IEAI Technical University of Munich
Munich Center for Technology in Society
Institute for Ethics in Artificial Intelligence

Research Brief – February 2022

Towards an Accountability Framework for AI: Ethical and Legal Considerations

By Auxane Boch, Ellen Hohma, Rainer Trauth

With the growth of what were once smaller AI applications into highly complex systems, the issue of who is responsible for the predictions or decisions made by these systems has become pressing. Using the example of autonomous driving, this Brief highlights major accountability problems for AI systems from a legal and ethical perspective. Implications for accountability, including explainability and responsibility requirements, can already be found in current guidelines. However, the transfer and application of these guidelines to the specific context of AI systems, as well as their comprehensiveness, requires further efforts, particularly in terms of societal demands. Therefore, further multidisciplinary investigations are required to strengthen the development of holistic and applicable accountability frameworks.

Application areas of artificial intelligence (AI) have grown rapidly in recent years. Patent registrations and AI-enabled inventions are increasing and are being adopted by industry as AI applications promise higher performance (Zhang, 2021). Small use cases are growing into larger and more complex systems that directly impact people's lives. Unlike previous technological advances, far-reaching decisions can be made without direct human intervention. This brings to light a new concern regarding how such systems can be made accountable, a major feature of which is improving the general understandability, or explainability, of such technologies. Explainability of these systems can help define and delineate accountability, and with that responsibility, more clearly.

Accountability is defined as "the fact of being responsible for what you do and able to give a satisfactory reason for it" (Cambridge Dictionary, 2022). Consequently, accountability consists of two components: (1) responsibility, defined as "something that it is your job or duty to deal with", and (2) explanation, i.e., "the details or reasons that someone gives to make something clear or easy to understand" (Cambridge Dictionary, 2022). In particular for AI systems, realizing those two concepts is a difficult task. As AI systems, such as neural networks, are often black box systems, i.e., the connection between input and output parameters is opaque, an explanation of how the system derived a certain prediction or conclusion is not always obvious. This problem may even intensify in the future, as explicability of algorithms decreases with their increasing complexity. Additionally, missing explanation exacerbates the issue of responsibility, making it hard to determine duties if the decision process or source of failure is not entirely clear.

In this Brief, we will investigate current challenges for accountability of AI systems from two perspectives; legal and ethical. The challenges for accountability from a practical perspective will be introduced using the example of autonomous vehicles. We will further examine how and which obligations to explainability arise from currently existing legal frameworks, how product liability standards can be translated into the specific use case of AI systems and the challenges that may arise. As an outlook, we will discuss the need for a broader accountability approach, the requirements for a multi-level explainability, and the bigger picture of social responsibility.

Autonomous Driving as a Use Case

Autonomous driving is arguably one of the most elaborate existing research areas. Many scientific disciplines and technologies are needed to realize the dream of a fully self-driving car. The interaction of the vehicle in the real world is a non-deterministic process, so that action sequences of the vehicle cannot be unambiguously determined. The high complexity of the driving task requires a high level of experience so that even difficult driving situations can be handled. AI and Machine Learning (ML) promise to solve these types of problems by using huge amounts of data. These large amounts of recorded data can be considered as experiences of the machine, which correspond to the past experiences of a brain. As with humans, decision-making based on historical experiences is not transparent or obvious to external individuals. In the technological context, opacity increases with system complexity. However, these challenges must be overcome for the technology to be widely adopted by customers and society (Amodei, 2016). In any case, to ensure acceptable products, each subsystem of the vehicle must be safeguarded in the process to avoid errors that can occur during the development phase and during operation.

This comes into conflict with the finding that users want techniques that are directly interpretable, comprehensible, and trustworthy (Zhu, 2018). The demand for ethical AI is increasing and has become thematically more important in society (Goodman, 2017). Therefore, trade-offs must assess between the performance of a model and its transparency, or explainability, in order to meet societal, and not just technological, needs (Dosilovic, 2018). One way to develop transparent but powerful algorithms is to use Explainable AI (XAI) methods. "XAI will create a suite of machine learning techniques that enables human users to

*This research is derived from the IEAI project „Towards an Accountability Framework for AI Systems: The Autonomous Vehicle Use Case", which is generously support by Fujitsu.

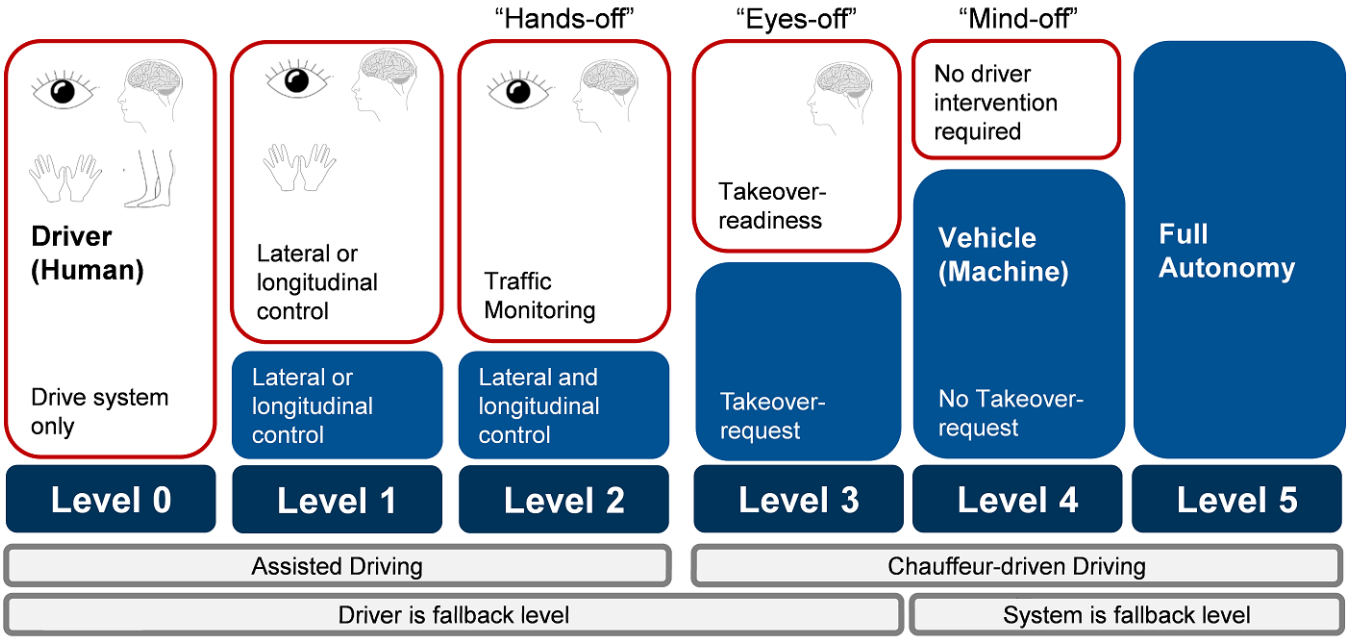

Figure 1: SAE Levels of Autonomous Driving (SAE International, 2018)

understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partner" (Gunning, 2017).

The essential goals of XAI are therefore understanding and trust. There are two technical ways to achieve this: either the development of transparent models from scratch or the post-hoc explainability of ML models (Arrieta, 2020). In transparent models, attention is paid to the requirements of the model already in the design process. Examples of techniques in this area include linear/logistic regression, decision trees or rule-based learning. Models from this domain are easy to understand, but limited in performance. Post-hoc explainability techniques, on the other hand, examine AI models where explainability was not considered during design process. Some of these algorithms analyze the black-box environment of the ML model to obtain information about the relationship between input and output through perturbations. With the help of these methods, the transparency of decisions can also be increased in the field of autonomous driving.

Figure 2 shows an attention map created in the temporal context of autonomous driving using camera recognition data (Kim, 2018).

Important areas of the camera that are essential for decision-making can be visualized. To make the decision process understandable for the user, other methods can be used, such as a linguistic explanation (Arrieta, 2020). The explanation in Figure 2 of the vehicle could then be as follows: "The car heads down the street, because the street is clear" (Zablocki, 2018).

XAI can contribute to minimizing and reporting negative impacts in the field of AI from a technical point of view. XAI methods are intended to provide the same level of transparency in the long term as other technical systems that operate without AI. In addition to decision-making, XAI can also help detect irregularities in the data. Unfair practices, such as discrimination against ethically marginalized groups, can be reduced. With these technical capabilities, responsible AI should thus be developed.

Nevertheless, further discussions are needed to actually achieve implementable accountability. Legal and ethical challenges, for instance, must be overcome to develop responsible AI for the future.

Legal Obligations to Accountability

The definition introduced above outlines that accountability consists of two components, being responsible for a decision or action, as well as giving appropriate explanation for it. As both of these concepts have already been brought up in other, digital or non-digital, contexts, obligations for AI systems can be examined with regard to

Figure2: Attention Map Generation (Zablocki, 2018, Kim, 2018)

existing legal guidelines. However, there are major difficulties for translating common general approaches towards regulating explainability and responsibility to the specific context of AI, some of which we will investigate below. systems can be examined with regard to major difficulties for translating common general approaches towards regulating explainability and responsibility to the specific context of AI, some of which we will investigate below.

Obligations to Explainability

Indicators for a right to explanation can be found in many legal frameworks, such as contract and tort law or consumer protection law (Sovrano et al., 2021). In particular regarding the use of personal data, the General Data Protection Regulation (Regulation 2016/679; GDPR) can be used as a first legal reference that regulates transparency obligations in the European Union (EU). Specifically, as AI systems are highly dependent on the use of data and, in some cases, the processing of personal data cannot be entirely avoided (e.g. recording high-resolution maps for autonomous driving including images of pedestrians in the streets), GDPR has a strong impact on data governance in AI systems.

If the GDPR is applicable, the data processor is obliged to provide the data subject with certain information on, for example, collection and use of personal data. Aiming now at deriving a general right to explanation for AI processes, essentially two main grounds can serve as a starting point in the GDPR (Ebers, 2020; Hacker & Passoth, 2021). First, Article 22 regulates "automated individual decision-making, including profiling" and, in certain cases, obliges the data controller to "implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision" (Art. 22(3), GDPR). A second anchor can be found in Article 15 regulating the "right of access by the data subject", which grants the data subject a right to "meaningful information about the logic involved" (Art. 15(1)(h), GDPR).

Autonomous vehicle systems are an ideal example of complex applications that could revolutionize their industry in the coming decades. However, there are already examples in which the system's opacity has led to long legal disputes due to the lack of clarity in regard to liability issues (Griggs & Wakabayashi, 2018). In 2018, for example, an Uber test vehicle hit a pedestrian even though a safety driver was on board of the vehicle. In the end, the test driver was held liable. However, this was not clear at the beginning.

However, major debates have emerged among scholars concerning whether those articles translate to a general right to explanation for AI processes (Bibal et al., 2021; Ebers, 2020; Felzmann et al., 2019). First, Article 22 applies to decisions "based solely on automated processing" (Art. 22(1), GDPR). Many AI-enabled systems still allow for human intervention. Therefore, they are not covered by this obligation (Ebers, 2020).

Further, it is controversial among legal scholars as to whether Article 22 actually grants a right to explanations of individual decisions (Bibal et al., 2021; Ebers, 2020). Recital 71, providing additional interpretive guidance on Article 22, mentions such a right. However, it has not been explicitly incorporated in the binding section of the GDPR (Ebers, 2020). Article 15 seems more promising, as it directly refers to "meaningful information". But it does not elaborate on the appropriate level of detail of the provided information. Some argue whether all information is needed to explain individual decisions or only the overall structure and functionalities of an AI-based system needs to be disclosed (Ebers, 2020; Hacker & Passoth, 2021).

Accountability consists of two components, being responsible for a decision or action, as well as giving appropriate explanation for it.

To obtain more clarity on the specific transparency requirements for AI systems, a first outlook can be found in the proposal for a Regulation of the European Parliament and of the council laying down harmonized rules on Artificial Intelligence (Regulation 2021/0106, AI Act), also referred to as AI Act. The European Commission proposes to take a risk-based approach and categorize AI applications into three levels, (1) prohibited AI practices, (2) high-risk AI systems and (3) AI systems of minimal risk. While prohibited AI practices bear an unacceptable risk and will therefore be banned from "placing on the market, putting into service or use" (Art. 5(1)(a), AI Act) in the EU, certain transparency requirements are imposed on high-risk or recommended for minimal-risk AI systems. For high-risk AI systems, the AI Act in its current implementation mainly demands transparency on data and data governance to prevent bias (Art. 10, AI Act), technical documentation of the general algorithmic logic to demonstrate the compliance with the AI Act requirements (Art. 11 & Annex VI, AI Act), record keeping to allow for monitoring and increase traceability (Art. 12, AI Act), as well as further transparency obligations to allow users to interpret and appropriately use the system's outputs (Art. 13, AI Act). While the current proposal of the AI Act is already significantly more concrete and tightly adapted to the specific circumstances of AI systems than more generic legal frameworks, there is still some criticism about its practical applicability. For instance, although the AI Act concretely elaborates on transparency measures to be put in place, the degree of required transparency is still left vague. Instead, it refers to an appropriate level, still allowing much room for variation and interpretation.

We see that legislation recognizes the issue of transparency and its intensified necessity in the context of AI systems. Current legislation can be adapted to the use of AI, as well as new directives have been put in place. However, the concept of explainability is targeted through transparency, which refers to transparent algorithmic processes rather than clearly traceable decision-making. Explainability issues have not been clarified in their entirety yet, as there are still unresolved questions in particular on the level of interpretability in the concrete application contexts.

Furthermore, a prevailing problem is the question of which rights are granted to whom and against whom, i.e., who can demand which explanation from whom. An example of this unresolved tension is the GDPR granting explanation rights to the data subject, who is, however, not necessarily the system's user. Exemplary is equally the AI Act, which does not grant the user any claim for reparation, only penalizes noncompliant behavior. Therefore, although more concrete guidelines that regulate the pressing matter of explanation and transparency are already in place or initiated, the issue has not yet been fully resolved and the main task now is to reconcile the theoretical conceptions with practice.

Implications of Liability

The second component of accountability from a legal perspective is understood as liability, i.e., "the state of being legally responsible for something" (Cambridge Dictionary, 2022). Liabilities for businesses are, of course, manifold and can result from many different obligations and regulations, such as from just described data abuse according to GDPR, or transparency duties from the newly proposed AI Act. An interesting research field for AI is liabilities due to system failure according to obligations derived from

product liability directives. On the one hand, this is the major legal starting point when studying responsibility distribution for a system error. On the other hand, it still needs to be investigated how AI systems fit into the directive's prerequisites.
Liability here refers to the "non-contractual, strict liability of the producer for damage caused by a defective product" (Borges, 2021, p. 33). It is therefore not a matter of faults, it instead depends on certain preconditions, such as which interests are affected or how the damage is caused. The approximation of the laws, regulations and administrative provisions of the Member States concerning liability for defective products (Council Directive 85/374/EEC, The Product Liability Directive, PLD), which came into force in the EU in 1985, gives guidance on these prerequisites and sets the standards for strict liability for a product's defects. Essentially, the producer of a product and respectively the manufacturer of the defect component are liable for damage to one of the rights protected under the directive. A complete and exhaustive list of rights that are protected is provided, mainly including death, personal injury and damage or destruction of property. Further, the damage must be caused by the product, in particular, by the defect of the product.

While this derivation sounds reasonable for most tangible products, questions arise if the directive is to be applied to AI systems. A first major precondition is that the damage is caused by a product, defined as "all movables […], even though incorporated into another movable or into an immovable" (Art. 1, Product Liability Directive (PLD)). In theory, it is therefore arguable if AI falls under the PLD, as it is not in line with this definition of movable objects and is usually provided as a service. In practice, however, software and, hence, AI created by software shall be treated like a product and be subject to the same liability standards (Cabral, 2020). Worth mentioning is also the scope that the PLD sets, limiting liabilities to damage to the health or property of private users. It excludes claims of commercial users, as well as mere financial loss due to the product failure (Borges, 2021). Furthermore, other rights, such as personality rights are currently not covered by product liability rights (Borges, 2021). This is particularly relevant for AI systems, as damage caused by incorrect assessment due to, for example, bias is not covered by product liability.

A second source of difficulties in applying the PLD to AI systems is the question of who accounts for which component. This is less obvious for AI than for physical products. According to the directive, the system manufacturer is responsible for the entire product and is jointly liable with the supplier of the defective component. In the case of software in general and AI in particular, however, it is more difficult to distinguish the individual components and therefore to derive concrete liabilities. An example is the distinction between the algorithmic conception of an AI and its trained implementation. Resolving interpretations suggest that an untrained AI model is seen as basic material while the network fed with data is considered a component equivalent (Borges, 2021). This highlights that a translation of current product liability standards to AI technologies is possible, however, more clarity on the concrete adaptation in this context is needed.

The lack of transparency surrounding the dataset, and its possible bias, and inner workings of the decision-making process of the systems building off of those data can result in the exclusion of certain populations from the AI decision-making process.

What this overview shows is that there already exists a backbone for both responsibility and explainability obligations of accountability in current legal frameworks. However, the transfer to the special use case of AI, for instance by drawing connections between standard products and AI systems, is still ongoing. The explainability component of accountability is particularly relevant as transparency (a necessary component of explainability) and responsibility are highly intertwined. Explanation of a decision can show reasoning and, hence, shed light on accountability of the engaged actors. Therefore, explainability becomes even more pertinent in legal considerations, as it can prove or disprove the liability of parties involved.

Ethical Challenges for Accountability

While the law may lay down good first steps for an accountability framework, an ethical framework goes beyond this and calls for more. Indeed, accountability, when considered a sum of responsibility and explainability, does not concern only the technical aspects of a tool, but also its global impact on society.

Accountability, on a higher level, can be defined as the relationship between an actor and the group (e.g. society) to which the actor holds an obligation to justify their conduct (Bovens, 2007).

For the autonomous driving example, explainability is an inevitable first step towards a holistic accountability framework. Installing a data recorder or 'black box' similar to the ones used in planes is often mentioned as a means to document inherent processes that led to a system failure. However, to record more complex parameters than a car's mere speed or geographic location that can give further direction on the responsibilities, knowledge about currently hidden details of opaque techniques is required. This opaqueness and the lack of strong experience in unifying them with recent legal regulations highlight the need for further investigations in this research field.

It is what allows critiques and praise regarding the performance of a stakeholder, and relates to their active choice to give information regarding their behaviors (Bovens et al., 2014). Using this approach, the need for explainability of the AI-powered tool implemented, and discussion relating to its use and impact, is quite clear. Additionally, a judgment entailing formal or informal positive or negative consequences can be passed onto the actor's choices, thus on the product proposed by said actor (Ackerman, 2005; Bovens, 2007; Olson, 2018).

However, some main questions remain for a proper application of the accountability concept in the context of AI – the identification of its different stakeholders and the way responsibility needs to be shared between them (Gevaert et al., 2021).

The Need for Explainability

As stated in the previous definitions, explainability in regard to a tool's quality and use is needed to build a good accountability evaluation. Even if a comprehensive accountability evaluation is broadly researched on the technical side of an AI application, to shed light on the inner workings of different types of algorithms and the data used to train them, more considerations need to be made to reach an acceptable global level of explainability (Gevaert et al., 2021).

Some main questions remain for a proper application of the accountability concept in the context of AI – the identification of its different stakeholders and the way responsibility needs to be shared between them.

The opacity of AI systems creates an imbalance in society for most vulnerable groups, in particular, due to the presence of bias towards said communities in the dataset used to train the tool. The lack of transparency surrounding the dataset, and its possible bias, and inner workings of the decision-making process of the systems building off of those data can result in the exclusion of certain populations from the AI decision-making process (Barocas & Selbst, 2016). The identification of such bias prior to implementation, and resolution of the issues through methodical approach would reduce this type of inequality. Thus, the use, evaluation of data and correction of possible bias by AI systems producers need to be able to be evaluated and punished or resolved in the court of public opinion.

A major issue faced to reach a proper acknowledgment and evaluation of such situation is the dissonant explanations given to humans, as compared to how one typically constructs explanations, impeding a good understanding of the problem (Miller, 2019). Indeed,

understandability differs from one individual to another depending on their personal context and the aim of the explanation. This influences how appropriate and useful a given "why" and "how" explanation is (Ribera & Lapedriza, 2019; Miller, 2019; Hoffman et al., 2018; The Alan Turing Institute, 2019).

In the case of autonomous driving, the recognition of traffic signs can be taken for example. Indeed, the passenger must be given the opportunity to react to the uncertainty if it occurs. If a traffic sign is intentionally manipulated, explainability from the accountability perspective, the developer perspective, and the user perspective helps to solve misunderstandings (Eykholt et al., 2018).

Gunning and Aha (2019), sum up three major needs for a good explainable AI: (1) to produce more explainable models, which would be necessary for a technical and social assessment of an AI product (2) to design better explanation interfaces to facilitate the interaction with the knowledge required, and (3) to understand the psychological requirements to deliver effective explanations to humans, which will impact the opportunity for technically literate or not individuals belonging to a given society to participate in the evaluation of a tool. This last point is of main interest in regard to our accountability approach, mainly, the importance of involving all actors of society's opinion and discussion in the definition of responsibility and consequences. In other words, this final point is of paramount importance for including communities served by the AI-producing stakeholders in the accountability distribution (van den Homberg et al., 2020).

More than Accountability, a Social Responsibility

More than its explainability requirement, accountability calls for an "ethical sense of who is responsible for the way AI works" (Floridi & Cowls, 2019, p.8). AI is not only a technical problem, but a social one due to its possible impact on society. Thus, the identification of responsible and accountable actors needs to be thoroughly approached to consider the global frame of social responsibility towards the groups impacted by the AI tool. Indeed, such technologies can impact communities' life, safety and well-being (Saveliev & Zhurenkov, 2020). As social responsibility can be understood as the consequential need for a stakeholder's actions to benefit society. A balance must be struck between economic/technological growth and well-being of the group. At this point, defining what is meant by "well-being of the group" differs by each culture and subculture accordingly.

In the case of autonomous vehicles, if it became the primary means of transport, social responsibility would translate as the reduction of 90% of all vehicles crashes, saving lives, and billions of dollars to societies (Bertoncello & Wee, 2015).

Building off 47 ethical principles for a socially beneficial AI, Floridi and Cowls (2019) proposed a five principles approach to AI ethical evaluation. Starting off with four core principles borrowed from bioethics – beneficence, non-maleficence, autonomy, and justice – the authors argued for the need of an additional one to support understandability and accountability; the principle of explainability (Saveliev & Zhurenkov, 2020). This last pillar enables other principles to be evaluated, allowing for a social impact, and thus social responsibility consideration for each AI-tool proposed.

Final Thoughts

In this research brief, we highlighted the most pressing questions regarding accountability of AI products. Today's frameworks and regulations do not provide clear answers to the questions of how we should deal with accountability problems. Even if new legislation, such as the AI Act, is introduced,

the specific application to business processes and technologies is not yet clear.

The increasing use of artificial intelligence in products also creates further accountability dependencies. The more complex the systems become, the greater the impact on people and society.

Moreover, providing a clearer path to understanding AI systems and stakeholders involved in their creation, implementation, and use is of paramount importance in the consideration of social responsibility and accountability. A lot more needs to be done in regard to accountability definition whether technically or socially.

The introduction of autonomous driving level 3 (SAE International, 2018) in Germany can be cited as an example of increase dependencies. In this case, the driving task is completely left to the system for the first time. A time transition period is defined, which must be granted to the driver to take back control. Handing power over to the machine and limiting human oversight and control creates new questions of how to deal with malfunctioning and system errors. Therefore, a clear definition and distribution of responsibilities through frameworks and regulation to mitigate potential risks is inevitable.

References

Ackerman, J. M. (2005). *Human rights and social accountability*. Participation and Civic Engagement, Social Development Department, Environmentally and Socially Sustainable Development Network, World Bank.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion*, *58*, 82-115.

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *Calif. L. Rev.*, *104*, 671.

Bertoncello, M., & Wee, D. (2015, June). *Ten ways autonomous driving could redefine the automotive world.* McKinsey & Company.

Bibal, A., Lognoul, M., De Streel, A., & Frénay, B. (2021). Legal requirements on explainability in machine learning. *Artificial Intelligence and Law*, *29*(2), 149-169.

Borges, G. (2021, June). AI systems and product liability. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law* (pp. 32-39).

Bovens, M. (2007). Analysing and assessing accountability: A conceptual framework 1. *European law journal*, *13*(4), 447-468.

Bovens, M., Goodin, R. E., & Schillemans, T. (Eds.). (2014). *The Oxford handbook public accountability*. Oxford University Press.

Cabral, T. S. (2020). Liability and artificial intelligence in the EU: Assessing the adequacy of the current Product Liability Directive. *Maastricht Journal of European and Comparative Law*, *27*(5), 615-635.

Cambridge Dictionary. (2022, February 23).*Accountability*. https://dictionary.cambridge.org/dictionary/english/accountability

Cambridge Dictionary. (2022, February 23).*Explanation*. https://dictionary.cambridge.org/dictionary/english/explanation

Cambridge Dictionary. (2022, February 23).*Liability*. https://dictionary.cambridge.org/dictionary/english/liability

Cambridge Dictionary. (2022, February 23).*Responsibility*. https://dictionary.cambridge.org/dictionary/english/responsibility

Cognizant, & Piroumian, V. (2014, April). *Why Risk Matters: Deriving Profit by Knowing the Unknown.*

Council Directive 85/374/EEC. *The approximation of the laws, regulations and administrative provisions of the Member States concerning liability for defective products.* European Parliament, Council of the European Union. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:31985L0374&from=EN

Došilović, F. K., Brčić, M., & Hlupić, N. (2018, May). Explainable artificial intelligence: A survey. In *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)* (pp. 0210-0215). IEEE.

Ebers, M. (2020). Regulating Explainable AI in the European Union. An Overview of the Current Legal Framework (s). *An Overview of the Current Legal Framework (s)(August 9, 2021). Liane Colonna/Stanley Greenstein (eds.), Nordic Yearbook of Law and Informatics*.

Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., ... & Song, D. (2018). Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1625-1634).

Felzmann, H., Villaronga, E. F., Lutz, C., & Tamò-Larrieux, A. (2019). Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns. *Big Data & Society*, *6*(1), 2053951719860542.

Floridi, L., & Cowls, J. (2021). A unified framework of five principles for AI in society. In *Ethics, Governance, and Policies in Artificial Intelligence* (pp. 5-17). Springer, Cham.

Gevaert, C. M., Carman, M., Rosman, B., Georgiadou, Y., & Soden, R. (2021). Fairness and accountability of AI in disaster risk management: Opportunities and challenges. *Patterns*, *2*(11), 100363.

Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a "right to explanation". *AI magazine*, *38*(3), 50-57.

Gunning, D., & Aha, D. (2019). DARPA's explainable artificial intelligence (XAI) program. *AI magazine*, *40*(2), 44-58.

Gunning, D. (2017). Explainable artificial intelligence (XAI)(2017). *Seen on*, *1*.Tech. rep., Defense Advanced Research Projects Agency (DARPA)

Hacker, P., & Passoth, J. H. (2021). Varieties of AI Explanations Under the Law. From the GDPR to the AIA, and Beyond. *From the GDPR to the AIA, and Beyond (August 25, 2021)*.

Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2018). Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608*.

Kim, J., Rohrbach, A, Darrell, T, Canny, JF, Akata, Z. (2018). Textual explanations for self-driving vehicles. ECCV

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, *267*, 1-38.

Griggs, T., & Wakabayashi, D. (2018, July 30). *How a Self-Driving Uber Killed a Pedestrian in Arizona*. The New York Times. https://www.nytimes.com/interactive/2018/03/20/us/self-driving-uber-pedestrian-killed.html

Olson, R. S. (2018). Establishing public accountability, speaking truth to power and inducing political will for disaster risk reduction:'Gcho Rios+ 25'. In *Environmental Hazards* (pp. 59-68). Routledge.

Regulation 2016/679. *General Data Protection Regulation*. European Parliament, Council of the European Union. https://eur-lex.europa.eu/legalcontent/EN/TXT/PDF/?uri=CELEX:32016R0679

Regulation 2021/0106. *Regulation of the European Parliament and of the council laying down harmonized rules on Artificial Intelligence*. European Parliament, Council of the European Union. https://eur-lex.europa.eu/legalcontent/EN/TXT/?uri=CELEX%3A52021PC0206

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135-1144).

Ribera, M., & Lapedriza, A. (2019, March). Can we do better explanations? A proposal of user-centered explainable AI. In *IUI Workshops* (Vol. 2327, p. 38).

Russell, S. J., & Norvig, P. (2016). *Artificial intelligence: a modern approach*. Malaysia. Pearson Education Limited.

SAE International (2018).*Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles (J3016)*. https://saemobilus.sae.org/content/J3016_201806

Saveliev, A., & Zhurenkov, D. (2020). Artificial intelligence and social responsibility: the case of the artificial intelligence strategies in the United States, Russia, and China. *Kybernetes*.

Sovrano, F., Sapienza, S., Palmirani, M., & Vitali, F. (2021). A Survey on Methods and Metrics for the Assessment of Explainability under the Proposed AI Act. *arXiv preprint arXiv:2110.11168*.

The Alan Turing Institute. (2019). *Explaning decisions made with AI: Part 1: The basics of explaning AI*. Information Commissioner's Office (ICO). https://ico.org.uk/for-organisations/guide-to-data-protection/key-dp-themes/explaining-decisions-made-with-artificial-intelligence/part-1-the-basics-of-explaining-ai/

van den Homberg, M. J., Gevaert, C. M., & Georgiadou, Y. (2020). The changing face of accountability in humanitarianism: Using artificial intelligence for anticipatory action. *Politics and Governance*, *8*(4), 456-467.

West, D. M. (2018). *The future of work: Robots, AI, and automation*. Brookings Institution Press.

Zablocki, É., Ben-Younes, H., Pérez, P., & Cord, M. (2021). Explainability of vision-based autonomous driving systems: Review and challenges. *arXiv preprint arXiv:2101.05307*.

Zhang, D., Mishra, S., Brynjolfsson, E., Etchemendy, J., Ganguli, D., Grosz, B., ... & Perrault, R. (2021). The ai index 2021 annual report. *arXiv preprint arXiv:2103.06312*.

Zhu, J., Liapis, A., Risi, S., Bidarra, R., & Youngblood, G. M. (2018, August). Explainable AI for designers: A human-centered perspective on mixed-initiative co-creation. In *2018 IEEE Conference on Computational Intelligence and Games (CIG)* (pp. 1-8). IEEE.