ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0. Available from http://www.i6doc.com/en/.

Societal Issues in Machine Learning: When Learning from Data is Not Enough

Davide Bacciu[1], Battista Biggio[2], Paulo J. G. Lisboa[3],
José D. Martín[4], Luca Oneto[1], Alfredo Vellido[5,6]

[1] Department of Computer Science - University of Pisa - Italy
[2] University of Cagliari - Italy, and Pluribus One - Italy
[3] Department of Applied Mathematics - Liverpool John Moores University - U.K.
[4] Department of Electronic Engineering - Universitat de València - Spain
[5] Intelligent Data Science and Artificial Intelligence Research Center (IDEAI)
Universitat Politècnica de Catalunya-BarcelonaTech - Spain
[6] Centro de Investigación Biomédica en Red - CIBER-BBN - Spain

Abstract. It has been argued that Artificial Intelligence (AI) is experiencing a fast process of commodification. Such characterization is on the interest of big IT companies, but it correctly reflects the current industrialization of AI. This phenomenon means that AI systems and products are reaching the society at large and, therefore, that societal issues related to the use of AI and Machine Learning (ML) cannot be ignored any longer. Designing ML models from this human-centered perspective means incorporating human-relevant requirements such as safety, fairness, privacy, and interpretability, but also considering broad societal issues such as ethics and legislation. These are essential aspects to foster the acceptance of ML-based technologies, as well as to ensure compliance with an evolving legislation concerning the impact of digital technologies on ethically and privacy sensitive matters. The ESANN special session for which this tutorial acts as an introduction aims to showcase the state of the art on these increasingly relevant topics among ML theoreticians and practitioners. For this purpose, we welcomed both solid contributions and preliminary relevant results showing the potential, the limitations and the challenges of new ideas, as well as refinements, or hybridizations among the different fields of research, ML and related approaches in facing real-world problems involving societal issues.

1 Introduction

Machine learning (ML) based technologies are increasingly tapping on information concerning many aspects of our lives, therefore affecting them in the context of society. This fact has forced institutions to regulate the way in which data are stored and manipulated (e.g., the European General Data Protection Regulation - GDPR) and to interrogate themselves on the ethical issues (e.g. discrimination) that may arise from the global adoption of these technologies.

For these reasons, ML, beyond being able to extract useful information from data, has now to take on board requirements that stem from its use in areas with clear societal impact, from governance to defence and from public health to urban planning, such as being able to guarantee the privacy of the data, or fairness with respect to minority groups, to name a few.

The remainder of the paper is structured as follows. First, different societal aspects of the use of ML and related techniques are separately considered and discussed, focusing on some of the most relevant aspects of each of the topics.

In the context of these different topics, we then summarily introduce the papers accepted for the current ESANN special session. Some general conclusions are finally drawn.

2 Societal issues of ML

2.1 Ethics, Legislation and ML

Artificial Intelligence in general and ML in particular were not originally conceived with an eye on ethical issues. It might be argued that ML aims at inspiring its methods on certain aspects of biological intelligence. It might also be argued that one of the aspects of the outcome of human biological intelligence is precisely (un)ethical behaviour. It is true, though, that ethics do only come into play in social interaction, where different human intelligences communicate and interact with each other.

We find ourselves at a crossroads in the development of AI and ML in which artificially intelligent entities are beginning to become tightly interwoven in the social fabric, in the form, for example, of intelligent home assistants or surveillance systems. Unsurprisingly, their societal impact is coming to the fore of public discussion. This does not imply by any means that elements of such an impact where not already there to consider. It is decades since ML was deployed in financial risk assessment or medical diagnostic assistance, for instance, both with clear socio-ethical implications [1, 2], but certainly not in the pervasive manner we are witnessing today.

Fortunately, the AI and ML community is waking up to the dangers for technological advancement that may come from not appropriately addressing the many ethical challenges posed by the use of intelligent systems in society. An example of that may be found in the ongoing effort by the European Union High-Level Expert Group on Artificial Intelligence to create consensus around "Draft Ethics Guidelines for Trustworthy AI" [3]. This document assumes that "Trustworthy AI [...] should respect fundamental rights, applicable regulation and core principles and values, ensuring an ethical purpose". In parallel, the AI4People project has investigated in detail the ethical implications of AI and ML practice and subsumed them in four general principles, defined as beneficence, non-maleficence, autonomy, and justice [4].

Consensual ethical guidelines become, at some point in the development of modern societies, congealed in the form of legislation. And legislation is becoming aware of the social implications of the use of ML. This is clear in the implementation, from May 2018, of the European Union directive for General Data Protection Regulation (GDPR). It mandates a "right to explanation" of decisions made on citizens by "automated or artificially intelligent algorithmic systems" [5]. Such right underscores the role of the "data controller", and actor that becomes legally bound by GDPR to provide citizens with "meaningful information about the logic involved, as well as the significance and the envisaged consequences of such [automated decision making] for the data subject". The implications of this regulation on ML are quite straightforward if this technology is used as the basis to provide automated decision making.

The areas of ML application within society in which this becomes relevant are countless, but GDPR implications are of special importance in any process of interaction between the individual and private or public institutions. Obvious examples can be found in any form of automated decision making involving contracts and transactions between individuals and private corporations (typical examples include denial of credit from financial companies due to automated risk assessment, or legal accountability in autonomous driving [6]), but also in interactions between the individual and public institutions (for instance in the use of automated face recognition for public security by police forces [7]), or in the use of ML in medicine and healthcare, where the medical experts (from nurses to specialists) and the institutions they belong to are the data controllers and, therefore, ultimately legally responsible for the use of medical decision support systems based on ML. These pressing matters are beginning to encourage the definition of guidelines for GDPR-compliant ML development [8].

2.2 Privacy, Fairness and ML

The problem of learning from data while preserving the privacy of individual observations has a long history and spans over multiple disciplines [9, 10, 11]. One way to preserve privacy is to corrupt the learning procedure with noise without destroying the information that we want to extract. Differential Privacy (DP) is one of the most powerful tools in this context [12, 11]. DP addresses the apparently self-contradictory problem of keeping private the information about an individual observation while learning useful information about a population. In particular, a procedure is DP if and only if its output is almost independent from any of the individual observations. In other words, the probability of a certain output should not change significantly if one individual is present or not, where the probabilities are taken over the noise introduced by the procedure.

The concept of DP has allowed to reach a milestone result by connecting the field of privacy-preserving data analysis and the generalization capability of a learning algorithm. On the one hand it proved that a learning algorithm which shows DP properties also generalizes well [13, 14], while, on the other hand, if an algorithm is not DP, it allowed to state the conditions under which a hold-out set can be reused without risk of false discovery, through a DP procedure called *Thresholdout* [14, 15, 16].

In recent years, there has been much interest on the topic of algorithmic fairness in ML (see, for instance, [17, 18, 19] and references therein). The central question is how to enhance supervised learning algorithms with fairness requirements, namely ensuring that sensitive information (e.g., knowledge about the ethnic group of an individual) does not "unfairly" influence the outcome of a learning algorithm. For example, if the learning problem is to decide whether a person should be offered a loan based on her previous credit card scores, we would like to build a model which does not unfairly use additional socially sensitive information such as race or sex.

Several notions of fairness and associated learning methods have been introduced in ML over the past few years, including Demographic Parity [17], Equal Odds and Equal Opportunities [18], Disparate Treatment, Impact, and Mistreatment [19]. The underlying idea behind such notions is to balance the decisions of a classifier among the different sensitive groups and label sets. Work on algorithmic fairness can be divided into four families. Methods in the first family modify a pretrained classifier in order to increase its fairness properties while maintaining as much as possible the classification performance: [20, 21, 18, 22] are examples of these methods. Methods in the second family enforce fairness directly during the training phase, e.g. [23] and references therein. The third family of methods implements fairness by modifying the data representation and then employs standard ML methods: [24, 25, 26, 27, 28, 29] are examples of implementation of these methods. Finally, the last family of methods faces the problem of fairness via causality [30, 31]: in this works, authors start from the idea of counterfactual fairness, which states that a decision is fair toward an individual if it coincides with the one that would have been taken in a counterfactual world in which the sensitive attribute were different, and in the context of causal inference they propose a way to compensate the biases along the unfair pathways.

2.3 Safety, Security and ML

As mentioned in the introduction, ML and AI have become socially pervasive. From self-driving cars to smart devices, almost every consumer application now leverages such technologies to make sense of the vast amount of data collected from its users. Recent deep learning (DL) algorithms for computer vision have even surpassed human performances on some well-defined benchmark datasets, including ImageNet [32]. It has thus been extremely surprising to discover that such algorithms can be easily fooled by *adversarial examples*, that is, imperceptible, adversarial perturbations to images, text and audio that mislead these systems into perceiving things that are not there [33, 34]. After this phenomenon has been largely echoed by the press, undermining the safety and security properties of such algorithms in the corresponding application domains, a large number of stakeholders have shown interest in understanding the risks associated to the misuses of ML, to develop proper mitigation strategies and incorporate them in their products. Despite such large interest, this challenging problem is still far from being solved. It is also clear that this problem is particularly sensible in the cybersecurity domain, where AI and ML are used to detect malware on infected hosts, network intrusions and attacks directed to web services. These problems have indeed an intrinsic adversarial nature, as cybercriminal organizations have a clear interest and economic incentive in bypassing such systems to reach their goals. If AI and ML become the weakest link in the security chain, sooner or later attackers will exploit their vulnerabilities to violate the overall system security.

For these reasons, the research community has proactively started investigating vulnerabilities of AI and ML techniques to well-crafted attacks against them since 2004, even though the research in the area of adversarial ML only boomed in 2014, after the vulnerability of DL techniques to adversarial examples was highlighted in [33]. However, adversarial examples (also known as evasion attacks) are not the only threat that an attacker can envision against AI and ML. In particular, adversarial examples are modifications of test samples that aim to cause their misclassification (e.g., malware samples or spam emails misclassified as legitimate).

Besides this threat, attackers can also tamper with the training data, e.g., in applications where the system is retrained online, on data collected during operation. In this case, the attacker can inject "poisoning points" in the training set to cause subsequent misclassification of test samples, either to cause a denial of service to legitimate users (e.g., by enforcing misclassification of legitimate data) or facilitate subsequent intrusions (so-called backdoor attacks). Another line of attacks aim to violate privacy of the target system or of its users. By querying ML systems provided as online services, an attacker may be able to reverse-engineer the learning algorithm, potentially stealing the trained model or gaining confidential information about its parameters and structure (model-extraction attacks). Some attacks are even able to identify if a sample was part of the training data of the target system (membership-inference attacks), or they can be used to reconstruct some training samples to a very high precision (model-inversion attacks). We refer the reader to [34] for a detailed categorization of such attacks and the corresponding publications, as well as an historical survey of the field.

2.4 Interpretability, Explainability and ML

In previous sections, we have highlighted one of the legal bottlenecks hampering the application of ML to real problems in the social domain, namely the “right to explanation” granted to European citizens by the GDPR directive of recent application. Such requirement is in direct course of collision with the limitations of many ML methods in terms of interpretability and explainability. These twin issues have of late come to the forefront of ML research [35, 36], mostly due to the widespread development and application of DL methods in systems with societal impact. As they amplify shallow neural networks, it comes as no surprise that DL may become an extreme case of *black box model*.

Interpretability and explainability in ML are difficult to conceptualize problems, because they involve human cognition. This has not precluded the development of methodologies for the improvement of ML model interpretability, or even the development of formal frameworks to analyze the problem of ML interpretability, as in [6]. In this study, it is suggested that interpretability might sometimes need to be judged according to specific requirements of the application area. An example of an area of social impact in which interpretability is paramount is medicine and health. The requirements in this area have recently been discussed in [37]. Methods based on DL have found in medicine and health a perfect playing field, specially in the analysis of medical images and text in the form of Electronic Health Records [38]. In [37], it is shown that most, if not all, of the recent reviews covering applications of DL in medicine and health [38, 39, 40] identify interpretability and explainability as major challenges.

In recent years, researchers have tried to characterize techniques for interpretable ML in terms of few structural and functional aspects. The former aspect relates to whether a ML model can be considered intrinsically self-explanatory or transparent [41], incorporating interpretability into its design. For instance, not without debate [6], rule based systems, Bayesian networks and decision trees are considered to be intrinsically more explainable than neural networks. The understanding of ML models in terms of linear methods, including nomograms and other forms of General Additive Models, has also been proposed as a promising area of research [42, 43].

The second aspect relates to whether explainability is a global property providing insight into how the model works, or if it is a local property providing an explanation for how an individual prediction is made. The issue of designing ML models that are intrinsically explainable at a global level has often come down to imposing constraints, either numerical or architectural. Sparsity constraints, such as LASSO penalization [44], have long been used to augment interpretability (in particular, in linear models). More recently, Capsule Networks [45] provided an example of how imposing a modular design in the neural network architecture can strengthen its transparency. From a local perspective, the use of attention mechanisms are thought to reduce opacity in neural models by providing a means to highlight the contribution of specific parts of the input information or of the network itself on the final prediction. When transparency is external to the design of the model, we refer to *post-hoc* interpretability [41]. At a global level it is worth to mention the use of adversarial learning techniques to generate explanatory inputs of neurons' preferential stimuli, such as in activation masks [46]. At a local level, it is quite popular the use of perturbation techniques that alter either parts of the original input [47], or generate a set of neighboring point by small alterations [48],to assess the behaviour of the model under such modifications.

Finally, it is important to pay attention to explainability in planning problems, an approach seldom considered. Explainable planning aims to come up with plausible interpretation of planning models in terms of how the problem is solved, i.e., it should answer questions related to the reason of taking from a given particular action to a whole policy. One the first relevant works on this topic proposed a generation of explanations for human-robot interaction [49]. This has also been studied in the field of cognitive science either statistically [50], or by analyzing the links between knowledge and new ideas [51].

3 When Learning from Data is Not Enough. A Summary of the contributions of the ESANN special session

A total of six studies were accepted in the special session. In "Privacy Preserving Synthetic Health Data" [52], authors examine the feasibility of using synthetic medical data generated by Generative Adversarial Networks (GANs) in the classroom, to teach data science in health informatics. This paper fits the topics broached in Section 2.2. Authors present an end-to-end methodology to retain instructional utility, while preserving privacy to a level, which meets regulatory requirements: a GAN is trained by a certified medical-data security-aware agent, inside a secure environment, while the GAN is used outside of the secure environment by external users (instructors or researchers) to generate synthetic data. This second step facilitates data handling for external users, by avoiding de-identification, which may require special user training, be costly, and/or cause loss of data fidelity. Authors benchmark the proposed GAN versus various baseline methods using a novel set of metrics. At equal levels of privacy and utility, GANs provide small footprint models, meeting the desired specifications of authors application domain. Data, code, and details of a challenge organized for educational purposes are made available by the authors.

The second accepted work is "Fairness and Accountability of machine learning Models in Railway Market: are Applicable Railway Laws Up to Regulate Them?" [53]. This paper discusses whether the law is up to regulate ML model-based decision-making in the context of railways systems operation. Authors especially deal with the fairness and accountability of these models when exploited in the context of train traffic management and, therefore address the topics broached both in Sections 2.1 and 2.2. Railway sector-specific regulation, in their quality as network industry, hereby serves as a pilot. Authors show that, even where technological solutions are available, the law needs to keep up to support and accurately regulate the use of the technological solutions and authors identify stumble points in this regard.

The third work is "Deep RL for Autonomous Robots: Limitations and Safety Challenges" [54]. With the rise of deep reinforcement learning (Deep RL), there has also been a string of success stories related to problems of control of increasingly complex agents in physics simulators. This has lead to some optimism regarding uses in autonomous robots and vehicles. However, a recent study showed that popular benchmarks are actually solvable by a linear policy, raising the question of how well these represent real problems [55]. Authors therefore analyze a popular deep RL approach on toy examples from robot obstacle avoidance, showing that these converge very slowly, if at all, to safe policies. Convergence issues in the presence of uncertainty or local minima are identified as aspects needing more attention for such safety-critical control applications.

The fourth accepted work is "Detecting Adversarial Examples through Nonlinear Dimensionality Reduction" [56]. Deep neural networks are vulnerable to adversarial examples, i.e., carefully-perturbed inputs aimed to mislead classification. This work proposes a detection method based on combining non-linear dimensionality reduction and density estimation techniques. Authors empirical findings show that the proposed approach is able to effectively detect adversarial examples crafted by non-adaptive attackers, i.e., not specifically tuned to bypass the detection method. Given the reported promising results, an extension of the analysis to adaptive attackers is proposed as future work.

The fifth work is "Explaining classification systems using sparse dictionaries" [57] and relates to the topics broached in Section 2.4. It addresses the relevant topic of finding ways to explain the decisions of ML systems to end users, data officers, and other stakeholders. These explanations must be understandable to human beings. Much work in this field focuses on image classification, as the required explanations can rely on images, therefore making communication relatively easy, and may take into account the image as a whole. Here, authors propose to exploit the representational power of sparse dictionaries to determine image local properties that can be used as crucial ingredients of understandable explanations of classification decisions.

The last work accepted in our special session is "Dynamic fairness - Breaking vicious cycles in automatic decision making" [58] and again remits to the topics discussed in Section 2.2. Past research has shown that ML may reproduce and even exacerbate human bias due to biased training data or flawed model assumptions, leading to discriminatory actions. To counteract this, researchers have proposed multiple mathematical definitions of fairness according to which classifiers can be optimized. However, it has also been shown that the outcomes generated by some fairness notions may be unsatisfactory. In this contribution, authors add to this research by considering decision making processes over time, establishing a theoretic model in which even perfectly accurate classifiers adhering to almost all common fairness definitions lead to stable long-term inequalities due to vicious cycles. Only demographic parity, enforcing equal rates of positive decisions in all groups, avoids these effects and establishes instead a virtuous cycle leading to perfectly accurate and fair classification in the long term.

4 Conclusions

The increasingly pervasive use of AI and ML based systems in our everyday life, together with socially sensitive conflicts associated to such systems making the news (think of autonomous driving, automated weapons, or real-time AI-based surveillance, to name a few), are forcing the research community and governing institutions to address the societal impact of the use of these technologies.

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0. Available from http://www.i6doc.com/en/.

In this ESANN special session, we have gathered an interesting collection of papers that address diverse aspects of the societal impact of ML technologies and propose solutions for their potentially negative effects.

A topic as broad as this cannot be covered in full in just a brief introductory paper. In section 2, we have, at least, tried to provide an informative glimpse into the sheer variety of topics involved, all of them relevant to ML practitioners.

Private companies are swiftly positioning themselves to dominate the landscape of technological development related to AI and ML. Governments and public institutions, responsible for regulating the public domain and the societies they represent, are dragging behind these advances. The ML research community cannot ignore any longer the pressing societal issues related to the impact of these realignments, and efforts such as those of the European Union High-Level Expert Group on Artificial Intelligence's "Draft Ethics Guidelines for Trustworthy AI" [3] and the AI4People project [4] illustrate a path forward for this community to remain relevant in the shaping of the future of socially responsible ML.

Acknowledgements

This research was partially funded by Spanish MINECO TIN2016-79576-R research project and Italian MIUR grant RBSI14STDE.

References

[1] L. Delamaire, H. A. H. Abdou, and J. Pointon. Credit card fraud and detection techniques: a review. *Banks and Bank Systems*, 4(2):57–68, 2009.

[2] D. G. Altman. The scandal of poor medical research. *British Medical Journal*, 308:283, 1994.

[3] High-Level Expert Group on Artificial Intelligence. Draft ethics guidelines for trustworthy AI. *European Commission Directorate-General for Communication. Working Document*, 2018.

[4] L. Floridi et al. AI4People - an ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, 28(4):689–707, 2018.

[5] B. Goodman and Flaxman S. European Union regulations on algorithmic decision making and a "right to explanation". *AI Magazine*, 38(3), 2017.

[6] F. Doshi-Velez et al. Accountability of AI under the law: the role of explanation. *arXiv preprint arXiv:1711.01134*, 2017.

[7] B. Davies, M. Innes, and A. Dawson. An evaluation of South Wales police's use of automated facial recognition. *Universities' Police Science Institute Crime and Security Research Institute. Report*, 2018.

[8] M. Veale, R. Binns, and M. Van Kleek. Some HCI priorities for GDPR-compliant machine learning. the general data protection regulation: An opportunity for the CHI community? In *(CHI-GDPR 2018) Workshop at ACM CHI'18, Montreal, Canada*, 2018.

[9] V. S. Verykios et al. State-of-the-art in privacy preserving data mining. *ACM Sigmod Record*, 33(1):50–57, 2004.

[10] S. Greengard. Privacy matters. *Commun. ACM*, 51(9):17–18, 2008.

[11] C. Dwork and A. Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):1–277, 2014.

[12] C. Dwork. Differential privacy: A survey of results. In *International Conference on Theory and Applications of Models of Computation*, 2008.

[13] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. Roth. Preserving statistical validity in adaptive data analysis. In *Annual ACM Symposium on Theory of Computing*, 2015.

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0.
Available from http://www.i6doc.com/en/.

[14] L. Oneto, S. Ridella, and D. Anguita. Differential privacy and generalization: Sharper bounds with applications. *Pattern Recognition Letters*, 89:31–38, 2017.

[15] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. Roth. Generalization in adaptive data analysis and holdout reuse. In *Neural Information Processing Systems*, 2015.

[16] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. Roth. The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349(6248):636–638, 2015.

[17] T. Calders, F. Kamiran, and M. Pechenizkiy. Building classifiers with independency constraints. In *IEEE international conference on Data mining*, 2009.

[18] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. In *Neural Information Processing Systems*, 2016.

[19] M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *International Conference on World Wide Web*, 2017.

[20] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger. On fairness and calibration. In *Neural Information Processing Systems*, 2017.

[21] A. Beutel, J. Chen, Z. Zhao, and E. H. Chi. Data decisions and theoretical implications when adversarially learning fair representations. In *Conference on Fairness, Accountability, and Transparency in Machine Learning*, 2017.

[22] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. In *International Conference on Knowledge Discovery and Data Mining*, 2015.

[23] M. Donini, L. Oneto, S. Ben-David, J. Shawe-Taylor, and M. Pontil. Empirical risk minimization under fairness constraints. In *Neural Information Processing Systems*, 2018.

[24] J. Adebayo and L. Kagal. Iterative orthogonal feature projection for diagnosing bias in black-box models. In *Conference on Fairness, Accountability, and Transparency in Machine Learning*, 2016.

[25] F. Calmon, D. Wei, B. Vinzamuri, K. Natesan Ramamurthy, and K. R. Varshney. Optimized pre-processing for discrimination prevention. In *Neural Information Processing Systems*, 2017.

[26] F. Kamiran and T. Calders. Classifying without discriminating. In *International Conference on Computer, Control and Communication*, 2009.

[27] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning fair representations. In *International Conference on Machine Learning*, 2013.

[28] F. Kamiran and T. Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, 2012.

[29] F. Kamiran and T. Calders. Classification with no discrimination by preferential sampling. In *Machine Learning Conference*, 2010.

[30] S. Chiappa and T. P. S. Gillam. Path-specific counterfactual fairness. *arXiv preprint arXiv:1802.08139*, 2018.

[31] M. J. Kusner, J. Loftus, C. Russell, and R. Silva. Counterfactual fairness. In *Neural Information Processing Systems*, 2017.

[32] O. Russakovsky et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, Dec 2015.

[33] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.

[34] B. Biggio and F. Roli. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84:317–331, 2018.

[35] A. Vellido, J. D. Martín-Guerrero, and P. J. G. Lisboa. Making machine learning models interpretable. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 2012.

[36] A. Adadi and M. Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, 6:52138–52160, 2018.

[37] A. Vellido. The importance of interpretability and visualization in machine learning for applications in medicine and health care. In *Neural Computing and Applications*, pages doi.org/10.1007/s00521–019–04051–w, 2019.

[38] B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi. Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *IEEE J Biomed Health Inform*, 5(22):1589–1604, 2018.

[39] D. Ravì, C. Wong, F. Deligianni, M. Berthelot, J. Andreu-Pérez, B. Lo, and G. Z. Yang. Deep learning for health informatics. *IEEE J Biomed Health*, 1(21):4–21, 2017.

[40] T. Ching et al. Opportunities and obstacles for deep learning in biology and medicine. *J R Soc Interface*, 141(15):20170387, 2018.

[41] Z. C. Lipton. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*, 2016.

[42] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad. Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'15), pp.1721-1730*, 2015.

[43] V. Van Belle et al. Explaining support vector machines: a color based nomogram. *PLoS ONE*, 11:317–331, 2016.

[44] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[45] S. Sabour, N. Frosst, and G. E. Hinton. Dynamic routing between capsules. In *Neural information processing systems*, 2017.

[46] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[47] R. Fong and A. Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the International Conference on Computer Vision*, 2017.

[48] M. T. Ribeiro, S. Singh, and C. Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[49] S. Rosenthal, S. P. Selvaraj, and M. Veloso. Verbalization: Narration of autonomous robot experience. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16)*, 2016.

[50] J. Koster-Hale and R. Saxe. Theory of mind: a neural prediction problem. *Neuron*, 79(5):836–848, 2013.

[51] R. W. Magid, M. Sheskin, and L. E. Schulz. Imagination and thegeneration of new ideas. *Cognitive Development*, 34:99–110, 2018.

[52] A. Yale, S. Dash, R. Dutta, I. Guyon, A. Pavao, and K. B Bennett. Privacy preserving synthetic health data. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 2019.

[53] C. Ducuing, L. Oneto, and R. Canepa. Fairness and accountability of machine learning models in railway market: are applicable railway laws up to regulate them? In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 2019.

[54] O. Andersson and P. Doherty. Deep rl for autonomous robots: Limitations and safety challenges. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 2019.

[55] A. Rajeswaran, K. Lowrey, E. V. Todorov, and S. M. Kakade. Towards generalization and simplicity in continuous control. In *Neural Information Processing Systems*, 2017.

[56] F. Crecchi, B. Bacciu, and B. Biggio. Detecting adversarial examples through nonlinear dimensionality reduction. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 2019.

[57] A. Apicella, F. Isgro, R. Prevete, A. Sorrentino, and G. Tamburrini. Explaining classification systems using sparse dictionaries. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 2019.

[58] P. Paasen, A. Bunge, C. Hainke, L. Sindelar, and M. Vogelsang. Dynamic fairness - breaking vicious cycles in automatic decision making. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 2019.