would you think to be most helpful? If a person is wondering why an insurance company demands a rather high monthly amount for health insurance, which type of explanation do you think to be most helpful?
2. Do you think that for AI applications to be ethical, it is necessary that it provides explanations? Do you think explanations are sufficient for trustworthiness of an AI system?
3. Discuss reasons why combining machine learning with knowledge-based approaches allows to perform machine learning from smaller data sets and yields more robust models.
4. Do you see problems which can arise from interactive/human-in-the-loop machine learning?
5. Is it possible to provide explainability and trustworthiness for the new generation of generative AI models (such as ChatGPT)?

Learning Resources for Students

1. A comprehensive and recent survey of XAI methods is given in
   Schwalbe, G. and Finzel, B. (2023). A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts. *Data Mining and Knowledge Discovery.* https://doi.org/10.1007/s10618-022-00867-8.
2. An introduction to hybrid/neurosymbolic AI is presented by
   Garcez, A. D. A. and Lamb, L. C. (2023). Neurosymbolic AI: The 3rd wave. *Artificial Intelligence Review, 56*(11), 12387–12406.
3. Requirements for designing interactive AI systems are presented in
   Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., and Horvitz, E. (2019). Guidelines for human-AI interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–13).
4. A discussion of trustworthy AI from a human-computer interaction perspective is
   Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction*, *36*(6), 495–504.
5. A highly readable book about shortcomings of purely data-driven approaches is
   Marcus, G. and Davis, E. (2019). *Rebooting AI: Building artificial intelligence we can trust*. Vintage.
6. The ethics guidelines for trustworthy AI of the European Commission can be found at
   https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

Acknowledgements This contribution is based on a previous publication in German: Schmid, Ute (2022). Vertrauenswürdige Künstliche Intelligenz (287–298). In: Frauke Rostalski (Hrsg). *Künstliche Intelligenz: Wie gelingt eine vertrauenswürdige Verwendung in Deutschland und Europa?* Tübingen: Mohr Siebeck. I am very grateful to Carlo Ghezzi and Erich Prem for their valuable comments and suggestions for a previous version of the manuscript. Furthermore, I want to thank Sebastian Krügel and Matthias Uhl for allowing me to make use of the cat images in Fig. 1 and Sonja Ruschhaupt for providing the figure.

References

Ai, L., Muggleton, S. H., Hocquette, C., Gromowski, M., & Schmid, U. (2021). Beneficial and harmful explanatory machine learning. *Machine Learning, 110*, 695–721.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K. R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS One, 10*(7), e0130140.

Bruckert, S., Finzel, B., & Schmid, U. (2020). The next generation of medical decision support: A roadmap toward transparent expert companions. *Frontiers in Artificial Intelligence, 3*, 507973.

Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters* (2018).

De Raedt, L., Dumančić, S., Manhaeve, R., & Marra, G. (2020). From statistical relational to neuro-symbolic artificial intelligence. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence* (IJCAI-20, pp. 4943–4950).

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine, 38*(3), 50–57.

Gunning, D., & Aha, D. (2019). DARPA's explainable artificial intelligence (XAI) program. *AI Magazine, 40*(2), 44–58.

Kambhampati, S. (2021). Polanyi's revenge and AI's new romance with tacit knowledge. *Communications of the ACM, 64*(2), 31–32.

Kersting, K., Lampert, C., & Rothkopf, C. (2019). *How machines learn: Artificial intelligence explained in an understandable way*. Springer.

Krizhevsky, A., Sutskever, I., & Hinton, G. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25* (NeurIPS 2012, pp. 1097–1105).

Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., & Müller, K. R. (2019). Unmasking Clever Hans predictors and assessing what machines really learn. *Nature Communications, 10*(1), 1096–1104.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46*(1), 50–80.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence, 267*, 1–38.

Müller, D., März, M., Scheele, S., & Schmid, U. (2022). An interactive explanatory AI system for industrial quality control. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 36, No. 11, pp. 12580–12586).

Rabold, J., Deininger, H., Siebers, M., & Schmid, U. (2020a). Enriching visual with verbal explanations for relational concepts–combining LIME with Aleph. In *Machine learning and knowledge discovery in databases: International workshops of ECML PKDD 2019, Proceedings, Part I* (pp. 180–192). Springer International Publishing.

Rabold, J., Schwalbe, G., & Schmid, U. (2020b). Expressive explanations of DNNs by combining concept analysis with ILP. In *KI 2020: Advances in artificial intelligence: 43rd German Conference on AI, Proceedings 43* (pp. 148–162). Springer International Publishing.

Rabold, J., Siebers, M., & Schmid, U. (2022). Generating contrastive explanations for inductive logic programming based on a near miss approach. *Machine Learning, 111*(5), 1799–1820.

Ribeiro, M. T., Singh, S., & Guestrin, C.. (2016, August). "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144).

Rich, E. (1983). *Artificial intelligence*. McGraw-Hill.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206–215.

Russell, S., & Norvig, P. (2020). *Artificial intelligence. A modern approach* (4th ed.). Pearson.

Schmid, U. (2018). Inductive programming as approach to comprehensible machine learning. In *DKB/KIK@ KI* (pp. 4–12).

Schmid, U. (2021). Interactive learning with mutual explanations in relational domains. In S. Muggleton & N. Chater (Eds.), *Human-like machine intelligence* (Chap. 17) (pp. 338–354). Oxford University Press.

Teso, S., & Kersting, K. (2019, January). Explanatory interactive machine learning. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 239–245).

Thaler, A. M., & Schmid, U. (2021). Explaining machine learned relational concepts in visual domains-effects of perceived accuracy on joint performance and trust. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 43, No. 43, pp. 1705–1711).

Tintarev, N., & Masthoff, J. (2012). Evaluating the effectiveness of explanations for recommender systems: Methodological issues and empirical studies on the impact of personalization. *User Modeling and User-Adapted Interaction, 22*, 399–439.

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law and Technology, 31*, 841.