INM708 Explainable Artificial Intelligence coursework

City, University of London

Kasper Groes Ludvigsen

1 Interpretability vs. Explainability

The terms *interpretability* and *explainability* are used interchangeably in the literature (Adadi & Berrada, 2018). According to Lewis (1986, p. 217), to explain an event means to provide information about its causal history, and an explanation can thus be considered an answer to a why-question (Miller, 2018). The focus on causes is echoed in Kim, Khanna & Koyejo (2016, p.) who define interpretability as the degree to which a human can consistently predict the model's result. Several researchers consider explainability and interpretability to be two sides of the same coin (Miller, 2018; Lipton, 2017; Molnar, 2021), and explainability is sometimes construed as post-hoc interpretability (Miller, 2018; Lipton, 2017). Techniques and model properties that either enable or comprise interpretability generally fall into two categories: 1) transparency, i.e. understanding how the model works, 2) Post-hoc explanations, i.e. what else can the model tell us. Transparency can be subdivided into *simulatability*, *decomposability* and *algorithmic transparency* (Lipton, 2017).

While some use *interpretability* and *explainability* interchangeably, others researchers have strong views on the difference between interpretability and explainability and which is desirable. Rudin (2019) describes an explanation as “a separate model that is supposed to replicate most of the behavior of a black box” (Rudin, 2019, p. 2), where the term “explanation” refers to an understanding of how a model works. She is a strong advocate for using machine learning models that are inherently interpretable rather than trying to explain black box models. The main problem with explainable AI methods, is that any explanation method for black box models is at risk of being an inaccurate representation of the original model in parts of the feature space (Rudin, 2019). This limits the trust in the explanation, and hence the trust in the black box model. This reasoning is echoed in Mittelstadt, Russel and Wachter (2018). On the other hand, inherently interpretable models provide their own explanations that are trustworthy (Rudin, 2019). Rudin’s distinction between interpretability and explanation mimics Molnar’s (2021) taxonomy of machine learning interpretability, which distinguishes between intrinsic and post hoc interpretability. In Molnar’s terminology, intrinsic interpretability is achieved for instance via short decision trees or sparse linear models, and post hoc interpretability refers to the application of interpretation methods after model training. Other ways to way to distinguish between interpretation methods are (Molnar, 2021): 1) to look at the results they produce , 2) to consider whether the method is applicable to all types of models or only some models (model-specific vs. model-agnostic), 3) to consider the scope of the interpretation. Is it local, i.e. it explains a single prediction, or is it global, i.e. it explains the entire model, or somewhere in between?

Although there is debate about the exact terminology, it seems researchers mostly agree that models can be explained (or “interpreted”) in two fundamentally different ways: Either because the models are inherently understandable, for instance due to their simplicity, or because a method can be applied to explain or interpret the results post hoc. Lipton (2017), Molnar (2021), Rudin (2019) and Adadi & Berrada (2018) all consider interpretable models to include rule based systems, decision trees and logistic and linear regression models and their extensions (e.g. GAM) which are constrained in model form, e.g. by obeying monotonicity or additivity. It should be noted, though, that very large decision trees and linear models with many features can become uninterpretable (Lipton, 2017).

2. Why XAI?

An often cited motivation for explainability is trust, i.e. that in order for us to trust a model, we must be able to receive an explanation for its decision. But why don't we just trust a model based on a performance metric such as accuracy? Explanations may come into play more often when the model does not perform as intended (Lipton, 2017; Miller, 2019). Adadi & Berrada (2018) found that at least four main motivations for explainable AI exist:

1. Explain to justify, which refers to the ability to explain an outcome with the aim of providing reasons or justifications for the outcome.
2. Explain to control, which refers to the ability to control the behavior or the AI system by knowing why it behaves the way it does. Understanding system behavior arguably provides more visibility over unknown flaws in the system design, enabling quicker correction of errors.
3. Explain to improve - if we understand why the system behaves in a certain manner, we can more easily improve it.
4. Explain to discover. In some cases, it is desirable to understand the behavior of a system so that we may learn from it. For instance, we might learn from gaining insights into the strategies of game playing systems.

These motivations are not exhaustive. For instance, they do not consider “explain to comply”. As a consultant on a public sector software development program, I experienced how this motivation led to the use of Generalized Additive Models rather than more advanced methods in order to comply with Danish regulation requiring explainability of predictions. While this legislation itself is probably driven by the motivation “explain to justify”, the implementation of the legislation is driven by “explain to comply”. Although there is debate among legal scholars about the extent to which articles 13-15 of the General Data Protection Regulation give individuals a “right to explanation” (Selbst & Powell, 2017; Wachter, Mittelstadt & Floridi, 2016), the law may encourage organizations to design explainable systems with the aim of achieving compliance with regulation.

Several researchers (e.g. Adadi & Berrada, 2018; Guidotti et al, 2017) point out that explainability may not always be necessary, which naturally leads one to ask: Which AI systems require explainability then?

While universally applicable rules seem hard to define, it strikes me as reasonable to consider the potential consequences of the decisions made by the system. Human harm caused by autonomous vehicles should require serious scrutiny; recommending the wrong product to a website visitor should in most cases not. The nature of the environment in which the system acts should also be considered. In environments where data and model drift are likely, you may want explainable models, in which case the motivation would be to explain to control and to improve.

Many times, the need for explainability is articulated through examples of machine learning models that have gone wrong (e.g. Larson, Surya Mattu, Lauren Kirchner and Julia Angwin, 2016; Rudin, Wang & Coker, 2019; Stochastic Programming Society, 2020). When such examples are used to motivate explainability, it conveys the message that it is okay to use machine learning for any type of problem as long as we use models that can be explained or interpreted. The debate about what distinguishes interpretable AI from explainable AI and why either is needed, removes attention from the debate of whether machine learning should even be used for high stakes decisions in the first place. So in addition to asking “Why XAI?”, one ought to also ask “Why AI?”.

3. Interpretable models vs. Model-agnostic methods: characteristics, pros and cons

Interpretable models include linear and logistic regression and their extensions such as GLM, and GAM, in addition to decision trees and decision rules and naive bayes and kNN classifiers. With the exception of kNN, these models are all interpretable on the modular level, which means that we can understand at least how some parts of the model affect predictions if not all parts (Molnar, 2021).

Model-agnostic methods separate the explanation from the machine learning model (Molnar, 2021). The strength of model-agnostic methods is that they can be applied to all types of machine learning models hence enabling comparisons between different model types. They also provide explanation flexibility, meaning that they are not limited to a certain form of explanation. Model-agnostic methods also give flexibility in terms of how to represent the model it explains (Ribeiro, Singh & Guestrin, 2016) - the representation must not be the same as that of the machine learning model being explained, which is an advantage for instance when the representation of the machine learning model input is abstract word embeddings (Molnar, 2021). Model-agnostic methods also have downsides. LIME (Ribeiro, Singh and Guestrin, 2016) is an example of a model-agnostic method that suffers from only being able to provide local explanations, and only linear models are used to approximate local behavior, so their explanations might be wrong if the actual model behavior is non-linear (Hulstaert, 2018).

It is often asserted that interpretable models come at the cost of reduced predictive performance (e.g. Molnar, 2021). However, this is not always the case. In particular, when data is structured and when features are naturally meaningful, there is often no significant difference in performance between more complex classifiers, e.g. artificial neural networks, and simpler classifiers, e.g. logistic regression (Rudin, 2019, appendix B). However, interpretable models are not applicable to all types of problems, or at least they suffer from a significant disadvantage compared to deep learning methods. Such problems include computer vision tasks, where convolutional neural networks have enabled advances that are out of the reach of simpler models (Simonyan & Zisserman, 2015), as well as natural language processing tasks where recurrent neural networks specialized for sequential data have long been state of the art (Hochreiter & Schmidhuber, 1997).

4. Study of ethical/legal responsibility, AI rights and liability

4.1. Case study introduction

In Denmark, all public sector employees are required by law to notify authorities if they suspect that a child is living under conditions that could jeopardise the health and wellbeing of the child (Borger.dk, n.d.). In 2019, social workers in Denmark received 137,986 such notifications, roughly 380 per day. This number is significant in a country with just shy of six million inhabitants, and it increases each year. Upon receiving a notification, a social worker opens a case with the aim of assessing if the child is in risk of abuse or other harm and should be removed from its family. Due to the sheer amount of notifications, the public sector has looked to technology for ways to make the workload more manageable. In 2018, social workers started to use a machine learning model that was intended as a tool to support their decisions about whether or not to remove children from their parents. It was not intended to decide cases autonomously. The machine learning model assigned a risk score to children about whom notifications had been made (Kulager, 2021).

200 notifications about alarming behavior among children were assessed by the machine learning model before it was taken out of production after it was discovered that the model was flawed. It turned out the algorithm was highly age biased, such that two children for whom all input features except age were the same would be scored differently, the older child being assigned a higher risk score. The probable explanation for this bias can be found in the training data which showed that older children were removed more frequently. This is likely because abused children's experiences become more visible with age because they gain the ability to tell others about it or because they start to exhibit behavior, e.g. criminal or violent, that indicates failure to thrive. This bias could mean that a child is wrongfully removed from its family, or that a child who is abused is not removed. In addition, some aspects of a person's life situation can be found ill-suited for formal representation by social workers, which can cause case workers to not record them in a computer system (Petersen, Christensen, Harper & Hildebrandt, 2021). As a result, important information may not be accessible to the AI, which interferes with its ability to make ethical decisions.

4.2. Analysis of ethical and legal agency

Legal agency

A legal agent is a legal subject "which can *control and change* its behavior and *understand* the legal consequences of its actions or omissions" (Mondragon, 2021, p. 16). The machine learning model in question does not live up to this requirement, because it can only ever do what it was designed to do, i.e. assign risk scores. It will not all of a sudden be able to cook hamburgers and estimate property values. Only if hamburger cooking and property value estimation models were embedded in the same “agent”, this system in its entirety could be considered an agent able to *control and change* its behavior. The question about whether a piece of computer software can *understand* the legal consequences of its actions or omissions comes down to how one defines *understanding*. AI is already used to make predictions about legal outcomes (Lu, 2019), and so it seems perfectly plausible that an AI system could be trained to anticipate the legal consequences of all actions. It seems there is broad agreement in academia that AI generally does not understand cause and effect - machine learning based AI merely identifies correlations (Marcus & David, 2019; Pearl & Mackenzie, 2018). Answering whether that amounts to *understanding* is outside of the scope of this coursework. I will instead say that whether an AI entity qualifies for legal agency is a matter of degrees, although currently only humans are assigned legal agency (Lior, n.d.).

Moral agency

Machine and computer actions can have a moral quality because they can produce actions that affect the moral rights and obligations of a human being (Stahl, 2004). The question then is: Can computers be ascribed moral responsibility such that they can be perceived as moral agents? Traditionally, the answer provided by moral philosophers has been “no”, because computers do not fulfil a number of conditions such as cognitive and emotional abilities, knowledge of the results of actions, and the power to change events. Another argument against computer moral agency ascription is that it can be used as an excuse for people to evade their responsibility (Stahl, 2004).

However, arguments in favour of ascribing some degree of moral agency to computers do exist. Computers often play a role in social interactions, for instance the social interactions that characterise the work of social workers, and such interactions have a moral nature (Stahl, 2004). In addition, computers can be made part of explicit moral decision making (Gotterbarn, 2002), which is clearly the case when social workers take into account the output of an algorithm when making decisions about removing a child from their family. The machine model in case might therefore be viewed as a moral agent in an “operational” way in the sense that it is an extension of its designers’ values that is more autonomous than other tools such as hammers, and has some sensitivity to ethical matters built into it. The niche between genuine moral agency and operational morality is called “functional morality” (Wallach & Allen, 2009).

Instead of discussing whether computers can be genuine moral agents (a discussion that could go on for eternity as the answer depends on the metaphysical convictions of the participants of the discussion), the question can be sought answered empirically through the Moral Turing Test (Allen, 2000). Such a test might be carried out with MedEthEx (Anderson, Anderson & Armen, 2006) which can answer moral questions.

4.3. Analysis of an AI rights

When using AI to decide whether to remove a child from its family, the decision by the AI entity could affect the rights to family life enjoyed by the forcefully removed child and its parents. If the AI assigns a risk score that is too high, the AI decision could be violating a fundamental human right. The right to family life is described in article 16 of the UN's human rights convention from 1948 and is legally binding. The European Convention on Human Rights from 1950 also asserts the right to family life. According to its article 8, public authorities can only interfere with family life if there is legal basis for doing so and if it is necessary in a democratic society. Interference with the right to family life could for instance be justified if it is necessary to protect the rights and freedom of other individuals. If a child is removed, it is in some cases because it has been abused by its parents. In such a case, the child's claim-right against abuse has been violated, and the parents have violated their duty not to abuse the child. In other cases, a child can be removed if its parents suffer from physical or mental handicap (e.g. see Kutzner v. Germany), and the child's claim-right to sufficient care is violated. No duty exists to not use machine learning in making decisions about limiting the rights to family life - but should AI be used for such decisions?

It has been argued that AI often stands in opposition to human welfare because it has contributed to the perpetuation of historical and social bias and injustice, to invasion of privacy and to exploitation of human labour. Using AI to make important decisions that affect human rights bears the risk of furthering these detrimental effects of AI, especially on socioeconomically disadvantaged groups (Birhane & van Dijk, 2020). One could in particular be worried that the algorithm in this case is not sufficiently sophisticated to take into account the plethora of factors that determine the ability of parents to care for their child. Previous cases might show that parents with physical or mental functional impairments have had their rights to family life affected, but how can the algorithm assess the ability of a particular parent to care for their child? My guess is that it cannot. When the model is used to assign a risk score, it is unable to understand whether functional impairment is the cause of a child's failure to thrive, because the model can only understand correlations. In addition, it does not seem far-fetched to imagine that factors such as race and income could play a disproportionate role in assigning risk scores given how such factors have previously been shown to lead to unjust decisions about human futures (Heaven, 2020; Larson, Mattu, Kirchner & Angwin, 2016).

I will not dwell on the idea of ascribing rights to AI, because I largely subscribe to Birhane and van Dijk's (2020, p. 2) stance on the question, summarized here: "Robot rights signal something more serious about AI technology, namely, that, grounded in their materialist techno-optimism, scientists and technologists are so preoccupied with the possible future of an imaginary machine, that they forget the very real, negative impact their intermediary creatures - the actual AI systems we have today - have on actual human beings."

4.4. Analysis of an AI liability

Let us assume that a social worker did not use their own judgement to decide in a case, but instead merely used the output of the algorithm. Later, it was found that the algorithm output had led to the wrongful removal of a child from its family. Who would be liable in such a case?

In order for any party to be liable, factual and legal causation must be established. Let's first consider the social worker. Factual causation can be established if the social worker's inaction was a necessary condition for the consequence. Although tough, the question would likely be settled by invoking the bonus pater familias concept and assessing if a reasonable social worker would have acted similarly. It is also necessary to establish that the omission of the social worker was free and deliberate, and that the case worker either knew or ought to have known the consequences of the inaction. Given the intended usage of the algorithm, it would likely be determined that a case worker ought to have known not to let the algorithm decide on its own. Finally, it must be established that there has been no *novus actus interveniens* (intervening act), which is an independent event - occurring after the case worker's negligence - which caused or contributed to the consequence (Hogan Lovells Publications, 2017). In conclusion, it seems possible to establish factual and legal causation. Another important aspect is the burden of proof. It will likely be difficult under most circumstances to prove that the case worker did not use their own opinion to assign a risk score.

Let's now consider the algorithm manufacturer. One legal analogy that might be applied is that of property, more specifically products. In order for the manufacturer to be liable, the injured party must prove at least one of three defect categories: A manufacturing defect, failure to provide adequate instructions or warnings, or a design defect (Lior, n.d., section 3-A-1).
A manufacturing defect is when the product "departs from its intended design even though all possible care was exercised in the preparation and marketing of the product." (Lior, n.d., In the case at hand, "all possible care" was *not* exercised in the manufacturing of the algorithm as will be discussed below. Failure to provide adequate instructions or warnings would likely not be provable. Since the algorithm was designed as a decision support tool, it must be assumed that reasonable instructions and warnings were given not to let case decisions rest solely on the algorithm output.
A design defect is when "the foreseeable risks of harm posed by the product could have been reduced by the adoption of a reasonable alternative design" (Lior, n.d., section 3-A-1) . In the case at hand, the damage is caused by age bias, which was identified by an undergrad student (Kulager, 2021). The fact that a student was able to identify the bias suggests that the university researchers who developed the algorithm should have been so too, but failed to do so due to negligence, and that alternative designs such as reducing the weight of age could have removed the bias.
Again, causation must be established. The decision of the algorithm seems to be a necessary condition for the removal of the child unless it can be proved that the social worker would have made the same decision which is doubtful cf. the arguments above. Whether the designers of the algorithm ought to have known that it would make wrong decisions is less clear as it is widely accepted that software is error prone (Goertzel, 2016; House of Lords, 2006).

Most writings on the product analogy conclude that it may not be sufficient to handle the unique features of AI systems such as the black box problem (Lior, n.d.). However, not all AI methods are black box, and for those that are, explainable methods exist cf. the discussion in question 3. It therefore may not be unrealistic to expect the product analogy to be applied in AI liability cases.

4.5. Conclusions and general discussion

Widespread use of AI is still a relatively new phenomenon, and many questions about ethics and legality regarding AI are still unanswered. The ethical questions in particular may remain unanswered for quite some time, if not forever. Or perhaps ethical discussions about AI will simply go out of fashion the same way we no longer discuss the dangers of having cars in our streets. As has been discussed above, no AI entity (or any other non-human "entity" for that matter) has yet been ascribed legal agency. In general, computer programs are considered mere tools in legal academia, although some debate exists about whether AI should be given the same status as corporations. In the EU, this debate has revolved around ascribing personhood to AI which essentially is analogous to corporate personhood (Hern, 2017). Addressing an "urgent need to create a robust legal framework" (Hern, 2017, paragraph 3) that would ensure that robots are and will remain in the service of humans, the EU parliament's legal affairs committee published a report in 2017 outlining a possible framework for granting legal personhood to robots (Committee on Legal Affairs, 2017). One might suspect that those and other similar endeavours by elected politicians were motivated by populistic aspirations fueled by mainstream media attention diverted to the topic of AI and its supposed threats around the time (Wallace, 2014; Adams, 2016; Cadwalladr, 2014). Aside from being grounded in fantasies about what AI might one day become, the debate about AI rights and personhood carries the risk of diverting attention from real issues pertaining to adverse effects of AI adoption (Birhane & van Dijk, 2020). In addition, the debate about what distinguishes interpretable AI from explainable AI and why either is needed, removes attention from the debate of whether machine learning should even be used for high stakes decisions in the first place. In cases where the use of AI does result in harm, existing regulatory frameworks suffice as they are built on principles that supersede the specific technological artifact they govern. Although the question about liability is not straightforward, it seems factual and legal causation could be established for the social worker and perhaps also for the manufacturer of the AI. While the adoption of AI poses many serious threats (cf. previous sections), none of them endangers the human species as a whole, and it is my firm belief that an AI-induced judgement day will continue to be but the plot of a Hollywood film.

Optional task - applying SHAP to CNN predictions on MNIST data

For this task, I opted for SHAP over LIME because the former has better support for PyTorch which is what we have been using in most other modules.

The SHAP method uses SHAP values as a measure of feature importance, which is Shapley values of a conditional expectation function of the original model (Lundberg & Lee, 2017). Shapley values, a concept from coalitional game theory, is a way to distribute “payout” between the features that are used in a prediction (Molnar, 2021). The methods used to estimate the SHAP values are better aligned with human intuition and are better at discriminating model output classes than other methods (Lundberg & Lee, 2017). The shapley value estimated by SHAP is the contribution of a feature value to the difference between the actual prediction and the mean prediction.

I used the SHAP method to explain the predictions made by a convolutional neural network trained to classify handwritten digits in the MNIST dataset. I originally implemented the model in INM702. All the code associated with this task (model implementation and SHAP interpretation) can be found in the notebook `shap.ipynb`. The output of applying the SHAP method to 10 predictions is seen below in Figure 1.

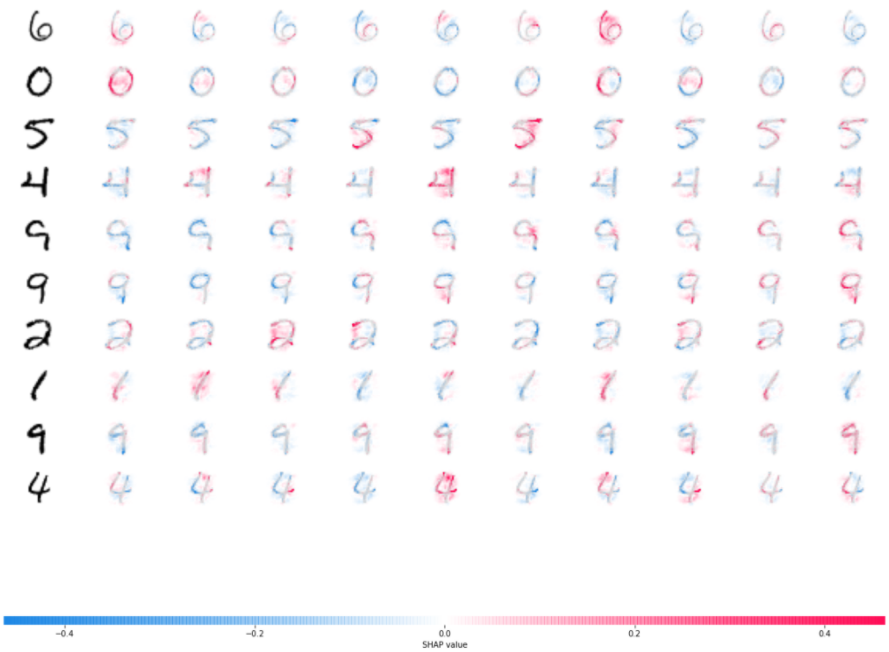

Figure 1 - SHAP output

The first column in Figure 1 shows the 10 predictions on which I applied SHAP. The other columns show the explanations for the predictions. Red pixels indicate areas of the image that increased the SHAP output, while blue pixels decrease the SHAP model output. The red pixels in the middle of the 0 indicate that a blank area inside a circle is predictive of the number 0. In the fourth and last row, the red pixels in the 6th column show that the lack of a vertical bar makes the number 4 stand out. In the row with ones, it also seems that the lack of a vertical bar on top and in the middle of the horizontal one makes the ones stand out from sevens.

The SHAP method has a number of benefits as described by Molnar (2021). It is theoretically solid, which cannot be said about LIME which assumes that the machine learning model to be explained behaves linearly locally without any theoretical justification for this assumption. Shapley values, on which SHAP are based, are characterized by being *efficient*, meaning that they distribute “payouts” fairly. In the context of XAI, this guarantees that the contribution to a prediction from each feature is distributed fairly. Such guarantees are not made by LIME. SHAP also allows for contrastive explanations, and such contrasts can be derived from a subset of the data or even a single datapoint. This is not possible with LIME.

A disadvantage of SHAP is that it is computationally expensive. There are 2^k possible coalitions, but the exponential increase in possible coalitions is handled by sampling coalitions, which reduces compute time, but increases the variance of the Shapley value (Molnar, 2021). It has also been shown that the SHAP method is susceptible to adversarial attacks (Slack, Hilgard, Jia, Singh, Lakkaraju, 2020).

References

Adadi, A. and Berrada, M. (2018) ‘Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)’, *IEEE Access*, PP, pp. 1–1. doi: 10.1109/ACCESS.2018.2870052.

Adams, T. (2016) *Nick Bostrom: ‘We are like small children playing with a bomb’*, *the Guardian*. Available at: http://www.theguardian.com/technology/2016/jun/12/nick-bostrom-artificial-intelligence-machine (Accessed: 20 May 2021).

Allen, C. et al. (2000), ‘Prolegomena to Any Future Artificial Moral Agent’, Journal of Experimental and Theoretical Artificial Intelligence 12, pp. 251–261.

Bendixen, C. *et al.* (2014) *Ret til at være forældre*. Kbh.: Institut for Menneskerettigheder.

Borger.dk (no date) *Børn i mistrivsel*. Available at: https://www.borger.dk/familie-og-boern/Udsatte-boern-og-unge/Boern-i-mistrivsel (Accessed: 24 May 2021).

Cadwalladr, C. (2014) *Are the robots about to rise? Google’s new director of engineering thinks so…*, *the Guardian*. Available at: http://www.theguardian.com/technology/2014/feb/22/robots-google-ray-kurzweil-terminator-singularity-artificial-intelligence (Accessed: 20 May 2021).

Case of Kutzner v. Germany (Application no. 46544/99), Judgment 26 February 2002 (Final 10/07/2002).

Committee on Legal Affairs (no date) *REPORT with recommendations to the Commission on Civil Law Rules on Robotics*. Available at: https://www.europarl.europa.eu/doceo/document/A-8-2017-0005_EN.html (Accessed: 20 May 2021).

Definition of EXPLAINING (no date). Available at: https://www.merriam-webster.com/dictionary/explaining (Accessed: 14 May 2021).

Definition of INTERPRETING (no date). Available at: https://www.merriam-webster.com/dictionary/interpreting (Accessed: 14 May 2021).

Doran, D., Schulz, S. and Besold, T. R. (2017) 'What Does Explainable AI Really Mean? A New Conceptualization of Perspectives', *arXiv:1710.00794 [cs]*. Available at: http://arxiv.org/abs/1710.00794 (Accessed: 3 February 2021).

Ethical Implications, Proceedings of the sixth ETHICOMP Conference, 13–15 November 2002, Lisbon, Portugal, Lisbon: Universidade Lusiada, pp. 125–141

Goertzel, K. (2016) 'Legal liability for bad software', *CrossTalk*, 29, pp. 23–28.

Gotterbarn, D. (2002), 'The Ethical Computer Grows Up: Automating Ethical Decisions', in I. Alvarez et al., eds., The Transformation of Organisations in the Information Age: Social and

Grint, K. and Woolgar, S. (1997), The Machine at Work: Technology, Work, and Organization, Cambridge: Blackwell.

Guestrin, M. T. R., Sameer Singh, Carlos (2016) *Local Interpretable Model-Agnostic Explanations (LIME): An Introduction*, *O'Reilly Media*. Available at: https://www.oreilly.com/content/introduction-to-local-interpretable-model-agnostic-explanations-lime/ (Accessed: 20 May 2021).

Guidotti, R. *et al.* (2018) 'A Survey Of Methods For Explaining Black Box Models', *arXiv:1802.01933 [cs]*. Available at: http://arxiv.org/abs/1802.01933 (Accessed: 26 January 2021).

Hern, A. (2017) *Give robots 'personhood' status, EU committee argues*, *the Guardian*. Available at: http://www.theguardian.com/technology/2017/jan/12/give-robots-personhood-status-eu-committee-argues (Accessed: 20 May 2021)

House of Lords (U.K.) Science and Technology Committee. (2007, July 24). 5th Report of Session 2006–07, Personal Internet Security, Volume I: Report, Chapter 4: “Appliances and applications.” HL Paper 165–I. Retrieved from http://www.publications.parliament.uk/pa/ld200607/ldselect/ldsctech/165/165i.pdf

Johnson, D.G. (2001), Computer Ethics, 3rd edition, Upper Saddle River, NJ: Prentice Hall.

Jordan, N. (1963), ‘Allocation of Functions Between Man and Machines in Automated Systems’, Journal of Applied Psychology 47(3), pp. 161–165.

Kulager, F. (2021) *Kan algoritmer se ind i et barns fremtid? I Hjørring og Silkeborg eksperimenterede man på udsatte børn* (2021) *Zetland*. Available at: https://www.zetland.dk/historie/s8YxAamr-aOZj67pz-e30df (Accessed: 20 May 2021).

Krizhevsky, A., Sutskever, I. and Hinton, G. E. (2017) ‘ImageNet classification with deep convolutional neural networks’, *Communications of the ACM*, 60(6), pp. 84–90. doi: 10.1145/3065386.

Lenk, H. (1994), Macht und Machbarkeit der Technik, Stuttgart: Philipp Reclam jun

Lent, M., Fisher, W. and Mancuso, M. (2004) *An Explainable Artificial Intelligence System for Small-unit Tactical Behavior.*, p. 907.

Lewis, D. K. (1986) ‘Causal Explanation’, in Lewis, D. (ed.) *Philosophical Papers Vol. Ii*. Oxford University Press, pp. 214–240.

Lior, A. (no date) ‘AI Entities as AI Agents: Artificial Intelligence Liability and the AI Respondeat Superior Analogy’, *Mitchell Hamline Law Review*. Available at: https://mhlawreview.org/law_review_article/ai-entities-as-ai-agents-artificial-intelligence-liability-and-the-ai-respondeat-superior-analogy/ (Accessed: 19 May 2021).

Lipton, Z. C. (2017) ‘The Mythos of Model Interpretability’, *arXiv:1606.03490 [cs, stat]*. Available at: http://arxiv.org/abs/1606.03490 (Accessed: 20 January 2021).

Lou, Y., Caruana, R. and Gehrke, J. (2012) ‘Intelligible models for classification and regression’, in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD ’12. the 18th ACM SIGKDD international conference*, Beijing, China: ACM Press, p. 150. doi: 10.1145/2339530.2339556.

Lu, D. (no date) *AI learns to predict the outcomes of human rights court cases*, *New Scientist*. Available at:

https://www.newscientist.com/article/2212953-ai-learns-to-predict-the-outcomes-of-human-rights-court-cases/ (Accessed: 19 May 2021).

Lundberg, S. M. and Lee, S.-I. (no date) ‘A Unified Approach to Interpreting Model Predictions’, p. 10.

Mattu, J. L., Julia Angwin,Lauren Kirchner,Surya (no date a) *How We Analyzed the COMPAS Recidivism Algorithm*, *ProPublica*. Available at: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm?token=SV45W9VHgigYbUE-m7o9xnvExqobnjcg (Accessed: 27 January 2021).

Mattu, J. L., Julia Angwin,Lauren Kirchner,Surya (no date b) *How We Analyzed the COMPAS Recidivism Algorithm*, *ProPublica*. Available at: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm (Accessed: 20 May 2021).

Miller, T. (2018) ‘Explanation in Artificial Intelligence: Insights from the Social Sciences’, *arXiv:1706.07269 [cs]*. Available at: http://arxiv.org/abs/1706.07269 (Accessed: 13 May 2021).

Mittelstadt, B., Russell, C. and Wachter, S. (2019) ‘Explaining Explanations in AI’, *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 279–288. doi: 10.1145/3287560.3287574.

Molnar, C. (2021) *Interpretable Machine Learning*. Available at: https://christophm.github.io/interpretable-ml-book/agnostic.html (Accessed: 10 February 2021).

Novus actus interveniens (2017) *www.hoganlovells.com*. Available at: http://www.hoganlovells.com/en/publications/novus-actus-interveniens (Accessed: 19 May 2021).

Mondragon, E. (2021), Lecture 3: Ethical and legal agency, Explainable Artificial Intelligence, City, University of London

Pagallo, U. (2013) *The Laws of Robots*. Dordrecht: Springer Netherlands. doi: 10.1007/978-94-007-6564-1.

Pearl, J. and Mackenzie, D. (2018), ‘The Book of Why: The new science of cause and effect’, New York: Basic Books, Published May 15, 2018

Petersen, A. *et al.* (2021) *‘We Would Never Write That Down’: Classifications of Unemployed and Data Challenges for AI*, *Proceedings of the ACM on Human-Computer Interaction*. doi: 10.1145/3449176.

Predictive policing algorithms are racist. They need to be dismantled. (no date) *MIT Technology Review*. Available at: https://www.technologyreview.com/2020/07/17/1005396/predictive-policing-algorithms-racist-dismantled-machine-learning-bias-criminal-justice/ (Accessed: 20 May 2021).

Razavian, N. *et al.* (2015) ‘Population-Level Prediction of Type 2 Diabetes From Claims Data and Analysis of Risk Factors’, *Big Data*, 3(4), pp. 277–287. doi: 10.1089/big.2015.0020.

Refugees, U. N. H. C. for (no date) *Refworld | Kutzner v. Germany*, *Refworld*. Available at: /cases,ECHR,58c166454.html (Accessed: 20 May 2021).

Ribeiro, M. T., Singh, S. and Guestrin, C. (2016a) ‘Model-Agnostic Interpretability of Machine Learning’, *arXiv:1606.05386 [cs, stat]*. Available at: http://arxiv.org/abs/1606.05386 (Accessed: 10 February 2021).

Ribeiro, M. T., Singh, S. and Guestrin, C. (2016b) ‘“Why Should I Trust You?”: Explaining the Predictions of Any Classifier’, *arXiv:1602.04938 [cs, stat]*. Available at: http://arxiv.org/abs/1602.04938 (Accessed: 3 February 2021).

Rudin, C. (2019) ‘Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead’, *arXiv:1811.10154 [cs, stat]*. Available at: http://arxiv.org/abs/1811.10154 (Accessed: 27 January 2021).

Rudin, C., Wang, C. and Coker, B. (2019) ‘The age of secrecy and unfairness in recidivism prediction’, *arXiv:1811.00731 [cs, stat]*. Available at: http://arxiv.org/abs/1811.00731 (Accessed: 27 January 2021).

Selbst, A. D. and Powles, J. (2017) ‘Meaningful information and the right to explanation’, *International Data Privacy Law*, 7(4), pp. 233–242. doi: 10.1093/idpl/ipx022.

Simonyan, K. and Zisserman, A. (2015) ‘Very Deep Convolutional Networks for Large-Scale Image Recognition’, *arXiv:1409.1556 [cs]*. Available at: http://arxiv.org/abs/1409.1556 (Accessed: 14 May 2021).

Dylan Slack*, Sophie Hilgard*, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. 2020. Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods. In Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society (AIES ’20), February 7–8, 2020, New York, NY, USA. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3375627.3375830

Stahl, B. C. (no date) ‘Information, Ethics, and Computers: The Problem of Autonomous Moral Agents’, p. 17.

Stanford University (2021) ‘Rights’, in *The Stanford Encyclopedia of Philosophy*. Spring 2021. Metaphysics Research Lab, Stanford University. Available at: https://plato.stanford.edu/archives/spr2021/entries/rights/ (Accessed: 20 May 2021).

Stochastic Programming Society (2020) *Interpretability vs. Explainability in Machine Learning*. Available at: https://www.youtube.com/watch?v=zsRKPxgHURQ&ab_channel=StochasticProgrammingSociety (Accessed: 20 May 2021).

Tibshirani, R. (1996) ‘Regression Shrinkage and Selection via the Lasso’, *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1), pp. 267–288.

Villasenor, J. (2019) ‘Products liability law as a way to address AI harms’, *Brookings*, 31 October. Available at: https://www.brookings.edu/research/products-liability-law-as-a-way-to-address-ai-harms/ (Accessed: 20 May 2021).

Vladeck, D. C. (no date) ‘Machines Without Principals: Liability Rules and Artificial Intelligence’, *WASHINGTON LAW REVIEW*, 89, p. 35.

Wachter, S., Mittelstadt, B. and Floridi, L. (2016) ‘Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation’. doi: 10.1093/idpl/ipx005.

Wallace, G. (2014) *Elon Musk warns against unleashing artificial intelligence ‘demon’*, *CNNMoney*. Available at: https://money.cnn.com/2014/10/26/technology/elon-musk-artificial-intelligence-demon/index.html (Accessed: 20 May 2021).

Wallach, W. and Allen, C. (no date) *CAN (RO)BOTS REALLY BE MORAL?*, *Moral Machines*. Oxford University Press. Available at: https://oxford.universitypressscholarship.com/view/10.1093/acprof:oso/9780195374049.001.0001/acprof-9780195374049-chapter-5 (Accessed: 18 May 2021).