# Model Transparency: Why do we care?

Ioannis Papantonis & Vaishak Belle
*University of Edinburgh, UK*

Keywords: Explainable AI

Abstract: Artificial intelligence (AI) and especially machine learning (ML) has been increasingly incorporated into a wide range of critical applications, such as healthcare, justice, credit risk assessment, and loan approval. In this paper, we survey the motivations for caring about model transparency, especially as AI systems are becoming increasingly complex leviathans with many moving parts. We then briefly outline the challenges in providing computational solutions to transparency.

## 1 Introduction

Artificial intelligence (AI) and especially machine learning (ML) has been increasingly incorporated into a wide range of critical applications, such as healthcare (Kononenko, 2001; Loftus et al., 2019), justice (Chouldechova, 2017; Christin, 2017; Kleinberg et al., 2017), credit risk assessment (Chen et al., 2016; Finlay, 2011), and loan approval (Finlay, 2011; Wu et al., 2019a). At the same time, automated systems have an effect on casual everyday decisions, by recommending news articles (Alvarado and Waern, 2018), movies (Bennett et al., 2007), and music (Mehrotra et al., 2018). In the core of this ML predominance lies the expectation that models can be more accurate than humans (Poursabzi-Sangdeh et al., 2021a), something that has already been demonstrated in various cases (Culverhouse et al., 2003; Goh et al., 2020; Hilder et al., 2009; Grove et al., 2000). Having said that, employing algorithms even as recommendation systems for cultural products (like movies) makes them part of the human culture, since they not only handle cultural products, but they also influence peoples' decisions and perceptions (Gillespie, 2016). This also means that they should not be viewed as mere tools (Bozdag, 2013), but rather as entities that hold their own values (Alvarado and Waern, 2018).

As such, it is paramount to make sure that their values align with those of human's, thus enabling ML's responsible integration to society (Russell et al., 2015; Gabriel, 2020; Christian, 2020). This need is further magnified by several recent instances of automated systems perpetuating undesired historical human biases, such as Amazon's recruitment algorithm exhibiting misogynistic behaviour (Meyer, 2018), or commercial systems utilized by the US criminal justice system being extremely biased against black defendants (Angwin et al., 2016; Dressel and Farid, 2018). Apart from that, ML failures can arise, for example, due to misuse, as in the case of an individual who spent an extra year in prison due to a typographical error in one of the inputs that was given to the ML system (Wexler, 2017). Of course, poor model design is another major source of catastrophic failures with far-reaching implications, such as putting people in danger due to inaccurate air quality assessment (McGough, 2018), or by providing life-threatening cancer treatment recommendations (Strickland, 2019; Ross and Swetlitz, 2018). These and other similar pitfalls, along with the consequences and confusion that come with them (Galanos, 2019; Aleksander, 2017), have led to some extreme arguments about ML potentially eroding the social fabric and even posing a threat to society's democratic foundation (Bozdag and Van Den Hoven, 2015).

In light of such concerns, it is becoming increasingly clear that proactive actions need to be taken on a large scale in order to avoid bleak future situations. The urgency of this matter is reflected, for example, in the Declaration of Cooperation on Artificial Intelligence signed by the members of the European Union (EU).[1] This development was followed by the formation of an expert group on AI,[2] with the goal of anchoring the development of AI that is both successful and ethically sound. *Transparency* was a central notion highlighted in this call, and especially its relationship with *trustworthiness*, which is one of the end goals of this initiative. As the Commission Vice-

[1] https://digital-strategy.ec.europa.eu/en/news/eu-member-states-sign-cooperate-artificial-intelligence

[2] https://ec.europa.eu/commission/presscorner/detail/en/IP_18_1381

President for the Digital Single Market, Andrus Ansip, noted: "*As always with the use of technologies, trust is a must*". Subsequent research outputs of the resulting group have further emphasized the importance of transparency,[3] listing it as one of the seven key requirements in order to achieve *trustworthy AI*.[4]

At the same time, several additional large scale initiatives regarding the responsible integration of AI have been taken, such as:

- The Asilomar AI Principles, with the support of the Future of Life Institute.[5]
- The Montreal Declaration for Responsible AI, with the support of the University of Montreal.[6]
- The General Principles, with contributions from 250 thought leaders throughout the world.[7]
- The Tenets of the Partnership on AI, with contributions from stakeholders coming from diverse fields that make use of AI (academia, industry, etc).[8]

It is worth noting that a comparative meta-analysis found the above sets of principles to greatly overlap, ensuring that the scientific/regulatory/investing communities have reached a satisfying level of consensus (Floridi et al., 2018). Again, transparency was recognized as a core component that should drive the integration of automated systems, present in all initiatives, although the terminology was somehow inconsistent[9]. Finally, an additional survey that analyzed 84 different ethical guidelines for AI found that transparency was the most common principle among them, called for in 73 of them (Jobin et al., 2019).

## 2 Dimensions of Transparency

Having established the need for transparency, as the ability to understand AI/ML, a natural next step is to define all the notions it should encompass. At this point, it should be noted that transparency is by no means a new concept by itself, rather it has a long history in an array of disciplines (Margetts, 2011; Hood and Heald, 2006). Despite that, AI/ML adds a unique dimension to it, since other disciplines rarely face the issue of employing tools with black-box design where the decision process itself is elusive (Floridi et al., 2018). This challenge has been a contributing factor to the surge in related publications, which increase by about 100% every other year (Larsson et al., 2019).

While this is an impressive rate, stakeholders outside the AI community have indicated that future developments should be predicated on further mutual engagement between the two parties (non-AI and AI) (Bhatt et al., 2020b; Bhatt et al., 2020a). This is a reasonable concern, since although the AI community has produced a vast literature during the last decade, the majority of the scientific output addresses only the technical side of transparency, through the lens of *model transparency* and *model explainability* (Arrieta et al., 2020). The former paradigm advocates in favour of utilizing "transparent" (or white-box) models, meaning that their design allows for readily inspecting their inner workings (Linardatos et al., 2020), such as rule-based classifiers or regression analysis. On the other hand, the latter approach, which is also known as *explainability in AI (XAI)*, develops post-hoc techniques that can provide explanations and information about the decision process of black-box models, i.e. models with an overly complex design that does not allow for gaining any meaningful insights, such as neural networks or random forests (Guidotti et al., 2018). These are both essential research directions, however, comparing them to the notion of transparency discussed so far, they seems rather narrow-scoped, focusing only on the technical side of achieving a transparent AI integration.

This observation has motivated a series of works that advocate in favour of expanding the scope of transparency, as used in the AI/ML community, to encompass a wider range of goals (Mittelstadt et al., 2019), in line with the calls mentioned earlier in this section. More specifically, some important directions that need to be incorporated into the AI community's agenda are related to:

- Providing guidelines regarding the appropriate way to utilize and explain AI systems, referred to as *competence*.
- Building an environment of trust,[10] which "*can only be achieved by ensuring an appropriate involvement by human beings in relation to high-risk AI applications*".[11]

[3] https://ec.europa.eu/info/publications/white-paper-artificial-intelligenceeuropean-approach-excellence-and-trust_en

[4] https://ec.europa.eu/digital-singlemarket/en/news/ethics-guidelines-trustworthy-ai

[5] https://futureofife.org/ai-principles

[6] https://www.montrealdeclaration-responsibleai.com/the-declaration

[7] http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html

[8] https://www.partnershiponai.org/∼tenets/

[9] The authors in (Floridi et al., 2018) introduce the term *Explicability*, to overcome this inconsistency, however they define it in terms of transparency.

[10] https://digital-strategy.ec.europa.eu/en/news/eu-member-states-sign-cooperate-artificial-intelligence

[11] https://ec.europa.eu/info/publications/white-paper-

This is not an exhaustive list, as, for example, additional dimensions that incorporate legal aspects can potentially be fostered under this expanded notion of transparency (Larsson, 2019). However, both of these aspects can have an immediate positive impact, considering that AI/ML systems are already deployed, so their correct and responsible use should be top priority.

# 3 Directions

Given the dimensions identified above, we briefly discuss some research directions in both the technical camp (transparent models) as well as the sociotechnical camp (stakeholder engagement), and outline challenges in both.

## 3.1 Transparent Models

Clearly, the use of transparent models – i.e., models that are transparent by design, have interpretable features, are human-readable, etc – is motivated by their ability to allow users to understand their inner workings. Let us examine a candidate, for concreteness, but also mention challenges that arise with it.

- **Probabilistic models.** For the sake of concreteness, consider Bayesian networks, and other types of graphical models (Pearl, 1988). Bayesian networks (BNs) are a class of probabilistic models that represent relationships between variables by using direct (usually acyclic) graphs (Darwiche, 2009). This has the very appealing advantage of clearly expressing dependencies in the data, by only drawing arrows between variables. Furthermore, once the BN is specified, graphical tests can accurately recover all conditional independencies, without the need to perform any algebraic manipulations (Geiger et al., 1990). Due to these properties BNs are arguably one of the most transparent model classes, since their internal representation (and its implications) can be easily inspected, by construction. It is this strength that has turned BNs into the backbone of causal inference, too; causal relationships are represented through a BN, while graphical criteria identify which causal effects can be estimated using observational data (Pearl, 2009). Naturally, BNs have found numerous applications in many integral applications (Kalet et al., 2015; Castelletti and Soncini-Sessa, 2007; Shenton et al., 2014; Uusitalo, 2007; Stewart-Koster et al., 2010; Friis-Hansen, 2000).
- **Expert knowledge.** In addition to the above, BNs allow for incorporating various forms of a priori constraints, such as temporal ones (Dechter et al., 1991). Of course, probabilistic independence constraints can be encoded as well during model design, by directly adjusting the topology of the directed graph. The combination of all these properties as well as the ability to infer causal relationships, instead of correlations, offers a powerful alternative to black-box models, especially when considering high-stakes applications (Rudin, 2019; Rudin et al., 2022).
- **Computational hurdles.** While BNs come with significant advantages, a downside is that inference using them is intractable, in the sense that computing marginal probabilities is NP-hard (Cooper, 1990). On top of that, specialized routines are required to perform the inferential step. This is the main motivation behind the recent emergence of so-called tractable probabilistic models (TPMs) (Poon and Domingos, 2011), as an alternative approach that generalizes traditional BNs. TPMs directly encode the joint distribution of a set of variables, in a way that allows for a simple mechanism for performing inference. Furthermore, they can potentially lead to exponential savings in both inference time and storing space (Darwiche, 2003). Consequently, TPMs have gathered significant attention in many applications (Bekker et al., 2014).

  A downside, however, is that TPMs are represented as computational graphs that do not allow for directly inspecting the relationships between the variables. Furthermore, incorporating probabilistic constraints is not immediate, as in the BN case. In fact, it is unclear whether it is at all feasible. These challenges effectively turn TPMs into black-box models, despite them being closely related to one of the most transparent model classes.

  The same kind of computation vs transparency dichotomy can be observed in so-called variational approaches (Srivastava and Sutton, 2017). These models avoid the explicit computation of probabilities at run-time by training, say, neural networks on the distribution encoded in the model. But the downside is that neural networks are not interpretable, and although there is considerable work on unwrapping the functionings of neural networks (Sharma et al., 2019; Belle and Papantonis, 2020), either by inspecting the internal nodes or doing post-hoc analysis, transparency as a first-class object is ultimately lost.

artificial-intelligenceeuropean-approach-excellence-and-trust_en

- **Balancing transparency and computation.** Perhaps a midpoint in between these extremes is the emerging work on statistical relational learning and neuro-symbolic AI (Gutmann et al., 2011; Hu et al., 2016). The idea is to empower probabilistic and deep learning models with logical templates, either as a specification language, a training function, or a classification target so that (respectively) experts can encode their knowledge using logic, perform data-efficient learning using domain-specific logical rules, or extract logical rules for post-hoc inspection. It remains to be seen whether this line of work will bear more fruit in the long run.

We refer readers interested in learning about other types of transparent models to (Arrieta et al., 2019; Mehrabi et al., 2021; Belle and Papantonis, 2020)

## 3.2 User Competence and Trustworthiness

Model transparency is an essential component for important applications, but, as argued earlier, it is rather narrow scoped, since it does not address the perplexing complexity of incorporating AI into society. Engaging with parties that either use or are affected by AI can lead to significant advances in terms of ensuring its proper use as well as establishing a trusting human-AI relationship. In fact, there seems to be a strong link between these two desiderata, supported by evidence suggesting that users' understanding and competence have a great influence on the amount of trust placed upon an automated system (Balfe et al., 2018; Sheridan and Telerobotics, 1992; Merritt and Ilgen, 2008). Fostering trust then, may have direct implications on the adoption of AI in practical applications (Linegang et al., 2006; Wright et al., 2019).

Having said that, trust is not a binary distinction where one can only trust or distrust a model. Rather than that, a user might trust a model's outcomes for a certain sub-population, while being suspicious of decisions concerning another sub-population (perhaps one that is under-represented in the dataset). This showcases that trust is a thing to be adjusted or *calibrated*, so models are employed appropriately (Zhang et al., 2020). Failing to do so, can potentially lead to an over-reliance on the model's decisions (Cummings, 2004) or model aversion, where users entirely dismiss a model after a few mistakes are made (Dietvorst et al., 2015).

There is already a considerable body of work that explores the use of XAI generated explanations as a means to enhance trust (Mahbooba et al., 2021; Guo, 2020; Gunning et al., 2019; Gunning and Aha, 2019). This is a fairly reasonable approach, resembling the way that humans justify their decisions by providing information that is relevant to their decision making process. Despite that, recent studies provide evidence both in favour (Lai et al., 2020; Lai and Tan, 2019) and against (Poursabzi-Sangdeh et al., 2021b; Chu et al., 2020; Carton et al., 2020) the utility of explanations in making a model's internal reasoning clear to users. This has raised concerns about the way users perceive XAI explanations overall, calling for additional surveys to shed some light on this topic (Doshi-Velez and Kim, 2017; Vaughan and Wallach, 2020).

Along this line, there is also alarming evidence suggesting that practitioners utilize XAI techniques in a wrong way (Kaur et al., 2020). An important observation here is that misuse may arise both due to an incomplete technical understanding of XAI, as well as due to misunderstandings regarding XAI's intended use. This situation clearly impedes trust calibration, and thus achieving transparency in AI's social integration. Here are some avenues by means of which we, as a community, can contribute to the understanding and embedding of trust:

- **Clarify use.** We need to establish frameworks that establish the proper use of XAI, while also developing a framework that can be used in order to calibrate trust between human users and AI.
- **Explicate limitations.** While there is a plethora of technical XAI contributions, studying the advantages and limitations of each explanation type, as well as ways they can be combined to convey a more complete picture of a model's decision making process, has not received as much attention by the AI community. We need to identify the most prominent explanation types and techniques, discuss the kind of insights each one offers, and suggest conceptual frameworks to further emphasize their distinctions. Finally, we need to propose ways to combine multiple explanations together in order to gain a more well-versed understanding of a model. Some recent surveys such as (Arrieta et al., 2019; Mehrabi et al., 2021; Wu et al., 2019b; Belle and Papantonis, 2020) are starting to paint such a picture.
- **Education in XAI.** As mentioned earlier, practitioners face various kinds of challenges when applying XAI techniques, most of them stemming from their incomplete understanding of the field. A natural step to address this issue would be to offer the affected parties sufficient education to appropriately understand and apply the right techniques. However, there is a stark lack of academic resources on XAI, such as university level

courses. Of course, there are online articles discussing related things, but this is not a holistic, systematic approach. In fact, there is only a single academic course on XAI, offered by Harvard University (Lakkaraju and Lage, 2019), as well as some tutorials (Samek and Montavon, 2020; Camburu and Akata, 2021), but they are usually intended for researchers. We need to provide guidelines for implementing and delivering courses, including coding assignments with concrete feedback.

- **Trust Calibration and Model Comprehension** One of the ultimate goals of XAI is to facilitate building trusting relationships between users and AI. While educating people on the technical details and underlying principles is a step towards this goal, there are additional factors to be considered to ensure proper use. A concerning finding is that data scientists might understand explanations, but instead of using them in order to further inspect a model, they use them to construct narratives to convince themselves that the model performs as it should (Kaur et al., 2020).

## 4 Conclusions

We have briefly surveyed the importance of model transparency and suggested some directions for future work. We consider both technical directions, discussing the computation vs expressiveness tradeoff, and socio-technical ones. We hope we convey the urgency of the matter to readers, and that they are encouraged to come up with novel solutions that promote the responsible and safe integration of AI into critical social applications.

## 5 Acknowledgements

This research was partly supported by a Royal Society University Research Fellowship, UK, and partly supported by a grant from the UKRI Strategic Priorities Fund, UK to the UKRI Research Node on Trustworthy Autonomous Systems Governance and Regulation (EP/V026607/1, 2020–2024).

## REFERENCES

Aleksander, I. (2017). Partners of humans: a realistic assessment of the role of robots in the foreseeable future. *Journal of Information Technology*, 32(1):1–9.

Alvarado, O. and Waern, A. (2018). Towards algorithmic experience: Initial efforts for social media contexts. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 1–12, New York, NY, USA. Association for Computing Machinery.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). Machine bias. In *Ethics of Data and Analytics*, pages 254–264. Auerbach Publications.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al. (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115.

Arrieta, A. B., Rodríguez, N. D., Ser, J. D., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., and Herrera, F. (2019). Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *CoRR*, abs/1910.10045.

Balfe, N., Sharples, S., and Wilson, J. R. (2018). Understanding is key: An analysis of factors pertaining to trust in a real-world automation system. *Human factors*, 60(4):477–495.

Bekker, J., Davis Leuven, J. K., Choi, A., Darwiche, A., and Van den Broeck, G. (2014). Tractable Learning for Complex Probability Queries.

Belle, V. and Papantonis, I. (2020). Principles and practice of explainable machine learning.

Bennett, J., Lanning, S., and Netflix, N. (2007). The netflix prize. In *In KDD Cup and Workshop in conjunction with KDD*.

Bhatt, U., Andrus, M., Weller, A., and Xiang, A. (2020a). Machine learning explainability for external stakeholders. *arXiv preprint arXiv:2007.05408*.

Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., Ghosh, J., Puri, R., Moura, J. M., and Eckersley, P. (2020b). Explainable machine learning in deployment. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 648–657.

Bozdag, E. (2013). Bias in algorithmic filtering and personalization. *Ethics and information technology*, 15(3):209–227.

Bozdag, E. and Van Den Hoven, J. (2015). Breaking the filter bubble: democracy and design. *Ethics and information technology*, 17(4):249–265.

Camburu, O.-M. and Akata, Z. (2021). Natural-xai: Explainable ai with natural language explanations. International Conference on Machine Learning.

Carton, S., Mei, Q., and Resnick, P. (2020). Feature-based explanations don't help people detect misclassifications of online toxicity. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 95–106.

Castelletti, A. and Soncini-Sessa, R. (2007). Bayesian networks and participatory modelling in water resource management. *Environmental Modelling & Software*, 22(8):1075–1088.

Chen, N., Ribeiro, B., and Chen, A. (2016). Financial credit risk assessment: A recent review. *Artif. Intell. Rev.*, 45(1):1–23.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2):153–163. PMID: 28632438.

Christian, B. (2020). *The alignment problem: Machine learning and human values*. WW Norton & Company.

Christin, A. (2017). Algorithms in practice: Comparing web journalism and criminal justice. *Big Data & Society*, 4(2):2053951717718855.

Chu, E., Roy, D., and Andreas, J. (2020). Are visual explanations useful? a case study in model-in-the-loop prediction. *arXiv preprint arXiv:2007.12248*.

Cooper, G. F. (1990). The computational complexity of probabilistic inference using bayesian belief networks. *Artificial intelligence*, 42(2-3):393–405.

Culverhouse, P. F., Williams, R., Reguera, B., Herry, V., and González-Gil, S. (2003). Do experts make mistakes? a comparison of human and machine indentification of dinoflagellates. *Marine ecology progress series*, 247:17–25.

Cummings, M. (2004). Automation bias in intelligent time critical decision support systems. In *AIAA 1st intelligent systems technical conference*, page 6313.

Darwiche, A. (2003). A differential approach to inference in bayesian networks. *Journal of the ACM (JACM)*, 50(3):280–305.

Darwiche, A. (2009). *Modeling and reasoning with Bayesian networks*. Cambridge university press.

Dechter, R., Meiri, I., and Pearl, J. (1991). Temporal constraint networks. *Artificial intelligence*, 49(1-3):61–95.

Dietvorst, B. J., Simmons, J. P., and Massey, C. (2015). Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1):114.

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

Dressel, J. and Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4(1):eaao5580.

Finlay, S. (2011). Multiple classifier architectures and their application to credit risk assessment. *European Journal of Operational Research*, 210(2):368–378.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., et al. (2018). Ai4people—an ethical framework for a good ai society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, 28(4):689–707.

Friis-Hansen, A. (2000). Bayesian networks as a decision support tool in marine applications.

Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and machines*, 30(3):411–437.

Galanos, V. (2019). Exploring expanding expertise: artificial intelligence as an existential threat and the role of prestigious commentators, 2014–2018. *Technology Analysis & Strategic Management*, 31(4):421–432.

Geiger, D., Verma, T., and Pearl, J. (1990). d-separation: From theorems to algorithms. In *Machine Intelligence and Pattern Recognition*, volume 10, pages 139–148. Elsevier.

Gillespie, T. (2016). *Trendingistrending: When Algorithms Become Culture*. Routledge.

Goh, Y., Cai, X., Theseira, W., Ko, G., and Khor, K. (2020). Evaluating human versus machine learning performance in classifying research abstracts. *Scientometrics*, 125.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., and Nelson, C. (2000). Clinical versus mechanical prediction: a meta-analysis. *Psychological assessment*, 12(1):19.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42.

Gunning, D. and Aha, D. (2019). Darpa's explainable artificial intelligence (xai) program. *AI magazine*, 40(2):44–58.

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., and Yang, G.-Z. (2019). Xai—explainable artificial intelligence. *Science Robotics*, 4(37):eaay7120.

Guo, W. (2020). Explainable artificial intelligence for 6g: Improving trust between human and machine. *IEEE Communications Magazine*, 58(6):39–45.

Gutmann, B., Thon, I., Kimmig, A., Bruynooghe, M., and De Raedt, L. (2011). The magic of logical inference in probabilistic programming. *TPLP*, 11:663–680.

Hilder, S., Harvey, R. W., and Theobald, B.-J. (2009). Comparison of human and machine-based lip-reading. In *AVSP*, pages 86–89.

Hood, C. and Heald, D. (2006). *Transparency in historical perspective*. Number 135. Oxford University Press.

Hu, Z., Ma, X., Liu, Z., Hovy, E. H., and Xing, E. P. (2016). Harnessing deep neural networks with logic rules. *CoRR*, abs/1603.06318.

Jobin, A., Ienca, M., and Vayena, E. (2019). The global landscape of ai ethics guidelines. *Nature Machine Intelligence*, 1(9):389–399.

Kalet, A. M., Gennari, J. H., Ford, E. C., and Phillips, M. H. (2015). Bayesian network models for error detection in radiotherapy plans. *Physics in Medicine & Biology*, 60(7):2735.

Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., and Wortman Vaughan, J. (2020). Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–14.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., and Mullainathan, S. (2017). Human Decisions and Machine Predictions*. *The Quarterly Journal of Economics*, 133(1):237–293.

Kononenko, I. (2001). Machine learning for medical diagnosis: history, state of the art and perspective. *Artif. Intell. Medicine*, 23(1):89–109.

Lai, V., Liu, H., and Tan, C. (2020). " why is' chicago'deceptive?" towards building model-driven tutorials for humans. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13.

Lai, V. and Tan, C. (2019). On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 29–38.

Lakkaraju, H. and Lage, I. (2019). Interpretability and explainability in machine learning.

Larsson, S. (2019). The socio-legal relevance of artificial intelligence. *Droit et societe*, (3):573–593.

Larsson, S., Anneroth, M., Felländer, A., Felländer-Tsai, L., Heintz, F., and Ångström, R. C. (2019). Sustainable ai: An inventory of the state of knowledge of ethical, social, and legal challenges related to artificial intelligence.

Linardatos, P., Papastefanopoulos, V., and Kotsiantis, S. (2020). Explainable ai: A review of machine learning interpretability methods. *Entropy*, 23(1):18.

Linegang, M. P., Stoner, H. A., Patterson, M. J., Seppelt, B. D., Hoffman, J. D., Crittendon, Z. B., and Lee, J. D. (2006). Human-automation collaboration in dynamic mission planning: A challenge requiring an ecological approach. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 50, pages 2482–2486. SAGE Publications Sage CA: Los Angeles, CA.

Loftus, T., Tighe, P., Filiberto, A., Efron, P., Brakenridge, S., Mohr, A., Rashidi, P., Upchurch, G., and Bihorac, A. (2019). Artificial intelligence and surgical decision-making. *JAMA Surgery*, 155.

Mahbooba, B., Timilsina, M., Sahal, R., and Serrano, M. (2021). Explainable artificial intelligence (xai) to enhance trust management in intrusion detection systems using decision tree model. *Complexity*, 2021.

Margetts, H. (2011). The internet and transparency. *The Political Quarterly*, 82(4):518–521.

McGough, M. (2018). *How Bad is Sacramento's Air, Exactly? Google Results Appear at Odds with Reality, some say*, chapter [online] Available: https://www.sacbee.com/news/california/fires/article216227775.html.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6).

Mehrotra, R., McInerney, J., Bouchard, H., Lalmas, M., and Diaz, F. (2018). Towards a fair marketplace: Counterfactual evaluation of the trade-off between relevance, fairness satisfaction in recommendation systems. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, pages 2243–2251, New York, NY, USA. Association for Computing Machinery.

Merritt, S. M. and Ilgen, D. R. (2008). Not all trust is created equal: Dispositional and history-based trust in human-automation interactions. *Human factors*, 50(2):194–210.

Meyer, D. (2018). *Amazon Reportedly Killed an AI Recruitment System Because It Couldn't Stop the Tool from Discriminating Against Women*, chapter [online] Available: https://fortune.com/2018/10/10/amazon-ai-recruitment-bias-women-sexist/.

Mittelstadt, B., Russell, C., and Wachter, S. (2019). Explaining explanations in ai. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 279–288.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.

Pearl, J. (2009). *Causality: Models, Reasoning and Inference*. Cambridge University Press, USA, 2nd edition.

Poon, H. and Domingos, P. (2011). Sum-product networks: A new deep architecture. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 689–690.

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J. W., and Wallach, H. (2021a). Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA. Association for Computing Machinery.

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J. W., and Wallach, H. (2021b). Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–52.

Ross, C. and Swetlitz, I. (2018). Ibm's watson supercomputer recommended 'unsafe and incorrect'cancer treatments, internal documents show. *Stat*, 25.

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.

Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., and Zhong, C. (2022). Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistics Surveys*, 16:1–85.

Russell, S., Dewey, D., and Tegmark, M. (2015). Research priorities for robust and beneficial artificial intelligence. *Ai Magazine*, 36(4):105–114.

Samek, W. and Montavon, G. (2020). Explainable ai for deep networks. European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases.

Sharma, S., Henderson, J., and Ghosh, J. (2019). CERTIFAI: counterfactual explanations for robustness, transparency, interpretability, and fairness of artificial intelligence models. *CoRR*, abs/1905.07857.

Shenton, W., Hart, B. T., and Chan, T. U. (2014). A bayesian network approach to support environmental flow restoration decisions in the yarra river, australia. *Stochastic Environmental Research and Risk Assessment*, 28(1):57–65.

Sheridan, T. B. and Telerobotics, A. (1992). *Human supervisory control*. Cambridge, MA: MIT Press.

Srivastava, A. and Sutton, C. (2017). AUTOENCODING VARIATIONAL INFERENCE FOR TOPIC MODELS.

Stewart-Koster, B., Bunn, S., Mackay, S., Poff, N., Naiman, R. J., and Lake, P. S. (2010). The use of bayesian networks to guide investments in flow and catchment restoration for impaired river ecosystems. *Freshwater Biology*, 55(1):243–260.

Strickland, E. (2019). Ibm watson, heal thyself: How ibm overpromised and underdelivered on ai health care. *IEEE Spectrum*, 56(4):24–31.

Uusitalo, L. (2007). Advantages and challenges of bayesian networks in environmental modelling. *Ecological modelling*, 203(3-4):312–318.

Vaughan, J. W. and Wallach, H. (2020). A human-centered agenda for intelligible machine learning. *Machines We Trust: Getting Along with Artificial Intelligence*.

Wexler, R. (2017). When a computer program keeps you in jail. *New York Times*.

Wright, J. L., Chen, J. Y., and Lakhmani, S. G. (2019). Agent transparency and reliability in human–robot interaction: the influence on user confidence and perceived reliability. *IEEE Transactions on Human-Machine Systems*, 50(3):254–263.

Wu, M., Huang, Y., and Duan, J. (2019a). Investigations on classification methods for loan application based on machine learning. In *2019 International Conference on Machine Learning and Cybernetics (ICMLC)*, pages 1–6.

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., and Yu, P. S. (2019b). A comprehensive survey on graph neural networks. *CoRR*, abs/1901.00596.

Zhang, Y., Liao, Q. V., and Bellamy, R. K. (2020). Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 295–305.