The Next Frontier: AI We Can Really Trust

Andreas Holzinger[1,2](✉)

[1] Human-Centered AI Lab, Medical University Graz, Graz, Austria
andreas.holzinger@human-centered.ai

[2] xAI Lab, Alberta Machine Intelligence Institute, Edmonton, Canada

Abstract. Enormous advances in the domain of statistical machine learning, the availability of large amounts of training data, and increasing computing power have made Artificial Intelligence (AI) very successful. For certain tasks, algorithms can even achieve performance beyond the human level. Unfortunately, the most powerful methods suffer from the fact that it is difficult to explain why a certain result was achieved on the one hand, and that they lack robustness on the other. Our most powerful machine learning models are very sensitive to even small changes. Perturbations in the input data can have a dramatic impact on the output and lead to entirely different results. This is of great importance in virtually all critical domains where we suffer from low data quality, i.e. we do not have the expected i.i.d. data. Therefore, the use of AI in domains that impact human life (agriculture, climate, health, ...) has led to an increased demand for trustworthy AI. Explainability is now even mandatory due to regulatory requirements in sensitive domains such as medicine, which requires traceability, transparency and interpretability capabilities. One possible step to make AI more robust is to combine statistical learning with knowledge representations. For certain tasks, it can be advantageous to use a human in the loop. A human expert can - sometimes, of course not always - bring experience, domain knowledge and conceptual understanding to the AI pipeline. Such approaches are not only a solution from a legal point of view, but in many application areas the "why" is often more important than a pure classification result. Consequently, both explainability and robustness can promote reliability and trust and ensure that humans remain in control, thus complementing human intelligence with artificial intelligence.

Keywords: Artificial intelligence · Trust · Explainable AI · Robustness · Human-in-the-loop

1 Success in Machine Learning Enabled a New AI Spring

At the interface between politics, industry and consumers, artificial intelligence is experiencing unprecedented popularity. Politicians around the world are declaring AI a strategic goal, industry sees it as a huge growth engine, and many

M. Kamp et al. (Eds.): ECML PKDD 2021 Workshops, CCIS 1524, pp. 427–440, 2021.
https://doi.org/10.1007/978-3-030-93736-2_33

application areas that impact human life (e.g. agriculture, climate, health, ...) see it as a great opportunity for multiple improvements in predictive modelling [4], diagnostics [39] and therapy [9].

For example, in July 2017, China's State Council published the country's artificial intelligence (AI) development strategy, titled "New Generation Artificial Intelligence Development Plan", with the goal of becoming a world leader in AI by 2030 [41]. Many other countries followed, e.g. in 2018 the German Federal Government with "AI made in Germany 2030"[1].

This recent AI spring was triggered by three main drivers: 1) the worldwide trend towards digitization, and thereby the 2) availability of big data, and above all 3) the remarkable advances in statistical machine learning and computational power.

The spread of AI solutions is accelerated by current events, a very recent example is the health domain: The potential for medical AI-based systems in the near future has increased enormously after the sad 200 million COVID-19 cases and 4 million deaths worldwide (as of 6 August 2021)[2]. To give another very recent example: A graph-based machine learning method enables the identification of bioactive anti-COVID-19 molecules in foods based on their ability to target the SARS-CoV-2 host gene (protein-protein) interactome. Based on this work, a "food map" was created that estimates the theoretical anti-COVID-19 potential of each ingredient based on the diversity and relative content of antivirally active candidates. Such approaches will play an important role in future clinical trials of precise nutritional interventions against COVID-19 and other viral diseases [32].

Indeed, the increased availability of data has reignited interest in AI algorithms for the medical domain, especially convolutional neural networks in image analysis and specifically in radiology and pathology. However, in order to use AI to solve problems in medicine and life sciences beyond the laboratory and routine, there is an urgent need to go beyond simple benchmarking and improve the performance of methods that only work with independent and identically distributed (i.i.d.) data. Independently and identically distributed random variables have the same distribution and do not affect each other - however, this is rarely the case with real data. Machine learning is learning from observed data by constructing stochastic models that can be used to make predictions and decisions. It sounds simple, but when do we have i.i.d. in reality, real-world data is highly non-linear, non-stationary and high-dimensional and often noisy. Data quality is therefore a basic requirement for the correct functioning of our data-driven algorithms. This often requires great efforts of data pre-processing, data cleansing, because malfunctions due to "dirty data" can have dramatic effects. However, this data cleaning can also have a negative impact on data quality, especially if not done carefully [48].

[1] https://www.ki-strategie-deutschland.de.

[2] https://www.worldometers.info/coronavirus/, accessed 6 August 2021.

Unfortunately, even the best current machine learning models do not generalize well, have difficulty with *small* training datasets ("little data") and are sensitive to even small perturbations as we will see later on.

Above all, the most successful approaches are so complex, so nonlinear, and so high-dimensional that they are difficult or impossible for human experts to interpret and, above all, can no longer derive causal relationships. Robustness and explainability have therefore been declared by the European Union to be definitely the most important properties for future trustworthy AI [17].

In this paper, we first define the terms trust and trustworthy AI, then explain what explainability and causability are and why this is important, briefly look at robustness, and conclude with how a human-in-the-loop can contribute to robustness and explainability.

2 Trust and Trustworthy AI

Before we dive into our topic, we need some definitions so that we can develop a common understanding of the terms used. Incidentally, this is also a good example of the *quality of explanations* and the question of when an explanation is good, which we will discuss in more detail later under the terminus Causability. As Jean Piaget (1896–1980) advocated in his tradition of human-centered and trans-disciplinary science, we first need a *common framework to enable mutual understanding* (see e.g. [37,38]).

2.1 What Is Trust?

Trust is a multidisciplinary concept that is very difficult to define [8], similarly as human intelligence [5], and this is why AI is also very difficult to define. The most common definition of intelligence is given by cognitive science as mental capability, and includes, among others, the ability to think abstract, to reason, and to solve problems from the real world. A hot topic in current AI/machine learning research is to find out whether and to what extent algorithms are able to learn such abstract thinking and reasoning similarly as humans can do - or whether the learning outcome remains on purely statistical correlation [22].

The concept of trust is also linked to several disciplines and influenced by many diverse factors. It can be understood better when we consider that trust has evolved genealogically from some fundamental features of human social life with the aim that we can *rely on other people to act cooperatively* [47].

As a social psychological construct, trust is a belief or an assessment, i.e. it is always *subjective* and dependent on personal attitudes and expectations. Nevertheless, trust has some consistent characteristics: subjectivity, dynamism, context awareness (risk situations, perceived domain importance, e.g. fields or situations that impact human life), incomplete transitivity, time decay, asymmetry and measurability. The basic factors include security, dependability, integrity, predictability, and reliability [53].

Trust evaluation is the process of quantifying trust with attributes that influence individual trust. A number of machine learning methods have even been used for trust assessment [51].

It is obvious that trust is very important for Human-AI interaction because trust is an attributional process and perceived trust is an important aspect of developing and maintaining interpersonal relationships. Successful cooperation between human communicators occurs when ambiguity and uncertainty in social perceptions are reduced through the development of trust. In particular, it is important to emphasise that the upcoming human-AI interaction represents a new paradigm in which human communication is augmented or even generated by an intelligent system. Here, trust develops differently than in classical early human-computer interaction or in interactions between humans [18].

2.2 What Is Trustworthy AI?

Despite all the successes and the recurring euphoria about AI, recent work shows that AI can unintentionally harm humans and that it is precisely the large-scale and wide introduction of AI technologies that holds enormous and unimagined potential for new types of unforeseen threats [26].

For example, AI can make unreliable decisions in safety-critical scenarios (e.g. in the medical domain) or undermine fairness by inadvertently discriminating against a group [16].

For this reason, the international research community has recently paid much attention to so-called *trustworthy AI*. Dimensions of trustworthy AI include: security, safety, privacy, non-discrimination, fairness, accountability (re-traceability, replicability), auditability and environmental Well-being, and most of all *robustness and explainability* [34]. These dimensions have also been included into the European Commissions ethics guidelines for trustworthy AI[3] [13,17]. For all these reasons, Trustworthy AI is a strongly emerging field in the international research community [7].

3 Explainability and Causability

3.1 What Is Explainable AI?

Although explainable AI (xAI) only emerged through the DARPA initiative [15] it is in principle not a new field. The problem of explainability is at least as old as AI itself, in fact it is the result of AI itself. DARPA's Explainable Artificial Intelligence (XAI) program aimed to develop AI systems whose models and decisions can be understood and trusted by end users. To this end, a large number of diverse methods have been developed by the growing xAI community. While also interpretable methods (aka ante-hoc methods) have been used, which are also known as glass-box models (such as decision trees, or graph-based methods),

[3] https://ec.europa.eu/commission/presscorner/detail/en/IP_19_1893, last accessed on August, 31, 2021.

the emphasis has been put by the community on the interpretation of so-called black-box methods (see below). The holistic integrative approach of the DARPA program is worth noting, as the realization of xAI has included not only methods for learning more explanatory models, but also the design of effective explanatory interfaces, as well as the psychological requirements for understanding effective explanations. Meanwhile xAI is an established vibrating field [1,44].

3.2 What Is Explainability?

Explainability, in the sense of the machine learning community focuses mostly on post-hoc methods, i.e. to make so-called "black-box" models explainable by a human [31]. Such methods highlight technically decision-relevant parts of machine representations and machine models, i.e., parts that contributed to model accuracy during training or to a particular prediction. A typical example of a method that does this very well is Layer Wise Relevance Propagation (LRP) [33]. With this method, heat maps can be used to visualize the parts that contributed to the given explanation. Graph Neural Networks (GNNs) are an increasingly popular approach for predicting graph-structured data, however, the input graphs are tightly entangled with the neural network structure, making traditional xAI approaches inapplicable on such graphs. Therefore, "GNN-LRP" [45] has already been further developed to provide explainability with graphs, or trees, as well. These methods can be very helpful in the biology, medicine and the life sciences, e.g. [24,25]. However, this "explainability" is a technical approach and does *not* relate to a human model. However, in certain domains, especially in the medical field, there is a need for causability, introduced by Holzinger et al. (2019) [23].

3.3 What Is Causability?

Causability is the measurable extent to which an explanation - resulting from an explainable AI method to a human expert achieves a specified level of causal understanding. Causability refers to a human model and can be measured with the System Causability Scale [21]. Causability is not a synonym for causality, instead the term causa-bil-ity was introduced in reference to usa-bil-ity. Whilst explainability (represented by the field of xAI) is about the technical implementation of transparency and traceability in AI approaches, causability is about measuring and ensuring the quality of explanations.

So let's briefly summarize: Explainability technically highlights decision relevant parts of machine representations and machine models, i.e. parts that have contributed to model accuracy in training or to a specific prediction for a given observation. This is already an important step and this is where the xAI community has already developed a variety of successful methods. However, explainability does not relate to a human model. Causability is the measurable extent to which an explanation (resulting from explainability) achieves a certain level of causal understanding for a human expert (or layperson, of course). Causal in the sense of Judea Pearl as relationship between cause and effect [35].

Why is this important? Because human understanding, especially checking whether and to what extent something has been understood, can only be guaranteed if we can map Explainability with Causability. Consequently, successful mapping between Explainability and Causability requires new human-AI interfaces that allow domain experts to interactively ask questions and counterfactuals to gain insight into the underlying explanatory factors of an outcome [20].

In an ideal world, statements originating from both "human intelligence" and "artificial intelligence" would be identical and congruent with the "ground truth," which is that it must be defined equally for humans and AI. That this is not easy and often does not work becomes quickly clear in the complex domain of medicine: medicine is a good exemplar of real-world challenges: (i) ground truth cannot always be precisely defined, especially for medical diagnoses; and (ii) human (scientific) models are often based on causality in the sense of Judea Pearl as the ultimate goal for understanding the underlying explanatory mechanisms.

While correlation is accepted as the basis for decisions in medical AI for a long time [42], it can only be considered as an intermediate step for causal considerations, which is relevant due to the importance of validity and necessary to build human trust [6].

Currently, there is much debate in the xAI community about avoiding bias and how to ensure fairness in AI decisions [30]. Bias is a core issue in causality, and causability is one possible measure of it. Validation of causal effects under particular causal structures is especially necessary when such effects are estimated in limited arrays. Randomized controlled trials are a good example. Such studies allow causal hypotheses to be tested because randomization by design is guaranteed, even with limited knowledge about the domain. A particular generalizability problem has been described by (Bareinboim & Pearl, 2013) [2], referred to as transportability, which can be viewed as a "data fusion framework" for external validation of intervention models and counterfactual queries. Transportability allows causal effects learned in experimental studies to be transferred to a new setup in which only observational studies can be conducted. Transportable models can be integrated into clinical guidelines to augment subject matter experts with "actionable" predictions to achieve better precision medicine [36].

The domain of artificial intelligence has tremendous potential to contribute to a better understanding of disease, which can lead to more accurate diagnoses, more rational disease prevention strategies, better treatment selection, and the development of new therapies. In addition, a better understanding of disease can contribute to the long-term goal of personalized precision medicine, which seeks to redefine the understanding of disease development and progression, treatment response, and health outcomes by measuring as precisely as possible the molecular, genetic, environmental, and behavioral individual factors that contribute to health and disease. Here, it is imperative that AI decisions be fully traceable across all modalities involved so that the medical professional has the ability to i) understand, ii) confirm, or iii) reject them. Whatever future human-AI interfaces look like, they must enable a domain expert to understand causal pathways in order to compute meaningful counterfactuals [28]. This is where the use of graphs and learning graph representations can be beneficial [24].

4 Robustness

4.1 Robustness in General

Many machine learning models achieve amazing performance on standard i.i.d. data. However when working with real data (e.g. from the medical domain) they fail miserably, being perturbed very easily. Robustness is generally defined as the property of a model to produce unperturbed results even if the input data is perturbed [50].

For example, it has been observed that commonly occurring image corruptions, such as random noise, contrast change, and blurring, can lead to significant performance degradation. Consequently, improving distributional robustness is an important step towards safely deploying models in complex, real-world settings [54]. Robustness is a ubiquitously observed property of biological systems and is considered a fundamental feature of complex evolvable systems. It is achieved by several underlying principles that apply to biological organisms as well as to sophisticated technical systems [29].

4.2 Robustness in Interventional Studies

Biomedical observational studies are affected by confounding and selection bias, which makes causal inference to be unfeasible if robust assumptions are not made. These require a priori domain knowledge, as data-driven prediction models may be used for drawing causal effects, but neither their parameters nor their predictions necessarily have a causal interpretation. The healthcare informatics communities are recommended to employ causal approaches and learn causal structures by using the *linchpins* to develop and test intervention models [40]: 1) target trials, 2) transportability, and 3) prediction invariance.

Target trials refer to algorithmic emulation of randomized studies. Transportability is a *license* to "transfer causal effects learned in experimental studies to a new population, in which only observational studies can be conducted." Akin to transportability is prediction invariance, where a "true causal model is contained in all prediction models whose accuracy does not vary across different settings".

When a causal structure is available or a target trial design can be devised, the evaluation of model transportability for a given set of action queries (e.g., treatment options or risk modifiers) is recommended; while for exploratory analyses where causal structures are to be discovered, prediction invariance could be used. In this way, as advocated by [40] transportability and prediction invariance could become guideline core tools and part of reporting protocols for intervention models, for a better alignment with the standards for prognostic and diagnostic models of medicine and biomedical practice today.

4.3 Robustness to Adversarial Attacks

Technically, we are talking about performance on unseen examples from the underlying distribution, and the goal is to train models so that the expected loss reaches a minimum:

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} = [\mathcal{L}(x,y;\theta)] \tag{1}$$

Szegedy et al. (2013) [49] made a fascinating discovery: several non-linear machine learning models, including state-of-the-art neural networks, misclassify well-known examples if they have been disturbed even slightly, e.g. with almost invisible salt and pepper noise. For example, a pig's face is suddenly classified as an airliner, a panda becomes a gibbon, or a benign melanoma becomes malignant (see picture) with extremely high confidence. Such total misclassifications can have dramatic effects in many application areas and do not contribute at all to trust building.

Several non-linear machine learning models, including state-of-the-art neural networks, also misclassify well-known examples if they have been disturbed even slightly, e.g. with almost invisible salt and pepper noise. The vehement however was that the output is completely misclassified, for example a pig face is suddenly classified as an airliner (see picture). This may be funny, however it can have dramatic effects in many application areas. For example, in medicine, such misclassifications can lead to serious consequences.

However, many non-linear ML models, particularly deep learning ones, falsely classify the so-called adversarial examples, i.e., inputs formed from small perturbations (e.g., salt-and-pepper noise) applied to training samples, which are usually not even visible for a time-limited human [10]. This results into dramatic effects, and completely wrong outputs with high confidence. A typical example is shown in Fig. 1.

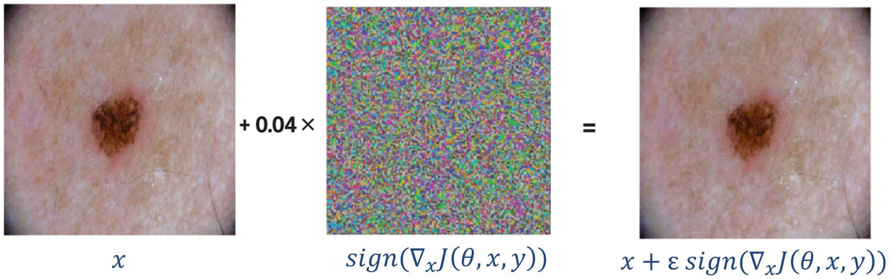
+ 0.04 ×
=
x
$sign(\nabla_x J(\theta, x, y))$
$x + \varepsilon\, sign(\nabla_x J(\theta, x, y))$

Fig. 1. Example of the susceptibility of our currently best performing deep learning models: One of the maybe most prominent adversarial example, cf. with [12]. Perturbations of just a few pixels ("salt-and-pepper noise") can dramatically change the classification and turn a malign melanoma into benign and vice versa.

Determining the appropriate Δ to use is a domain specific question, and therefore a human-in-the-loop [19] can be of help because humans, even if they also make mistakes, can be considered a robust proxy in decision making (see next chapter).

In Fig. 1 the θ are the feature parameters, x is the the input to the model, y is the targeted output function associated with x, and $J(\theta, x, y)$ is the cost function (loss function) used to train the neural network. The cost function around the current value of θ can be linearized, obtaining an optimal max-norm constrained pertubation of

$$\epsilon = sign(\nabla_x J(\theta, x, y)) \tag{2}$$

This is called "fast gradient sign method" of generating adversarial examples, the required gradient can be computed using backpropagation. The $sign(x)$ as real sign function maps from $\mathbb{R}$ to a Bit and adds a sign to it. The Nabla Operator denotes the vector representation of the differential operators (gradient, divergence, rotation). The property to resist such disturbances is called robustness, and achieving it can mean training models with *low expected adversarial loss*:

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} = [\max_{\delta\in\Delta} \mathcal{L}(x+\delta, y; \theta)]. \tag{3}$$

5 How Can a Human-in-the-Loop Help?

When we compare human learning and problem solving with the capabilities of our most advanced learning algorithms, some serious differences immediately stand out: Supervised learning requires a lot of labelled data while model-free reinforcement learning requires far too many trials. Humans, on the other hand, are able to generalise quickly and surprisingly well even in complex situations with little prior experience. Humans can generalise in a way that is different and more powerful than ordinary i.i.d. generalisation, namely these can correctly interpret novel combinations of existing concepts even when these combinations are extremely unlikely under training distribution, at least as long as they take into account higher-level syntactic and semantic patterns that have already been learned [46]. Humans are often very robust to change and can adapt quickly to change even with little training. Current Deep Learning is most successful in perceptual tasks and, more generally, in the previously mentioned System 1 tasks. The use of Deep Learning for System 2 tasks that require a deliberate sequence of steps is still in its infancy [3]. Humans are also very adept at inferring new causal relationships from even a few observations. Prior knowledge about the probability of occurrence of causal relationships of different kinds and the nature of the mechanisms linking causes and effects plays a crucial role in these inferences [14].

Let's stay again with an example from network medicine. Interactions of human experts on graphs clearly need to be based on the low-dimensional input features to efficiently discover, reject, or confirm causal links between biomedical modalities. Once these connections are computed and the structure of the input graphs is updated accordingly, methods for explainable GNN can be applied [52,55]. Human domain knowledge in the loop thereby enhances the model building process. Furthermore, human interactions can be realized here by "what-if" queries (counterfactuals) to the system, leading to a graph of counterfactuals in which features are defined as nodes and edges refer to combinations of features. Such a counterfactual graph can be generated in a purely data-driven manner: Given a test set that includes a sufficient number of samples, an algorithm traverses the feature space and exchanges feature values between nearest neighbors of a different class of results until the class of the instance itself changes. The nearest neighbor-based sampling leads to counterexamples of realistic patient profiles and is thus based on plausible counterfactuals. Of course, providing such counterfactuals, roughly based on the internals of a model, is not yet sufficient for explainability. The plausibility of the counterfactual change is therefore a must, i.e., the "counterfactual path" leading to the label change should have a real chance of occurring in practice for the counterfactual to be realistic. In this regard, recent attempts to find plausible counterfactuals for image classification should be extended to models for graph data. The sampled feature path leading to the class change is stored and forms a *contrafactual decision path*. Repeating this procedure results in a graph consisting of *multiple decision paths* that can be used as a communication channel back to the black box model. Recent work has shown how a DF can be efficiently reduced to a single decision tree [11,43], from which counterfactuals can be easily observed by the leaf nodes, so that it could be used as a model for *global* explanations. In such an approach, the *human-in-the-loop* will be able to study this consensus decision tree and thus adopt the changes to the counterfactual graph. Studying the impact of modifications to the counterfactual graph on the decision trees can facilitate the definition of symbolic rules to revise the internal structure of the input graph. Possible modifications include adding or deleting semantic links between modalities, however also adjusting their edge weights (reference) [24,36].

To implement such robust, explainable and thus trustworthy AI applications, we need an iterative, agile and human-centred AI design process. These processes have long been known in traditional software engineering as agile user-centred design methods [27] and now need to be taken to the next level of future AI engineers.

6 Conclusion

Explainability offers the great opportunity not only to meet the legal requirement for transparency and traceability of "black boxes", but also to promote trust in AI and, above all, to foster a deeper understanding of previously unknown connections - in other words, to contribute to the discovery of knowledge. Just think of the enormous support that doctors can draw from the combination of human intelligence and AI (e.g. in diagnosis): Humans show very good intuition in low-dimensional problems, can generalise amazingly well from a small amount of data, and recognise connections thanks to their everyday intelligence. For example, doctors can apply AI to "interesting" data and interrogate it interactively. Conversely, machine-generated results can be reconstructed from high-dimensional data spaces that no human could ever have found and checked for plausibility. The most important contribution of explainability is to clarify what is cause and what is effect in order to avoid falsely including artefacts and surrogates. This is desirable in many application domains and even mandatory in safety-critical domains.

The major challenges of our field for the future is to merge the two AI approaches, e.g. logic-based ontologies with probabilistic machine learning with one (or more or even many) humans in the loop into a hybrid multi-agent interaction model, used as a kind of "power steering for the brain". This would not only mean an extension (augmentation) of human intelligence by machine intelligence, but also, conversely, an extension of artificial intelligence by human intuition and thus an important contribution to making algorithms more robust.

Acknowledgements. The author declares that there are no conflicts of interests. This work does not raise any ethical issues. Parts of this work have been funded by the Austrian Science Fund (FWF), Project: P-32554, explainable AI.

References

1. Arrieta, A.B., et al.: Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Inf. Fusion **58**, 82–115 (2020). https://doi.org/10.1016/j.inffus.2019.12.012
2. Bareinboim, E., Pearl, J.: A general algorithm for deciding transportability of experimental results. arXiv:1312.7485 (2013)
3. Bengio, Y., Lecun, Y., Hinton, G.: Deep learning for AI. Commun. ACM **64**(7), 58–65 (2021). https://doi.org/10.1145/3448250
4. Biecek, P.: Dalex: explainers for complex predictive models in r. J. Mach. Learn. Res. **19**(1), 3245–3249 (2018)
5. Binet, A.: L'étude expérimentale de l'intelligence. Schleicher frères and cie, Paris (1903)
6. Cabitza, F., Campagner, A., Balsano, C.: Bridging the "last mile" gap between AI implementation and operation: "data awareness" that matters. Ann. Transl. Med. **8**(7), 501 (2020). https://doi.org/10.21037/atm.2020.03.63
7. Chatila, R., et al.: Trustworthy AI. In: Braunschweig, B., Ghallab, M. (eds.) Reflections on Artificial Intelligence for Humanity. LNCS (LNAI), vol. 12600, pp. 13–39. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-69128-8_2
8. Corazzini, J.G.: Trust as a complex multi-dimensional construct. Psychol. Rep. **40**(1), 75–80 (1977). https://doi.org/10.2466/pr0.1977.40.1.75
9. Donsa, K., Spat, S., Beck, P., Pieber, T.R., Holzinger, A.: Towards personalization of diabetes therapy using computerized decision support and machine learning: some open problems and challenges. In: Holzinger, A., Röcker, C., Ziefle, M. (eds.) Smart Health. LNCS, vol. 8700, pp. 237–260. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-16226-3_10

10. Elsayed, G.F., et al.: Adversarial examples that fool both human and computer vision. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Neural Information Processing Systems (NIPS 2018), pp. 1–11. NIPS Foundation (2018)
11. Fernández, R.R., De Diego, I.M., Aceña, V., Fernández-Isabel, A., Moguerza, J.M.: Random forest explainability using counterfactual sets. Inf. Fusion **63**(11), 196–207 (2020). https://doi.org/10.1016/j.inffus.2020.07.001
12. Finlayson, S.G., Bowers, J.D., Ito, J., Zittrain, J.L., Beam, A.L., Kohane, I.S.: Adversarial attacks on medical machine learning. Science **363**(6433), 1287–1289 (2019). https://doi.org/10.1126/science.aaw4399
13. Floridi, L.: Establishing the rules for building trustworthy AI. Nat. Mach. Intell. **1**(6), 261–262 (2019). https://doi.org/10.1038/s42256-019-0055-y
14. Griffiths, T.L., Sobel, D.M., Tenenbaum, J.B., Gopnik, A.: Bayes and blickets: effects of knowledge on causal induction in children and adults. Cogn. Sci. **35**(8), 1407–1455 (2011). https://doi.org/10.1111/j.1551-6709.2011.01203.x
15. Gunning, D., Aha, D.W.: Darpa's explainable artificial intelligence program. AI Mag. **40**(2), 44–58 (2019). https://doi.org/10.1609/aimag.v40i2.2850
16. Hajian, S., Bonchi, F., Castillo, C.: Algorithmic bias: from discrimination discovery to fairness-aware data mining. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM (2016). https://doi.org/10.1145/2939672.2945386
17. Hamon, R., Junklewitz, H., Sanche, I.: Robustness and Explainability of Artificial Intelligence - From technical to policy solutions. Publications Office of the European Union, Luxembourg (2020). https://doi.org/10.2760/57493
18. Hohenstein, J., Jung, M.: Ai as a moral crumple zone: the effects of AI-mediated communication on attribution and trust. Comput. Hum. Behav. **106**(2020). https://doi.org/10.1016/j.chb.2019.106190
19. Holzinger, A.: Interactive machine learning for health informatics: when do we need the human-in-the-loop? Brain Inf. **3**(2), 119–131 (2016). https://doi.org/10.1007/s40708-016-0042-6
20. Holzinger, A.: Explainable ai and multi-modal causability in medicine. Wiley i-com J. Interact. Media **19**(3), 171–179 (2020). https://doi.org/10.1515/icom-2020-0024
21. Holzinger, A., Carrington, A., Müller, H.: Measuring the quality of explanations: the system causability scale (SCS). KI - Künstliche Intelligenz **34**(2), 193–198 (2020). https://doi.org/10.1007/s13218-020-00636-z
22. Holzinger, A., Kickmeier-Rust, M., Müller, H.: KANDINSKY patterns as IQ-test for machine learning. In: Holzinger, A., Kieseberg, P., Tjoa, A.M., Weippl, E. (eds.) CD-MAKE 2019. LNCS, vol. 11713, pp. 1–14. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-29726-8_1
23. Holzinger, A., Langs, G., Denk, H., Zatloukal, K., Mueller, H.: Causability and explainability of artificial intelligence in medicine. Wiley Interdisc. Rev. Data Min. Knowl. Discov. **9**(4), 1–13 (2019). https://doi.org/10.1002/widm.1312
24. Holzinger, A., Malle, B., Saranti, A., Pfeifer, B.: Towards multi-modal causability with graph neural networks enabling information fusion for explainable AI. Inf. Fusion **71**(7), 28–37 (2021). https://doi.org/10.1016/j.inffus.2021.01.008
25. Holzinger, A., Mueller, H.: Toward human-AI interfaces to support explainability and causability in medical AI. IEEE Comput. **54**(10) (2021). https://doi.org/10.1109/MC.2021.3092610

26. Holzinger, A., Weippl, E., Tjoa, A.M., Kieseberg, P.: Digital transformation for sustainable development goals (SDGs) - a security, safety and privacy perspective on AI. In: Holzinger, A., Kieseberg, P., Tjoa, A.M., Weippl, E. (eds.) CD-MAKE 2021. LNCS, vol. 12844, pp. 1–20. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-84060-0_1
27. Hussain, Z., Slany, W., Holzinger, A.: Investigating agile user-centered design in practice: a grounded theory perspective. In: Holzinger, A., Miesenberger, K. (eds.) USAB 2009. LNCS, vol. 5889, pp. 279–289. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-10308-7_19
28. Kahneman, D.: Varieties of counterfactual thinking. In: Roese, N.J., Olson, J.M. (eds.) What might have been: The social psychology of counterfactual thinking. Taylor and Francis, New York (1995)
29. Kitano, H.: Biological robustness. Nat. Rev. Genet. **5**(11), 826–837 (2004). https://doi.org/10.1038/nrg1471
30. Kusner, M.J., Loftus, J.R.: The long road to fairer algorithms. Nature **578**, 34–36 (2020). https://doi.org/10.1038/d41586-020-00274-3
31. Lakkaraju, H., Arsov, N., Bastani, O.: Robust and stable black box explanations. In: Daumé, H., Singh, A. (eds.) International Conference on Machine Learning (ICML 2020), pp. 5628–5638. PMLR (2020)
32. Laponogov, I., et al.: Network machine learning maps phytochemically rich "hyperfoods" to fight covid-19. Human genomics **15**(1), 1–11 (2021). https://doi.org/10.1186/s40246-020-00297-x
33. Lapuschkin, S., Binder, A., Montavon, G., Mueller, K.R., Samek, W.: The LRP toolbox for artificial neural networks. J. Mach. Learn. Res. (JMLR) **17**(1), 3938–3942 (2016)
34. Liu, H., et al.: Trustworthy ai: A computational perspective. arXiv:2107.06641 (2021)
35. Pearl, J.: Causality: Models, Reasoning, and Inference, 2nd edn. Cambridge University Press, Cambridge (2009)
36. Pfeifer, B., Saranti, A., Holzinger, A.: Network module detection from multimodal node features with a greedy decision forest for actionable explainable AI. arXiv:2108.11674 (2021)
37. Piaget, J.: On the Development of Memory and Identity. Clark University Press, Worchester (1961)
38. Piaget, J., Inhelder, B.: Memory and Intelligence. Routledge, London (1973)
39. Ploug, T., Holm, S.: The four dimensions of contestable AI diagnostics-a patient-centric approach to explainable AI. Artif. Intell. Med. **107**(2020). https://doi.org/10.1016/j.artmed.2020.101901
40. Prosperi, M., et al.: Causal inference and counterfactual prediction in machine learning for actionable healthcare. Nat. Mach. Intell. **2**(7), 369–375 (2020). https://doi.org/10.1038/s42256-020-0197-y
41. Roberts, H., Cowls, J., Morley, J., Taddeo, M., Wang, V., Floridi, L.: The Chinese approach to artificial intelligence: an analysis of policy, ethics, and regulation. AI Soc. **36**(1), 59–77 (2020). https://doi.org/10.1007/s00146-020-00992-2
42. Roque, F.S., et al.: Using electronic patient records to discover disease correlations and stratify patient cohorts. PLoS Comput. Biol. **7**(8) (2011). https://doi.org/10.1371/journal.pcbi.1002141
43. Sagi, O., Rokach, L.: Explainable decision forest: transforming a decision forest into an interpretable tree. Inf. Fusion **61**, 124–138 (2020). https://doi.org/10.1016/j.inffus.2020.03.013

44. Samek, W., Montavon, G., Vedaldi, A., Hansen, L.K., Müller, K.-R. (eds.): Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. LNCS (LNAI), vol. 11700. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-28954-6
45. Schnake, T., et al.: Xai for graphs: Explaining graph neural network predictions by identifying relevant walks. arXiv:2006.03589 (2020)
46. Shepard, R.N.: Toward a universal law of generalization for psychological science. Science **237**(4820), 1317–1323 (1987). https://doi.org/10.1126/science.3629243
47. Simpson, J.A.: Psychological foundations of trust. Curr. Dir. Psychol. Sci. **16**(5), 264–268 (2007). https://doi.org/10.1111/j.1467-8721.2007.00517.x
48. Stoeger, K., Schneeberger, D., Kieseberg, P., Holzinger, A.: Legal aspects of data cleansing in medical AI. Comput. Law Secur. Rev. **42**(2021). https://doi.org/10.1016/j.clsr.2021.105587
49. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. arXiv:1312.6199 (2013)
50. Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., Madry, A.: Robustness may be at odds with accuracy. In: 7th International Conference on Learning Representations, ICLR 2019 (2019)
51. Wang, J., Jing, X., Yan, Z., Fu, Y., Pedrycz, W., Yang, L.T.: A survey on trust evaluation based on machine learning. ACM Comput. Surv. (CSUR) **53**(5), 1–36 (2020). https://doi.org/10.1145/3408292
52. Xu, K., Hu, W., Leskovec, J., Jegelka, S.: How powerful are graph neural networks? arXiv:1810.00826 (2018)
53. Yan, Z., Holtmanns, S.: Trust modeling and management: from social trust to digital trust. In: Subramanian, R. (ed.) Computer Security, Privacy and Politics: Current Issues, Challenges and Solutions, pp. 290–323. IGI Global (2008)
54. Yin, D., Lopes, R.G., Shlens, J., Cubuk, E.D., Gilmer, J.: A fourier perspective on model robustness in computer vision. arXiv:1906.08988 (2019)
55. Ying, Z., Bourgeois, D., You, J., Zitnik, M., Leskovec, J.: Gnnexplainer: generating explanations for graph neural networks. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alche Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems, pp. 9244–9255 (2019)