Identifier le "bon" niveau d'explicabilité pour une situation donnée

Valérie Beaudouin, Isabelle Bloch, David Bounie, Stéphan Clémençon, Florence d'Alché-Buc, James Eagan, Winston Maxwell, Pavlo Mozharovskyi, Jayneel Parekh

IDENTIFIER LE « BON » NIVEAU D'EXPLICABILITÉ POUR UNE SITUATION DONNÉE

Valérie Beaudouin[1] et **Isabelle Bloch**[2] et **David Bounie**[1] et **Stéphan Clémençon**[2] et Florence dAlché-Buc[2] et James Eagan[2] et Winston Maxwell[1] et **Pavlo Mozharovskyi**[2] et **Jayneel Parekh**[2 1]

Résumé. Nous présentons un cadre de définition du « bon » niveau d'explicabilité, fondé sur des considérations techniques, juridiques et financières. Notre approche passe par une suite logique en trois étapes : *Premièrement,* la définition des principaux facteurs contextuels, tels que les destinataires visés par l'explication, le contexte opérationnel, les préjudices potentiels liés à l'utilisation du système et le cadre juridique et réglementaire. Cette étape va permettre de contribuer à la caractérisation des besoins d'explication sur le plan opérationnel et juridique et les avantages sociaux qui leur sont associés. *Deuxièmement,* le recensement des outils techniques pouvant être mobilisés, y compris des modèles a*d hoc* (perturbation des entrées, cartes de saillance ...) et des modèles hybrides d'IA. *Troisièmement,* en fonction des deux première étapes, le choix des bons niveau d'explicabilité en sorties globales et locales, prenant en compte leurs coûts. Nous identifions sept catégories de coûts et insistons sur le fait que les explications sont d'utilité sociale que dans la mesure où leurs avantages l'emportent sur leurs coûts.

1 INTRODUCTION

Cet article reprend les conclusions d'une étude plus longue [1] sur les explications propres au contexte dans une approche multidisciplinaire. L'explicabilité est une exigence aussi bien opérationnelle qu'éthique. L'exigence opérationnelle d'explicabilité découle de l'exigence d'une meilleure robustesse, notamment dans le cadre d'une application critique ; elle améliore également son niveau d'acceptation par l'utilisateur du système. L'exigence éthique d'explicabilité vise les atteintes aux droits fondamentaux et d'autres problématiques sociales qui pourraient ne pas être prises suffisamment en compte par les seules exigences opérationnelles. Les travaux de recherche actuels sur l'explicabilité de l'IA portent sur le volet informatique [18], ou sur le volet législation et politiques [20]. L'originalité du présent article est d'intégrer les démarches technique, juridique et financière dans une méthodologie unique pour atteindre le niveau optimal d'explicabilité. La dimension technique nous permet de cerner les explications possibles et l'équilibre à atteindre entre explicabilité et performance des algorithmes. Cependant, les explications dépendent toujours du contexte, qui lui-même découle du cadre règlementaire et d'une analyse des avantages et coûts que nous évoquons plus loin.

Notre approche passe par une suite logique en trois étapes : *Premièrement,* la définition des principaux facteurs contextuels, tels que les destinataires visés par l'explication, le contexte opérationnel, les préjudices potentiels liés à l'utilisation du système et le cadre juridique et règlementaire. Cette étape va permettre de contribuer à la caractérisation des besoins d'explication sur le plan opérationnel et juridique et les avantages sociaux qui leur sont associés. *Deuxièmement,* le recensement des outils techniques pouvant être mobilisés, y compris des démarches *ad hoc* (perturbation des entrées, cartes de saillance ...) et des modèles hybrides d'IA. *Troisièmement,* en fonction des deux première étapes, le choix des bons niveaux d'explicabilité en sorties globales et locales, prenant en compte leurs coûts.

Le recours à des solutions hybrides qui conjuguent l'apprentissage statistique à l'IA symbolique est un champ de recherche prometteur pour des applications critiques ou des applications telles que celles employées en médecine, pour lesquelles les décisions algorithmiques doivent s'appuyer sur un socle important de connaissances du domaine. Nous prévoyons que l'équilibre entre explicabilité et performance deviendra plus facile à atteindre au fur et à mesure que les solutions techniques d'explicabilité convergeront vers des modèles hybrides d'IA. L'explicabilité fera ainsi partie intégrante de la performance. Par ailleurs, l'explicabilité devenant une condition nécessaire à la certification en matière de sécurité, nous pouvons nous attendre à voir s'aligner les besoins d'explicabilité : les besoins opérationnels et de sécurité d'une part et les besoins éthiques et de droits fondamentaux d'autre part. Certaines solutions d'explicabilité opérationnelle peuvent remplir ces deux objectifs.

2 DÉFINITIONS

Nous considérons que les termes explicabilité et interprétabilité sont synonymes [16], même si on leur trouve d'autres définitions [1], et nous nous concentrons plutôt sur la différence importante qui existe entre l'explicabilité/l'interprétabilité dite « globale » et « locale ». L'explicabilité globale signifie la capacité à expliquer le fonctionnement d'un algorithme dans sa

[1] 1. I3, Télécom Paris, CNRS, Institut Polytechnique de Paris, France
2. LTCI, Télécom Paris, Institut Polytechnique de Paris, France – email: isabelle.bloch@telecom-paris.fr

globalité, alors que l'explicabilité locale signifie la capacité à expliquer une décision algorithmique en particulier. [7]. On emploie également le terme d'« explicabilité *ad hoc* » au lieu d'explicabilité locale.

La transparence est un concept qui dépasse celui de l'explicabilité [6], puisque la transparence englobe l'idée de l'accès à des données brutes, sans que celles-ci soient forcément compréhensibles. A l'inverse, l'explicabilité implique la transformation des données brutes afin de les rendre intelligibles par des humains. L'explicabilité devient ainsi un élément à valeur ajoutée de la transparence. La transparence et l'explicabilité ne se suffisent pas à elles-mêmes. En effet, elles agissent plutôt en tant que facilitatrices des autres fonctions telles que la traçabilité et l'auditabilité, toutes deux essentielles lorsqu'il s'agit de responsabilité. Ainsi, on peut considérer la responsabilité comme le summum de la gouvernance algorithmique [15] qui se nourrit d'autres concepts, dont l'explicabilité.

3 TROIS FACTEURS SERVENT À IDENTIFIER LE « BON » NIVEAU D'EXPLICABILITÉ

Notre approche recense trois considérations qui vont contribuer à établir le bon niveau d'explicabilité : les facteurs contextuels (entrées), les solutions techniques disponibles (entrées), et les choix d'explicabilité portant sur le format et le niveau des explications (sorties).

3.1 Facteurs contextuels

Nous identifions quatre catégories de facteurs contextuels qui vont nous permettre d'établir les différentes origines du besoin d'explication et de sélectionner l'explication qui convient le mieux (sortie) en fonction des moyens techniques et de leurs coûts. Les quatre facteurs contextuels sont répertoriés ci-dessous.

- Facteurs de destination : Qui est le destinataire de l'explication ? Quel est son niveau d'expertise ? Quelles sont ses contraintes de temps ? Ces questions génèrent des conséquences importantes en termes de volume de précisions fournies et de la temporalité de l'explication. [5, 7].
- Facteurs d'impact : Quels sont les préjudices que l'algorithme pourrait occasionner et comment les explications pourraient-elles remédier à ce problème ? Les réponses à cette question détermineront le niveau des avantages sociaux qu'offre l'explication. De manière générale, un impact important de l'algorithme va de pair avec celui des avantages découlant de l'explication [8].
- Facteurs règlementaires : Quel est le contexte règlementaire de cette application ? Quels sont les droits fondamentaux en jeu ? Nous examinons à la partie 5 ces facteurs qui vont nous permettre de caractériser les avantages sociaux conférés par l'explication dans un contexte donné.
- Facteurs opérationnels : Dans quelle mesure l'explication constitue-t-elle un facteur opérationnel incontournable ? Et en ce qui concerne la certification en matière de sécurité ? Et quid de la confiance des utilisateurs ? Ces facteurs pourront contribuer à établir des solutions permettant de réaliser des objectifs à la fois opérationnels, éthiques ou juridiques.

3.2 Solutions techniques

La disponibilité de solutions d'explication constitue un autre facteur en entrée. Les modèles *post hoc* telles que LIME [18], Kernal- SHAP [14] et les cartes de saillance [21] ont généralement pour but de calquer approximativement le fonctionnement d'un modèle de type « boîte noire », en ayant recours à un modèle d'explication distinct. Les modèles hybrides ont tendance à intégrer le besoin d'explication au sein même du modèle. Parmi ces approches figurent :

- La modification de fonctions objectives ou prédictives ;
- La mise en œuvre de règles floues, se rapprochant du langage naturel ;
- Les modèles en sortie [22] ;
- Les modèles en entrée, qui prétraitent les entrées du système d'apprentissage statistique, pour améliorer la pertinence et la structure des entrées. [1] ;
- Les règles floues génétiques.

La gamme des modèles hybrides possibles est quasi illimitée, les modèles pouvant associer apprentissage statistique et symbolique, ou bien une approche logique. Les exemples cités ci-dessus n'en représentent qu'un échantillon. La plupart des modèles sont en mesure de contribuer à l'explicabilité, même s'ils le font chacun de façon différente, qu'il s'agisse d'entrées, de sorties, ou de contraintes internes au système. L'explicabilité *by design* a principalement pour but d'intégrer cette dernière dans le système de prédiction.

3.3 Choix d'explication en sortie

La sortie d'explication consiste en ce qui est réellement communiqué au destinataire effectif de l'explication, qu'il agisse d'une explication globale du fonctionnement de l'algorithme ou de l'explication locale d'une décision en particulier.

Parmi les choix de sortie pour les explications *globales* figurent :

- L'adoption d'une approche de type « guide d'utilisation » pour présenter le fonctionnement de l'algorithme dans sa globalité [10] ;
- Le volume de précisions à fournir dans le guide d'utilisation ;
- La question de l'accès au code source, qui doit prendre en compte la protection du secret commercial mais aussi de l'usage limité du code source que pourraient en faire les destinataires de l'explication [10, 20] ;
- Des indications sur les données de formation et éventuellement la mise à disposition de ces données [10, 13, 17] ;
- Des informations sur l'algorithme d'apprentissage, y compris sa fonction objective ;
- Des informations sur les biais connus et les autres faiblesses que présente l'algorithme ; le signalement de

restrictions d'utilisation et d'avertissements.

Parmi les choix de sortie pour les explications *locales* figurent :

- Les tableaux de bord d'évaluation contrefactuelle, les utilisateurs finaux ayant la possibilité de tester des scénarios hypothétiques [20, 24] ;
- Les cartes de saillance qui montrent les facteurs principaux contribuant à la décision ;
- Le volume de précisions à fournir, y compris le nombre de facteurs et de coefficients pertinents à présenter aux utilisateurs finaux ;
- Les outils d'explication en plusieurs couches qui permettent aux utilisateurs d'accéder à des niveaux de complexité supérieure.
- L'accès au fichier journal propre à une décision [11, 26] ;
- Quelles sont les informations à conserver dans les rapports et combien de temps doivent-elles être stockées ?

4 L'EXPLICABILITÉ EN TANT QU'EXIGENCE OPÉRATIONNELLE

La recherche des années 90 sur l'explicabilité, tout comme l'intérêt récent que porte l'industrie à l'explicabilité, traitent essentiellement des explications qui ont pour objet de satisfaire les exigences opérationnelles des utilisateurs. Par exemple, un client peut avoir besoin d'explications dans le cadre de la validation de la sécurité d'un système d'IA et de l'accréditation de son processus ou demander que le système fournisse des informations supplémentaires à l'utilisateur final pour aider ce dernier (un radiologue, par exemple) à inscrire la décision dans un contexte clinique.

Des exigences opérationnelles d'explicabilité peuvent être demandées afin d'obtenir la certification d'applications critiques, puisque le système ne pourrait être mis sur le marché en l'absence de ces certifications. Par ailleurs, certains clients sont très demandeurs d'explications, afin de rendre le système plus facile à utiliser et d'augmenter la confiance des utilisateurs. Bien connaître les conséquences des différents facteurs accroît l'utilité du système, puisque les décisions prises s'accompagnent d'informations exploitables. Cette approche peut s'avérer beaucoup plus pertinente que la simple présence de prédictions qui ne sont pas expliquées. [25]. La compréhension de la causalité peut également renforcer la qualité en développant des systèmes plus robustes dans des domaines avec des entrées fluctuantes. Les clients sont de plus en plus nombreux à considérer l'explicabilité comme un critère de qualité du système d'IA. Ces exigences opérationnelles se distinguent des demandes d'explicabilité en ce qui concerne la règlementation, que nous évoquons à la partie 5 ; elles pourront néanmoins mener à une convergence des outils mis en œuvre pour répondre aux différents besoins.

L'explicabilité joue un rôle important dans l'assurance qualité des algorithmes, tant avant qu'après la mise sur marché du système, parce qu'elle contribue à révéler les faiblesses d'un algorithme, par exemple un bais qui serait passé inaperçu faute d'explication [9]. Elle contribue également aux modèles de type « cycle de vie entier du produit » [23] ou « sécurité en continu » [12] en matière de qualité et de sécurité algorithmiques.

La qualité des méthodes d'apprentissage statistique est souvent jugée à l'aune de son taux de précision lors de l'analyse des données de test. Ce critère de qualité ne suffit pas à refléter l'impact des faiblesses de l'algorithme sur sa qualité, notamment les biais et l'impossibilité d'une généralisation. Les solutions d'explicabilité présentées ici peuvent contribuer à identifier des zones de données en entrée pour lesquelles les algorithmes sont peu performants, ainsi que des défaillances dans les données d'apprentissage menant à des prévisions inexactes. Les démarches classiques de vérification et de validation (V&V) des logiciels sont mal adaptés aux réseaux neuronaux [3, 17, 23]. Le défi tient à la nature non-déterministe des réseaux neuronaux, qui complique la démonstration d'une absence de fonctionnalité non prévue, et à la capacité d'adaptation des algorithmes d'apprentissage statistique [3, 23]. Concernant les démarches traditionnelles de V&V et de certification, la définition d'un ensemble d'exigences servant à décrire intégralement le comportement d'un réseau neuronal représente le défi le plus important [2, 3]. Les exigences formulées de façon incomplète posent problème, car l'un des objectifs de la V&V est de comparer le comportement d'un logiciel à un document décrivant de façon précise et exhaustive le comportement attendu du système. [17]. En présence de réseaux neuronaux, une certaine incertitude subsiste quant à la nature exacte de la sortie pour une entrée donnée.

5 L'EXPLICABILITÉ EN TANT QU'EXIGENCE LÉGALE

Les modèles d'explication juridiques diffèrent selon qu'il s'agit de décisions émanant du secteur public ou privé. L'obligation qu'ont les États de fournir des explications repose sur le droit constitutionnel, par exemple, le respect du droit découlant de la Constitution des États-Unis ou le droit de s'opposer à une décision administrative en vertu de la législation européenne portant sur les droits fondamentaux. Ces droits stipulent que les personnes et les tribunaux doivent être en mesure de comprendre les raisons qui sous-tendent les décisions algorithmiques, de reproduire ces décisions afin d'effectuer des tests de défaillance et d'évaluer la proportionnalité des systèmes au regard d'autres droits fondamentaux tels que le droit à la vie privée. Aux États-Unis, le cas *Houston Teachers*[2] illustre le lien entre l'explicabilité et la garantie constitutionnelle de respect du droit. En Europe, la décision du Tribunal de grande instance de La Haye sur l'algorithme SyLI[3] montre que l'explicabilité entretient un lien étroit

[2] *Local 2415 v. Houston Independent School District*, 251 F. supp. 3d 1168 (S.D. Tex. 2017).

[3] *NJCM v. les Pays-Bas*, Tribunal de grande instance de La Haye, Cas n. C-09- 550982-HA ZA 18-388, 5 février 2020.

avec le principe constitutionnel de proportionnalité. La France a promulgué une loi sur les algorithmes déployés par l'État[4], qui contient des dispositions particulièrement strictes sur l'explicabilité du traitement algorithmique : communication du niveau et des modalités de contribution à la décision ; données utilisées pour le traitement et origine des données ; paramètres utilisés et coefficients intervenant dans les traitements individuels ; opérations effectuées pendant le traitement.

Pour les organismes privés, l'obligation d'explication se manifeste généralement lorsqu'un organisme est tenu à un devoir accru d'honnêteté et de loyauté, notamment s'il occupe une position anticoncurrentielle ou lorsque ses activités créent une relation de confiance ou de dépendance vis à vis de ses usagers. Il existe des lois spécifiques s'appliquant au secteur privé qui imposent des explications algorithmiques. Parmi elles, le Règlement européen pour les services d'intermédiation en ligne (UE) 2018/1150) est l'une des plus récentes ; elle impose une obligation d'explication au regard du classement des algorithmes à destination de ces services et des moteurs de recherche. Les éléments de langage du règlement illustre le fragile équilibre entre des principes qui s'opposent : mise à disposition des informations complètes, protection du secret commercial, précautions à prendre quant aux informations qui permettraient une manipulation entachée de mauvaise foi du classement des algorithmes, intelligibilité et utilité des explications vis à vis des usagers. Entre autres dispositions, les services d'intermédiation en ligne et les moteurs de recherche doivent fournir une « description raisonnée des principaux paramètres » qui peuvent avoir un impact sur le classement dans la plateforme, y compris « les critères et processus généraux ainsi [que les] signaux spécifiques intégrés dans les algorithmes ou [d']autres mécanismes d'ajustement ou de rétrogradation utilisés en relation avec le classement. »[5] Ces exigences sont plus détaillées que celles du Règlement général de la protection des données UE 2016/679 (RGPD), qui nécessite seulement « des informations des informations utiles concernant la logique sous-jacente ».[6] Aux États-Unis, les banques ont déjà l'obligation de fournir la raison principale d'un refus de prêt.[7] Une loi américaine actuellement à l'étude intitulée *Algorithmic Accountability Act* imposerait des obligations d'explicabilité concernant certains algorithmes à fort impact, y compris l'obligation de fournir des « descriptions détaillées de la conception, de la formation et des données et objectifs relatifs au système de décision automatisé. »[8]

6 AVANTAGES ET COÛTS DE L'EXPLICATION

En général, les lois et les règlements imposent des explications lorsque celles-ci sont bénéfiques à la société, autrement dit, lorsque le total des avantages qu'elles offrent l'emporte sur leurs coûts. Lorsqu'on s'intéresse à l'explicabilité algorithmique, une analyse des coûts et avantages permet de pallier l'indication du type d'explicabilité requis et du contexte, lorsque ces derniers n'ont pas encore été identifiés par la loi, en déterminant le bon niveau d'explication. Cette analyse sert à établir le pourquoi et le comment des explications à fournir, permettant ainsi de souligner et de gérer certains compromis. Pour que l'explicabilité soit utile à la société, les avantages doivent toujours l'emporter sur les coûts . Ces avantages sont étroitement liés au niveau d'impact de l'algorithme sur les droits individuels et collectifs [5, 8]. Concernant les algorithmes à impact faible, comme les algorithmes de recommandation de morceaux de musique, les avantages de l'explication sont mineurs. Concernant les algorithmes à impact fort, tel qu'un algorithme de reconnaissance d'images dans un véhicule autonome, les avantages de l'explication sont majeurs si par exemple on cherche à déterminer la cause d'une collision.

Les explications engendrent de nombreux coûts qui ne sont pas toujours faciles à déceler. Nous avons identifié sept catégories de coûts :

- Le coût de la conception et de l'intégration peut être élevé, parce que les exigences d'explicabilité varient en fonction des applications, des contextes et des territoires, ce qui signifie que les solutions d'explication unique et polyvalente suffisent rarement [9] ;
- Préférer l'explicabilité aux dépens de prévisions précises peut peser sur la performance, générant ainsi des coûts d'opportunité [5] ;
- La création et la conservation de journal de décision engendrent non seulement des coûts mais elles rentrent également en conflit avec les principes de protection des données qui nécessitent souvent que ces fichiers de journal soient détruits dès que possible [11, 26] ;
- La communication obligatoire du code source ou d'autres précisions concernant les algorithmes peut avoir un impact négatif sur le secret commercial protégé par le droit constitutionnel [4] ;
- Des explications détaillées sur le fonctionnement d'un algorithme peuvent favoriser le détournement du système avec pour conséquence potentielle un niveau de sécurité moindre ;
- Les explications sont à l'origine de règles implicites et de précédents, que les décisionnaires devront prendre en compte à l'avenir, limitant ainsi sa marge de manœuvre quant à ses décisions [19] ;
- La prescription de l'explicabilité peut allonger les délais de mise sur le marché, ralentissant d'autant l'innovation [9].

Pour des décisions algorithmiques à impact fort, ces coûts cèdent souvent la place aux avantages que présentent les explications. Cependant, ils méritent d'être pris en compte systématiquement, afin de s'assurer que le format et le volume de précisions relatives aux explications obligatoires sont bien adaptés à la situation.

4 Code des relations entre le public et l'administration, articles L. 311-3-1 et suivants

5 Règlement 2018/1150, considérant 24.

6 Règlement 2016/679, article 13(2)(f).

7 12 CFR Part 1002.9.

8 *Algorithmic Accountability Act* (Loi portant sur la « responsabilité algorithmique »), H.R. 2231 : projet de loi soumis le 10 avril 2019.

L'avantage social net (le total des avantages moins le total des coûts) doit être supérieur à zéro.

7 CONCLUSION : EXPLICATIONS D'IA SPÉCIFIQUES À LEUR CONTEXTE *BY DESIGN*

La règlementation de l'explicabilité d'IA reste à un stade embryonnaire, les efforts les plus ambitieux dans cette direction se trouvant en effet actuellement dans une loi française portant sur l'explicabilité des algorithmes déployés par l'État et dans une règlementation de l'UE sur les services d'intermédiation. Or, la loi laisse la place à de nombreuses interprétations de l'explicabilité, même dans ces textes. Le format de l'explication et le volume de précisions fournies seront largement influencés par les quatre catégories de facteurs contextuels décrits dans cet article : facteurs de destinataires, d'impact, de règlementation et opérationnels. Le volume de précisions des explications globales ou locales fournirait ainsi une échelle glissante selon le contexte, et selon les avantages et les coûts en jeu. La conservation des fichiers de journal concernant des décisions individuelles représente un des coûts majeurs des explications locales. Le type de données conservées dans les fichiers de journal et la durée de leur stockage sont des questions essentielles quant à l'établissement du bon niveau d'explicabilité. Les solutions hybrides tentent de mettre en place des modèles d'explicabilité *by design*, notamment en intégrant cette explicabilité dans le modèle prédictif. Ces solutions portent certes sur les besoins opérationnels, mais elles peuvent également satisfaire les besoins éthiques et juridiques d'explicabilité. Notre méthode en trois étapes s'appuie sur les facteurs contextuels, les solutions techniques et les sorties d'explicabilité permettant d'arriver au « bon » niveau d'explicabilité dans une situation donnée.

REFERENCES

[1] Valérie Beaudoin, Isabelle Bloch, David Bounie, Stéphan Clémencon, Florence d'Aché Buc, James Eagan, Maxwell Winston, Pavlo Mozharovskyi, and Jayneel Parekh, 'Flexible and context-specific AI explainability: a multidisciplinary approach', Technical report, ArXiv, (2020).

[2] Siddhartha Bhattacharyya, Darren Cofer, D Musliner, Joseph Mueller, and Eric Engstrom, 'Certification considerations for adaptive systems', in *2015 IEEE International Conference on Unmanned Aircraft Systems (ICUAS)*, pp. 270–279, (2015).

[3] Markus Borg, Cristofer Englund, Krzysztof Wnuk, Boris Duran, Christoffer Levandowski, Shenjian Gao, Yanwen Tan, Henrik Kaijser, Henrik Lönn, and Jonas Törnqvist, 'Safely entering the deep: A review of verification and validation for machine learning and a challenge elicitation in the automotive industry', *Journal of Automotive Software Engineering*, 1(1), 1–19, (2019).

[4] Jenna Burrell, 'How the machine thinks: Understanding opacity in machine learning algorithms', *Big Data & Society*, 3(1), 2053951715622512, (2016).

[5] Finale Doshi-Velez, Mason Kortz, Ryan Budish, Chris Bavitz, Sam Gershman, David O'Brien, Stuart Schieber, JamesWaldo, David Weinberger, and Alexandra Wood, 'Accountability of ai under the law: The role of explanation', *arXiv preprint arXiv:1711.01134*, (2017).

[6] European Commission, 'Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions - Building trust in human centric artificial intelligence (com(2019)168)', Technical report, (2019).

[7] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi, 'A survey of methods for explaining black box models', *ACM Computing Surveys (CSUR)*, 51(5), 93, (2018).

[8] AI HLEG, 'High-level expert group on artificial intelligence', *Ethics Guidelines for Trustworthy AI*, (2019).

[9] ICO, 'Project ExplAIn interim report', Technical report, Information Commissioners Office, (2019).

[10] IEEE, 'Ethically aligned design: A vision for prioritizing human wellbeing with autonomous and intelligent systems', *IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems*, (2019).

[11] Joshua A Kroll, Solon Barocas, Edward W Felten, Joel R Reidenberg, David G Robinson, and Harlan Yu, 'Accountable algorithms', *U. Pa. L. Rev.*, 165, 633, (2016).

[12] Zeshan Kurd and Tim Kelly, 'Safety lifecycle for developing safety critical artificial neural networks', in *Computer Safety, Reliability, and Security*, eds., Stuart Anderson, Massimo Felici, and Bev Littlewood, pp.77–91, Berlin, Heidelberg, (2003). Springer Berlin Heidelberg.

[13] David Lehr and Paul Ohm, 'Playing with the data: what legal scholars should learn about machine learning', *UCDL Rev.*, 51, 653, (2017).

[14] Scott M Lundberg and Su-In Lee, 'A unified approach to interpreting model predictions', in *Advances in Neural Information Processing Systems*, pp. 4765–4774, (2017).

[15] OECD, *Artificial Intelligence in Society*, 2019.

[16] OECD, *Recommendation of the Council on Artificial Intelligence*, 2019.

[17] Gerald E Peterson, 'Foundation for neural network verification and validation', in *Science of Artificial Neural Networks II*, volume 1966, pp. 196–207. International Society for Optics and Photonics, (1993).

[18] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin, 'Why should I trust you?: Explaining the predictions of any classifier', in *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144, (2016).

[19] Frederick Schauer, 'Giving reasons', *Stanford Law Review*, 633–659, (1995).

[20] Andrew Selbst and Solon Barocas, 'The intuitive appeal of explainable machines', *SSRN Electronic Journal*, 87, (01 2018).

[21] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman, 'Deep inside convolutional networks: Visualising image classification models and saliency maps', *arXiv preprint arXiv:1312.6034*, (2013).

[22] Philip S. Thomas, Bruno Castro da Silva, Andrew G. Barto, Stephen Giguere, Yuriy Brun, and Emma Brunskill, 'Preventing undesirable behavior of intelligent machines', *Science*, 366(6468), 999–1004, (2019).

[23] US Food and Drug Administration, 'Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML)-based software as a medical device', Technical report, (2019).

[24] Sandra Wachter, Brent Mittelstadt, and Chris Russell, 'Counterfactual explanations without opening the black box: Automated decisions and the gpdr', *Harv. JL & Tech.*, 31, 841, (2017).

[25] Max Welling, 'Are ML and statistics complementary?', in *IMS-ISBA Meeting on Data Science in the Next 50 Years*, (2015).

[26] Alan FT Winfield and Marina Jirotka, 'The case for an ethical black box', in *Annual Conference Towards Autonomous Robotic Systems*, pp. 262–273. Springer, (2017).